Apprentissage des statistiques avec Jamovi

Danielle Navarro, David Foxcroft, Jean-Marc Meunier

11/12/2019

Apprentissage des statistiques avec Jamovi

Un tutoriel pour les étudiants en psychologie et autres débutants

(Version française 0.70.2)

Danielle Navarro, University of New South Wales, d.navarro@unsw.edu.au

David Foxcroft, Oxford Brookes University , david.foxcroft@brookes.ac.uk

Traduction : **Jean-Marc Meunier**, Université Paris 8, jmeunier@univ-paris8.fr

INSTITUT D'ENSEIGNEMENT A DISTANCE - UNIVERSITE PARIS 8

Apprentissage des statistiques avec Jamovi : Un tutoriel pour les étudiants en psychologie et autres débutants

Danielle Navarro, David Foxcroft, Jean-Marc Meunier (Trad)

2020-10-06

Vue d'ensemble

L'apprentissage des statistiques avec Jamovi couvre le contenu d'un cours d'introduction à la statistique, tel qu'il est généralement enseigné aux étudiants de premier cycle en psychologie. Le livre aborde la façon de commencer dans Jamovi et donne une introduction à la manipulation des données. D'un point de vue statistique, l'ouvrage traite d'abord des statistiques descriptives et la représentation graphique, puis de la théorie des probabilités, de l'échantillonnage et de l'estimation et de la vérification des hypothèses nulles. Après avoir présenté la théorie, le livre couvre l'analyse des tableaux de contingence, la corrélation, les tests t, la régression, l'ANOVA et l'analyse factorielle. Les statistiques bayésiennes sont présentées à la fin du livre.

Citation

Citation de la version française

Navarro D.J., Foxcroft, D.R. (2020). Apprentissage des statistiques avec Jamovi : un tutoriel pour les étudiants en psychologie et autres débutants. (Version 0.70.2). (J.M. Meunier, Trad.) https://jmeunierp8.github.io/ManuelJamovi/index.html

Citation de la version anglaise

Navarro, D. J., & Foxcroft, D. R. (2019). Learning statistics with Jamovi: a tutorial for psychology students and other beginners. (Version 0.70). http://www.learnstatswithJamovi.com

Pour signaler une erreur ou une coquille, merci d'utiliser ce fil d'annotation. Un bref tutoriel est disponible ici.

Préface à la version 0.70

Cette mise à jour de la version 0.65 introduit de nouvelles analyses. Dans les chapitres ANOVA, nous avons ajouté des sections sur les mesures répétées ANOVA et l'analyse de la covariance (ANCOVA). Dans un nouveau chapitre, nous avons présenté l'analyse factorielle et les techniques connexes. Espérons que le style de ce nouveau matériel est cohérent avec le reste du livre, bien que les lecteurs attentifs pourraient vouloir mettre un peu plus

l'accent sur les explications conceptuelles et pratiques, et un peu moins sur l'algèbre. Je ne suis pas sûr que ce soit une bonne chose d'ajouter l'algèbre un peu plus tard. Cela reflète à la fois mon approche de la compréhension et de l'enseignement de la statistique, ainsi que les commentaires que j'ai reçus des étudiants lors d'un cours que j'enseigne. En cohérence avec cela dans le reste du livre, j'ai essayé de séparer les parties concernant l'algèbre en le mettant dans une boîte ou un cadre. Ce n'est pas que ce n'est pas important ou utile, mais pour certains élèves, ils peuvent vouloir passer outre et donc le cadre pour ces parties devrait aider certains lecteurs.

Pour cette version, je suis très reconnaissante pour les commentaires et réactions de mes étudiants et collègues, notamment Wakefield Morys-Carter, ainsi que de nombreuses personnes dans le monde entier qui m'ont envoyé de petites suggestions et corrections - très appréciées, et qui continuent à venir ! Une nouveauté assez intéressante est que les fichiers de données utilisés en exemple dans le livre peuvent maintenant être téléchargés dans Jamovi en tant que module additionnel - merci à Jonathon Love pour son aide à cet égard.

David Foxcroft, 1er février 2019

Note du traducteur

Souhaitant moi-même utiliser Jamovi avec mes étudiants de l'institut d'enseignement à distance de l'université Paris 8, il me fallait un manuel en français à mettre à leur disposition. Comme toujours, avant de se lancer dans ce travail passionnant, mais il faut bien le dire un peu ingrat parce que peu valorisé dans nos activités de chercheurs, un tour d'horizon sur la toile permet de voir ce que les collègues ont déjà réalisés. C'est ainsi que j'ai découvert cet excellent manuel. Je n'ai pas trouvé d'équivalent en français. Ce contenu va bien au-delà de ce que je leur enseigne habituellement parce que j'ai tendance, comme bien des collègues à réduire l'ambition à mesure que se réduisent les heures consacrées à la méthodologie. C'est bien sûr une erreur, mais il faut bien trouver un compromis. L'autre voie consiste à concevoir un cours non pas comme une synthèse des connaissances à un moment donné, mais comme un trousseau de clés permettant l'explorer un domaine de connaissances. C'est toutes la philosophie du projet Ontostats que je mène avec quelques collègues (Ces clés sont les concepts et les méthodes, mais aussi les controverses qui animent la communauté scientifique. De mon point de vue cet ouvrage, permet de relever ce défi. Il est plus qu'un manuel d'utilisation de Jamovi. Si vous regardez bien la part qui est consacrée au logiciel est d'ailleurs assez congrue et surtout bien séparée de la présentation conceptuelle des méthodes. Il pourrait donc parfaitement être mixé avec la présentation d'autres logiciels sans que cela nuise à la cohérence de l'ensemble. Cela a déjà été fait puisqu'à l'origine ce manuel s'appuyait sur le logiciel R (Daniel Joseph Navarro 2014) et qu'il existe une version de ce même manuel pour le logiciel JASP (Danielle J Navarro, Foxcroft, and Faulkenberry 2019). De mon point de vue, c'est une grande qualité.

Près de cinq cents pages pour un manuel, même s'il couvre les trois années du premier cycle (et un peu au-delà), c'est énorme, bien plus que la plupart des manuels. Je comprendrais que certains étudiants soient effrayés par l'ampleur de la tâche surtout sur un domaine aussi ardu et à certains égards rebutant que les statistiques. A ces étudiants, je recommande de lire l'excellent chapitre introductif et surtout l'épilogue. Je veux aussi

rappeler à tous les étudiants qu'un manuel n'est pas une bible, mais un outil permettant de répondre à un certain nombre de questions et de résoudre des problèmes, notamment ceux que vous rencontrerez dans vos exercices ou dans votre future activité professionnelle. Il ne présente pas de vérité absolue et les auteurs le rappellent régulièrement dans l'ouvrage. Cependant, pour qu'il devienne instrument, vous devez vous approprier l'ouvrage et acquérir les concepts nécessaires à sa manipulation efficiente. En d'autres termes, ce qui est attendu de vous n'est pas votre capacité à restituer le contenu du manuel, mais votre capacité à vous en servir pour résoudre des problèmes. C'est pour cela que vous avez encore des cours en plus du manuel. Ce dernier ne contient pas d'exercice. Ce sera probablement pour une prochaine version. Mais parallèlement au cours ou après lors d'un travail de recherche, un ouvrage tel que celui-ci ne se consulte pas uniquement de manière linéaire, même si les chapitres ont été pensés pour répondre à une progression et même si votre enseignant vous a prescrit la lecture de certains chapitres. Les multiples renvois dans les chapitres facilitent largement une telle lecture et vous pouvez aussi l'enrichir avec vos propres signets ou notes personnelles d'autant plus facilement que les outils informatiques d'annotations permettent de travailler collaborativement autour d'un même document (voir par exemple les possibilités offertes par outil comme hypothesis que nous utilisons dans un autre projet à l'institut).[1]

Ne me sentant ni l'envie, ni les compétences de le singer en réécrivant quelque chose de similaire à ma sauce, il m'est apparu évident qu'il fallait en faire une traduction, choix d'autant plus facile que les auteurs ont eu la gentillesse de mettre l'ouvrage sous licence creative commons. La traduction que je vous propose ici a été réalisée durant mes congés d'été, faute d'avoir le temps avec le reste de mes multiples activités. Mes compétences en anglais étant modestes, je confesse m'être appuyé pour certains passages sur des outils de traduction automatiques (deepl.com) pour dégrossir le travail. Il en reste peut-être quelques scories, je vous prie de m'en excuser et de me les signaler gentiment pour les prochaines versions. Je n'ai également pas pris le temps de traduire les graphiques ou certains tableaux. En ce qui concerne l'utilisation de Jamovi, J'ai fait le choix de ne pas traduire les références aux différents menus et commandes tant que nous n'avons pas de version française afin de ne pas perdre les lecteurs lorsqu'ils sont face du logiciel. Enfin, au cours de ce travail, j'ai découvert ou redécouvert beaucoup de choses. Il est très probable que cette traduction contienne des erreurs et sûrement beaucoup de coquilles, mais comme les erreurs de traduction, si vous me les signalez gentiment. Celles-ci sont recensées dans le fil d'annotation hypothesis permettant de proposer le correctif au fur et à mesure. Bien sûr, celles-ci seront corrigées dans la prochaine version.

Jean-Marc Meunier, le 5 octobre 2020

[1] Il s'agit du projet PEPE. Pour hypothesis, c'est ici que cela se passe

Pourquoi apprenons-nous les statistiques ?

« Tu ne répondras pas à des questionnaires ou à des quiz sur les affaires du monde, ni ne passeras de tests. Tu ne t'assiéras pas avec des statisticiens, et tu ne commettras pas une science sociale »

- W.H. Auden[2]

Sur la psychologie des statistiques

À la surprise de nombreux étudiants, les statistiques représentent une part assez importante de l'enseignement en psychologie. À la surprise de personne, la statistique n'est que très rarement la partie *préférée* de l'enseignement en psychologie. Après tout, si vous aimiez vraiment l'idée de faire des statistiques, vous seriez probablement inscrit à un cours de statistique en ce moment, pas à un cours de psychologie. Il n'est donc pas surprenant qu'une proportion assez importante de la population étudiante ne soit pas satisfaite du fait que la psychologie comporte autant de statistiques. Dans ce contexte, j'ai pensé que le bon point de départ pourrait être de répondre à certaines des questions les plus courantes que les gens se posent au sujet des statistiques.

Une grande partie de cette question est liée à l'idée même de statistiques. Qu'est-ce que c'est ? C'est pour quoi faire ? Et pourquoi les scientifiques sont-ils si obsédés par ça ? Ce sont toutes de bonnes questions, quand on y pense. Commençons par la dernière. En tant que groupe, les scientifiques semblent bizarrement obsédés par la réalisation de tests statistiques sur tout. En fait, nous utilisons si souvent les statistiques que nous oublions parfois d'expliquer aux gens pourquoi nous le faisons. C'est une sorte d'acte de foi parmi les scientifiques - et en particulier les spécialistes des sciences sociales - que l'on ne peut pas faire confiance aux découvertes tant qu'on n'a pas fait quelques statistiques. On pourrait pardonner aux étudiants de premier cycle de penser que nous sommes tous complètement fous, parce que personne ne prend le temps de répondre à une question très simple :

Pourquoi faites-vous des statistiques ? Pourquoi les scientifiques ne font-ils pas preuve de bon sens ?

C'est une question naïve à certains égards, mais la plupart des bonnes questions le sont. Il y a beaucoup de bonnes réponses,[3] mais pour moi, la meilleure réponse est très simple : nous ne nous faisons pas assez confiance. Nous nous inquiétons d'êtres humains et sensibles à tous les préjugés, tentations et fragilités dont souffrent les humains. Une grande partie des statistiques est essentiellement une sauvegarde. Utiliser le « bon sens » pour évaluer les preuves, c'est se fier à son instinct, s'appuyer sur des arguments verbaux et utiliser la force brute de la raison humaine pour trouver la bonne réponse. La plupart des scientifiques ne pensent pas que cette approche puisse fonctionner.

---

[2] La citation provient du poème d'Auden de 1946 *Under Wich Lyre : A Reactionary Tract for the Times*, prononcé dans le cadre d'un discours d'ouverture à l'Université Harvard. L'histoire du poème est intéressante : http://harvardmagazine.com/2007/11/a-poets-warning.html

[3] Y compris la suggestion que le bon sens fait défaut parmi les scientifiques.

En fait, à bien y penser, cela ressemble beaucoup à une question psychologique pour moi, et puisque je travaille dans un département de psychologie, il me semble que c'est une bonne idée de la creuser un peu plus ici. Est-il vraiment plausible de penser que cette approche de « bon sens » est très fiable ? Les arguments verbaux doivent être construits avec le langage, et toutes les langues ont des préjugés - certaines choses sont plus difficiles à dire que d'autres, et pas nécessairement parce qu'elles sont fausses (par exemple, l'électrodynamique quantique est une bonne théorie, mais difficile à expliquer en mots). Les intuitions de notre « instinct » ne sont pas faites pour résoudre des problèmes scientifiques, elles sont faites pour gérer des inférences quotidiennes - et comme l'évolution biologique est plus lente que les changements culturels, nous devrions dire qu'elles sont faites pour résoudre des problèmes quotidiens dans *un monde différent* de celui où nous vivons. Plus fondamentalement, le raisonnement exige des gens qu'ils s'engagent dans une « induction », qu'ils fassent des suppositions sages et qu'ils aillent au-delà de l'évidence immédiate des sens pour faire des généralisations sur le monde. Si vous pensez que vous pouvez le faire sans être influencé par divers distracteurs, eh bien, j'ai un pont à Londres que j'aimerais vous vendre. Comme nous le montrons dans la section suivante, nous ne pouvons même pas résoudre des problèmes « déductifs » (ceux pour lesquels il n'est pas nécessaire de deviner) sans être influencés par nos biais préexistants.

La malédiction des biais de croyance

Les gens sont plutôt intelligents. Nous sommes certainement plus intelligents que les autres espèces avec lesquelles nous partageons la planète (bien que beaucoup de gens puissent être en désaccord). Nos esprits sont des choses tout à fait étonnantes, et nous semblons être capables des exploits les plus incroyables de pensée et de raison. Mais ça ne nous rend pas parfaits. Et parmi les nombreuses choses que les psychologues ont montrées au fil des ans, il y a le fait que nous avons vraiment de la difficulté à être neutres, à évaluer les preuves de façon impartiale et sans être influencés par des préjugés préexistants. Un bon exemple en est **le biais de croyance** dans le raisonnement logique : si vous demandez aux gens de décider si un argument particulier est logiquement valide (c.-à-d. si la conclusion est vraie si les prémisses sont vraies), nous avons tendance à être influencés par la crédibilité de la conclusion, même lorsque nous ne le devrions pas. Par exemple, voici un argument valide dont la conclusion est crédible :

Toutes les cigarettes sont chères (Prémisse 1) Certaines choses qui créent une dépendance sont peu coûteuses (Prémisse 2) Par conséquent, certaines choses qui créent une dépendance ne sont pas des cigarettes (Conclusion)

Et voici un argument valide dont la conclusion n'est pas crédible :

Toutes les choses qui créent une dépendance coûtent cher (Prémisse 1) Certaines cigarettes sont bon marché (Prémisse 2) Par conséquent, certaines cigarettes ne créent pas de dépendance (Conclusion)

La structure logique de l'argument #2 est identique à celle de l'argument #1, et les deux sont valides. Toutefois, dans le deuxième argument, il y a de bonnes raisons de penser que la prémisse 1 est incorrecte et, par conséquent, il est probable que la conclusion est également incorrecte. Mais cela n'a rien à voir avec le sujet à l'étude ; un argument est déductiblement valable si la conclusion est une conséquence logique des prémisses. C'est-à-

dire qu'un argument valide n'a pas besoin d'impliquer de vraies déclarations. Considérons maintenant un argument invalide qui a une conclusion crédible :

Toutes les choses qui créent une dépendance coûtent cher (Prémisse 1) Certaines cigarettes sont bon marché (Prémisse 2) Par conséquent,certaines choses qui créent une dépendance ne sont pas des cigarettes (Conclusion)

Et enfin, un argument invalide avec une conclusion non crédible :

Toutes les cigarettes sont chères (Prémisse 1) Certaines choses qui créent une dépendance sont peu coûteuses (Prémisse 2) Par conséquent,certaines cigarettes ne créent pas de dépendance (Conclusion)

Supposons maintenant que les gens soient parfaitement capables de mettre de côté leurs préjugés préexistants sur ce qui est vrai et ce qui ne l'est pas, et d'évaluer purement un argument sur ses mérites logiques. Nous nous attendrions à ce que 100 % des gens disent que les arguments valides sont valides, et à ce que 0 % des gens disent que les arguments invalides sont valides. Donc, si vous faite une expérience avec ces exemples, vous vous attendez à voir des données comme celle-ci :

	Conclusion perçue comme vraie	Conclusion perçue comme fausse
Argument valide	100 % disent qu'il est valide	100 % disent qu'il est valide
Argument non valide	0 % disent qu'il est valide	0 % disent qu'il est valide

Si les données psychologiques ressemblaient à ceci (ou même à une bonne approximation de ceci), nous pourrions nous sentir en sécurité en faisant simplement confiance à nos intuitions. Autrement dit, il serait tout à fait acceptable de laisser les scientifiques évaluer les données en fonction de leur bon sens, et de ne pas se préoccuper de toutes ces statistiques obscures. Cependant, vous avez pris des cours de psycho, et maintenant vous savez probablement où cela nous mènent.

Dans une étude classique, Evans, Barston et Pollard (1983) ont mené une expérience portant exactement sur cette question. Ce qu'ils ont découvert, c'est que lorsque les préjugés préexistants (c.-à-d. les croyances) étaient en accord avec la structure des données, tout allait comme on l'espérait :

	Conclusion perçue comme vraie	Conclusion perçue comme fausse
Argument valide	92 % disent qu'il est valide	
Argument non valide		8 % disent qu'il est valide

Ce n'est pas parfait, mais c'est assez bon. Mais regardez ce qui se passe quand nos croyances sur la vérité de la conclusion vont à l'encontre de la structure logique de l'argument :

	Conclusion perçue comme vraie	Conclusion perçue comme fausse
Argument valide	92 % disent qu'il est valide	**44 % disent qu'il est valide**

Argument non valide	**92 % disent qu'il est valide**	8 % disent qu'il est valide

Mince, ce n'est pas aussi bon. Apparemment, lorsqu'on présente aux gens un argument solide qui contredit nos croyances préexistantes, nous trouvons qu'il est assez difficile de le percevoir comme un argument solide (les gens ne le faisaient que 46 % du temps). Pire encore, lorsqu'on présente aux gens un argument faible qui correspond à nos préjugés préexistants, presque personne ne peut voir que l'argument est faible (les gens se trompent 92 % du temps !).[4]

Si vous y réfléchissez, ce n'est pas comme si ces données étaient terriblement accablantes. Dans l'ensemble, les gens ont fait mieux que par hasard pour compenser leurs préjugés antérieurs, car environ 60 % des jugements des gens étaient exacts (on s'attendrait à ce que 50 % le soient avec le hasard). Malgré tout, si vous étiez un « évaluateur professionnel des données probantes « et que quelqu'un vous offrait un outil magique qui augmente vos chances de prendre la bonne décision de 60 % à 95 %, vous sauteriez probablement sur l'occasion, non ? Bien sûr que vous le feriez. Heureusement, nous avons un outil qui nous permet de le faire. Mais ce n'est pas de la magie, ce sont des statistiques. C'est donc la raison pour laquelle les scientifiques adorent les statistiques. C'est *trop facile* pour nous de « croire ce que nous voulons croire ». Donc, si nous voulons « croire aux données », nous allons avoir besoin d'un peu d'aide pour garder nos préjugés personnels sous contrôle. C'est ce que font les statistiques, ça nous aide à rester honnêtes.

La mise en garde contre le paradoxe de Simpson

Ce qui suit est une histoire vraie (je pense !). En 1973, l'Université de Californie, à Berkeley, s'inquiétait de l'admission d'étudiants dans leurs cours de troisième cycle. Plus précisément, ce qui a causé le problème, c'est la répartition par sexe de leurs admissions, qui ressemblait à ceci :

	Nombre de candidats	Pourcentage d'admis
Hommes	8442	44%
Femmes	4321	35%

Compte tenu de cela, ils craignaient d'être poursuivis en justice ![5] Étant donné qu'il y avait près de 13 000 candidats, une différence de 9 % dans les taux d'admission entre les hommes et les femmes est beaucoup trop grande pour que ce soit une coïncidence. Des

---

[4] Dans mes moments les plus cyniques, j'ai l'impression que ce seul fait explique 95% de ce que je lis sur internet.

[5] Des versions antérieures de ces notes laissaient entendre, à tort, qu'elles faisaient l'objet d'une poursuite. Mais ce n'est pas vrai. Il y a un joli commentaire à ce sujet ici : https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html. Un grand merci à Wilfried Van Hirtum de me l'avoir signalé.

données assez convaincantes, n'est-ce pas ? Et si je vous disais que ces données reflètent *en fait* un faible biais en faveur des femmes (en quelque sorte !), vous penseriez probablement que je suis soit folle, soit sexiste.

Bizarrement, c'est en fait en partie vrai. Lorsque les gens ont commencé à examiner plus attentivement les données sur les admissions, ils ont rapporté une histoire assez différente (Bickel, Hammel, and O'Connell 1975). Plus précisément,lorsqu'ils l'ont examiné département par département, il s'est avéré que la plupart des départements avaient en fait un taux de réussite légèrement *plus élevé* pour les femmes que pour les hommes. Le tableau ci-dessous indique le nombre d'admissions pour les six plus grands départements (les noms des départements ont été supprimés pour des raisons de confidentialité) :

	Hommes		Femmes	
Département	Candidats	Pourcentage d'admissions	Candidates	Pourcentage d'admissions
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Fait remarquable, la plupart des départements avaient un taux d'admission *plus élevé* pour les femmes que pour les hommes ! Pourtant, le taux global d'admission à l'université était *plus faible* chez les femmes que chez les hommes. Comment est-ce possible ? Comment ces deux affirmations peuvent-elles être vraies en même temps ?

Voilà ce qui se passe. Tout d'abord, notons que les départements *ne sont pas* égaux entre eux en termes de pourcentage d'admission : certains départements (par exemple, A, B) avaient tendance à admettre un pourcentage élevé de candidats qualifiés, alors que d'autres (par exemple, F) avaient tendance à rejeter la plupart des candidats, même si ceux-ci étaient de grande qualité. Ainsi, parmi les six départements indiqués ci-dessus, notez que le département A est le plus généreux, suivi de B, C, D, E et F dans cet ordre. Ensuite, remarquez que les hommes et les femmes avaient tendance à candidater dans des départements différents. Si nous classons les départements en fonction du nombre total de candidats masculins reçus, nous obtenons A>B>D>C>F>E (les départements « faciles » sont en gras). Dans l'ensemble, les hommes avaient tendance à présenter une demande dans les départements où le taux d'admission était élevé. Maintenant, comparez ceci à la façon dont les candidates se sont distribuées. En classant les départements en fonction du nombre total de candidatures féminines, on obtient un classement tout à fait différent : C>E>D>F>A>B. En d'autres termes, ce que ces données semblent suggérer, c'est que les candidates avaient tendance à postuler dans des départements « plus durs ». En fait, si nous regardons la Figure 1-1, nous constatons que cette tendance est systématique et assez frappante. Cet effet est connu sous le nom de **paradoxe de Simpson**. Ce n'est pas courant, mais cela arrive dans la vraie vie, et la plupart des gens en sont très surpris lorsqu'ils le

rencontrent pour la première fois, et beaucoup de gens refusent même de croire que c'est réel. C'est très réel. Bien que beaucoup de leçons statistiques très subtiles sont dissimulées derrière ce fait, je veux m'en servir pour souligner un point beaucoup plus important : il est difficile de faire de la recherche et il y a beaucoup de *pièges* subtils et contre-intuitifs qui attendent les personnes imprudentes. C'est la deuxième raison pour laquelle les scientifiques adorent les statistiques et pour laquelle nous enseignons les méthodes de recherche. Parce que la science est difficile et que la vérité est parfois astucieusement cachée dans les coins et les recoins de données complexes.

Avant d'en terminer avec ce sujet, j'aimerais souligner une autre chose vraiment critique qui est souvent négligée dans un cours de méthodologie de la recherche. Les statistiques ne résolvent qu'*une partie du* problème. Rappelez-vous que nous avons commencé tout cela avec la crainte que les processus d'admission de Berkeley pourraient être injustement biaisés à l'encontre des candidates. Lorsque nous avons examiné les données « agrégées », il nous a semblé que l'université faisait de la discrimination à l'égard des femmes, mais lorsque nous « désagrégeons » et que nous examinons le comportement individuel de tous les départements, il s'avère que les départements eux-mêmes étaient, le cas échéant, légèrement biaisés en faveur des femmes.

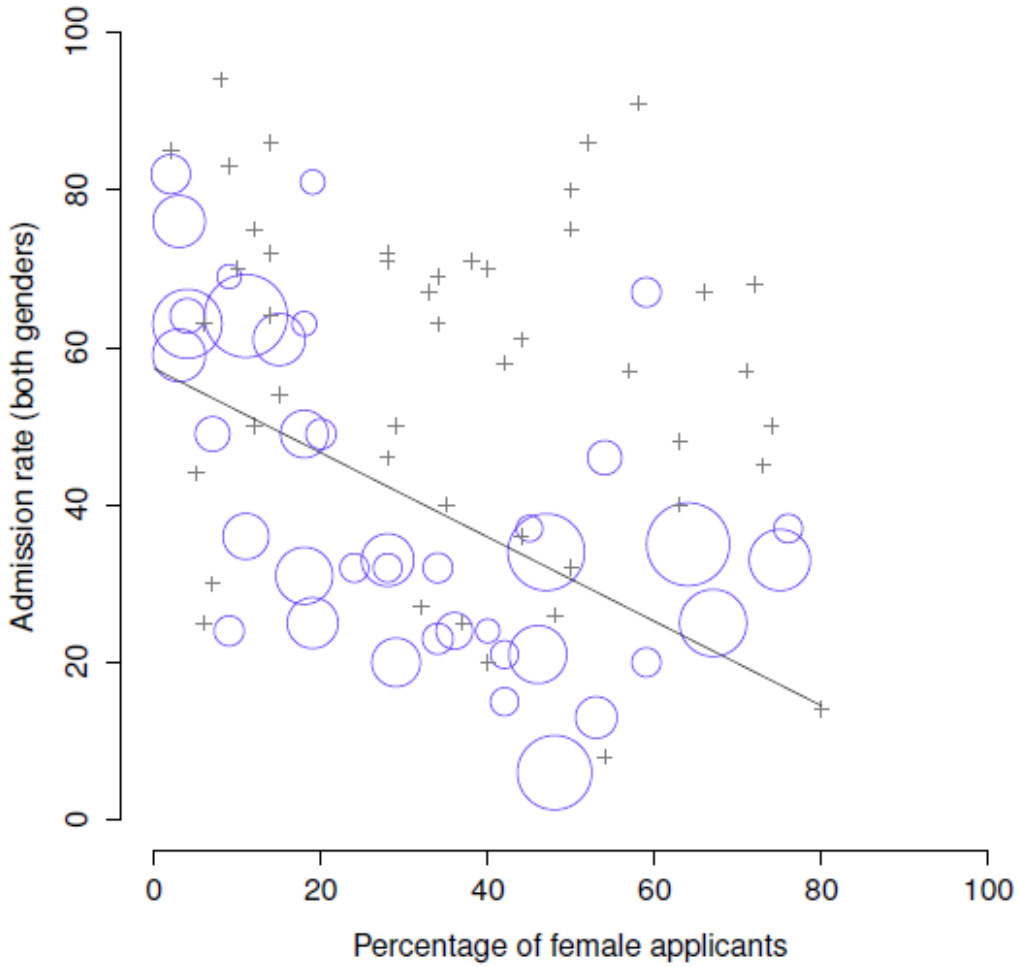


Figure 1-1 : Les données de Berkeley 1973 sur les admissions à l'université. Ce chiffre représente le taux d'admission pour les 85 départements qui comptaient au moins une femme candidate, en fonction du pourcentage de femmes candidates. Le graphique est un remaniement de la figure 1 de Bickel, Hammel et O'Connell (1975). Les cercles représentent les départements qui comptent plus de 40 candidats ; la superficie du cercle est proportionnelle au nombre total de candidats. Le graphique représente les départements qui comptent moins de 40 candidats.

Le biais sexiste dans le nombre total d'admissions était dû au fait que les femmes avaient tendance à candidater dans les départements plus difficile. D'un point de vue juridique, cela met probablement l'université à l'abri. Les admissions aux études supérieures sont déterminées au niveau de chaque département et il y a de bonnes raisons de le faire. Au niveau des différents départements, les décisions sont plus ou moins impartiales (le biais en faveur des femmes à ce niveau est faible et n'est pas uniforme dans tous les départements). Étant donné que l'université ne peut pas dicter les départements auxquels les gens choisissent de s'adresser et que la prise de décision se fait au niveau du département, elle ne peut guère être tenue responsable des biais que ces choix produisent.

C'est ce qui a motivé les remarques quelque peu désinvoltes que j'ai faites précédemment, mais ce n'est pas le problème. Après tout, si cela nous intéresse d'un point de vue plus sociologique et psychologique, nous pourrions nous demander *pourquoi* il y a de si grandes différences entre les genres en ce qui concerne les candidatures. Pourquoi les hommes ont-ils tendance à s'inscrire plus souvent que les femmes au programme d'ingénieur, et pourquoi cette tendance s'inverse-t-elle pour le département d'anglais ? Pourquoi les départements qui ont tendance à avoir un biais pour les demandes d'admission des femmes ont-ils tendance à avoir des taux d'admission globaux inférieurs à ceux des départements qui ont un biais pour les demandes des hommes ? Ne s'agit-il pas là d'un préjugé sexiste, même si tous les départements sont eux-mêmes impartiaux ? C'est possible. Posons, hypothétiquement, que les hommes préfèrent les « sciences dures » et que les femmes préfèrent les « sciences humaines ». Supposons en outre que la raison pour laquelle les départements des humanités ont de faibles taux d'admission est que le gouvernement ne veut pas financer les humanités (les places de doctorat, par exemple, sont souvent liées aux projets de recherche financés par le gouvernement). Est-ce que cela constitue un préjugé sexiste ? Ou simplement une vision non éclairée de la valeur des sciences humaines ? Que ce passerait-il si quelqu'un de haut placé au sein du gouvernement coupait les fonds des humanités parce qu'il estimait que les humanités sont des « choses inutiles pour les nanas » ? Cela semble assez *ouvertement* sexiste. Rien de tout cela ne relève de la statistique, mais c'est important pour le projet de recherche. Si vous vous intéressez aux effets structurels globaux des subtils préjugés sexistes, vous souhaiteriez probablement examiner à la *fois* les données agrégées et désagrégées. Si vous êtes intéressé par le processus de prise de décision à Berkeley même, vous n'êtes probablement intéressé que par les données désagrégées.

Bref, il y a beaucoup de questions critiques auxquelles vous ne pouvez pas répondre avec des statistiques, mais les réponses à ces questions auront un impact énorme sur la façon dont vous analyserez et interpréterez les données. C'est la raison pour laquelle vous devriez toujours considérer les statistiques comme un *outil* pour vous aider à mieux connaître vos

données. Ni plus ni moins. C'est un outil puissant à cette fin, mais rien ne peut remplacer une réflexion approfondie.

Statistiques en psychologie

J'espère que la discussion ci-dessus a contribué à expliquer pourquoi la science en général est si axée sur les statistiques. Mais je suppose que vous vous posez beaucoup plus de questions sur le rôle que jouent les statistiques en psychologie, et plus précisément pourquoi les programmes de psychologie consacrent toujours autant de cours aux statistiques. Voici donc ma tentative pour répondre à quelques-unes d'entre elles…

Pourquoi la psychologie a-t-elle autant de statistiques ?

Pour être tout à fait honnête, il y a plusieurs raisons dont certaines sont meilleures que d'autres. La raison la plus importante est que la psychologie est une science statistique. Ce que j'entends par là, c'est que les « choses » que nous étudions sont des *personnes*. Des gens réels, compliqués, glorieusement désordonnés, furieusement pervers. Les « choses » de la physique incluent les objets comme les électrons, et bien qu'il y ait toutes sortes de complexités qui surgissent en physique, les électrons n'ont pas leur propre esprit. Ils n'ont pas d'opinions, ils ne diffèrent pas les uns des autres de façon bizarre et arbitraire, ils ne s'ennuient pas au milieu d'une expérience, et ils ne se fâchent pas contre l'expérimentateur et n'essaient pas délibérément de saboter les données (non pas que je l'ai jamais fait !). Au fond, la psychologie est plus difficile que la physique.[6]

Fondamentalement, nous vous enseignons les statistiques en tant que psychologues parce que vous devez être meilleurs en statistiques que les physiciens. Il y a en fait un dicton utilisé parfois en physique, selon lequel « *si votre expérience a besoin de statistiques, vous auriez dû faire une meilleure expérience* ». Ils ont le luxe de pouvoir dire cela parce que leurs objets d'étude sont d'une simplicité pathétique par rapport au vaste désordre auquel sont confrontés les spécialistes des sciences sociales. Et ce n'est pas seulement la psychologie. La plupart des sciences sociales dépendent désespérément des statistiques. Pas parce que nous sommes de mauvais expérimentateurs, mais parce que nous avons choisi un problème plus difficile à résoudre. Nous vous enseignons les statistiques parce que vous en avez vraiment, vraiment besoin.

Quelqu'un d'autre ne peut-il pas faire les statistiques ?

Dans une certaine mesure, mais pas complètement. Il est vrai qu'il n'est pas nécessaire de devenir un statisticien pleinement formé uniquement pour faire de la psychologie, mais qu'il faut atteindre un certain niveau de compétence statistique. À mon avis, il y a trois raisons pour lesquelles tout chercheur en psychologie devrait être en mesure de faire des statistiques de base :

---

[6] Ce qui pourrait expliquer pourquoi la physique est un peu plus avancée que nous en tant que science.

- Premièrement, il y a la raison fondamentale : les statistiques sont étroitement liées à la conception de la recherche. Si vous voulez être bon dans la conception d'études psychologiques, vous devez à tout le moins comprendre les bases des statistiques.
- Deuxièmement, si vous voulez être bon du point de vue de la recherche en psychologie, vous devez être capable de comprendre la littérature psychologique. Mais presque tous les articles de la littérature psychologique rapportent les résultats d'analyses statistiques. Donc, si vous voulez vraiment comprendre la psychologie, vous devez être capable de comprendre ce que d'autres personnes ont fait avec leurs données. Et cela signifie comprendre un certain nombre de statistiques.
- Troisièmement, il y a un gros problème pratique à dépendre d'autres personnes pour faire toutes vos statistiques : l'analyse statistique *coûte cher*. Si vous vous ennuyez et que vous voulez savoir combien le gouvernement australien demande pour les frais universitaires, vous remarquerez quelque chose d'intéressant : les statistiques sont désignées comme une catégorie « priorité nationale », et donc les frais sont beaucoup, beaucoup plus bas que pour tout autre domaine d'études. C'est parce qu'il y a une pénurie massive de statisticiens. Donc, de votre point de vue de chercheur en psychologie, les lois de l'offre et de la demande ne sont pas exactement de votre côté ! Par conséquent, chaque fois que vous voudrez faire de la recherche en psychologie, la réalité cruelle vous rappellera que vous n'avez pas assez d'argent pour payer un statisticien. L'économie de la situation signifie donc qu'il faut être assez autosuffisant.

Il est à noter qu'un grand nombre de ces raisons se généralisent au-delà des chercheurs. Si vous voulez être un psychologue praticien et rester à la pointe du domaine, il est utile de pouvoir lire la littérature scientifique, qui repose en grande partie sur les statistiques.

- Je me fiche des emplois, de la recherche ou du travail clinique. Ai-je besoin de statistiques ?

D'accord, maintenant vous vous moquez de moi. Pourtant, je pense que ça devrait compter pour vous aussi. Les statistiques devraient compter pour vous de la même façon que les statistiques devraient compter pour *tout le monde*. Nous vivons au XXIe siècle, et les données sont *partout*. Franchement, étant donné le monde dans lequel nous vivons de nos jours, une connaissance de base des statistiques s'apparente à un outil de survie ! C'est le sujet de la section suivante.

Statistiques dans la vie quotidienne

"Nous nous noyons dans l'information, mais nous sommes affamés de connaissances" - Divers auteurs, original probablement John Naisbitt

Lorsque j'ai commencé à rédiger mes notes de cours, j'ai pris les 20 articles les plus récents affichés sur le site Web d'ABC. Sur ces 20 articles, il s'est avéré que 8 d'entre eux portaient sur un sujet que j'appellerais un sujet statistique et que 6 d'entre eux comportaient une erreur. L'erreur la plus courante, si vous êtes curieux, était de ne pas avoir rapporté les données de base (par exemple, l'article mentionne que 5 % des personnes dans la situation X ont une caractéristique Y, mais ne dit pas à quel point cette caractéristique est commune à tous les autres ! Ce que j'essaie de dire ici, ce n'est pas que les journalistes sont mauvais en statistiques (bien qu'ils le soient presque toujours), c'est qu'une connaissance de base des

statistiques est très utile pour essayer de comprendre quand quelqu'un d'autre fait une erreur ou même vous ment. En fait, l'une des plus importantes choses qu'une connaissance des statistiques vous apporte est de vous mettre en colère contre un journal ou un site Internet beaucoup plus souvent. Vous en trouverez un bon exemple à la section 4.1.5. Dans les versions ultérieures de ce livre, j'essaierai d'inclure plus d'anecdotes en ce sens.

Les méthodes de recherche ne se limitent pas aux statistiques

Jusqu'à présent, j'ai surtout parlé de statistiques, et il vous serait donc pardonné de penser que les statistiques sont tout ce qui m'importe dans la vie. Pour être juste, vous n'auriez pas tort, mais la méthodologie de recherche est un concept plus large que les statistiques. Ainsi, la plupart des cours de méthodologie de la recherche couvriront un grand nombre de sujets qui se rapportent beaucoup plus à la pratique de la conception de la recherche, et en particulier les questions que vous rencontrez lorsque vous essayez de faire de la recherche avec des humains. Cependant, environ 99 % des *craintes* des étudiants ont trait à la partie statistique du cours, alors je me suis concentré sur les statistiques dans cette discussion, et j'espère vous avoir convaincu que les statistiques sont importantes et, surtout, qu'il ne faut pas en avoir peur. Cela dit, il est assez typique que les cours d'introduction aux méthodes de recherche soient très riches en statistiques. Ce n'est pas (habituellement) parce que les enseignants sont mauvais. C'est plutôt le contraire. Les cours d'introduction se concentrent beaucoup sur les statistiques parce que vous avez presque toujours besoin de statistiques avant d'avoir besoin d'une formation sur les autres méthodes de recherche. Pourquoi ? Parce que presque tous vos travaux dans d'autres classes reposeront sur la formation en statistique, dans une bien plus grande mesure qu'ils ne reposent sur d'autres outils méthodologiques. Il n'est pas courant que les travaux de premier cycle exigent que vous conceviez votre propre étude à partir de toutes pièces (auquel cas vous auriez besoin d'en savoir beaucoup sur la conception de la recherche), mais il est courant que vous deviez analyser et interpréter des données recueillies dans une étude qu'un tiers a conçue (auquel cas vous devez disposer de statistiques). En ce sens et afin de vous permettre de réussir dans tous vos autres cours, connaitre les statistiques est urgent.

Mais notez que « urgent » est différent de « important » - les deux sont importants. Je tiens vraiment à souligner que la conception de la recherche est tout aussi importante que l'analyse des données, et ce livre y consacre beaucoup de temps. Cependant, bien que les statistiques aient une sorte d'universalité et fournissent un ensemble d'outils de base qui sont utiles pour la plupart des types de recherche, en psychologie, les méthodes de recherche ne sont pas aussi universelles. Il y a quelques principes généraux auxquels tout le monde devrait réfléchir, mais une grande partie de la conception de la recherche est très idiosyncrasique et propre au domaine de recherche dans lequel vous voulez vous engager. Dans la mesure où ce sont les détails qui sont significatifs, ils n'apparaissent généralement pas dans les statistiques d'introduction et les cours de méthodologie de recherche.

Brève introduction à la conception de la recherche

Consulter le statisticien à la fin d'une expérience, c'est souvent simplement lui demander de faire un examen post mortem. Il peut peut-être dire de quoi l'expérience est morte

- Sir Ronald Fisher[7]

Dans ce chapitre, nous allons commencer à réfléchir aux idées de base qui entrent dans la conception d'une étude, la collecte de données, la vérification de l'efficacité de votre collecte de données, etc. Il ne vous donnera pas assez d'information pour vous permettre de concevoir vos propres études, mais il vous donnera un grand nombre des outils de base dont vous avez besoin pour évaluer les études faites par d'autres personnes. Cependant, comme ce livre est beaucoup plus axé sur l'analyse des données que sur la collecte de données, je ne donne qu'un bref aperçu. Notez que ce chapitre est « spécial » de deux façons. Premièrement, c'est beaucoup plus spécifique à la psychologie que les chapitres suivants. Deuxièmement, il se concentre beaucoup plus sur le problème scientifique de la méthodologie de la recherche, et beaucoup moins sur le problème statistique de l'analyse des données. Néanmoins, les deux problèmes sont liés l'un à l'autre, de sorte qu'il est de tradition que les manuels de statistiques discutent du problème de façon un peu plus détaillée. Ce chapitre s'appuie fortement sur (Campbell and Stanley 1967) pour l'analyse de la conception de l'étude, et sur Stevens (1946) pour l'analyse des échelles de mesure.

Introduction à la mesure psychologique

La première chose à comprendre est que la collecte de données peut être considérée comme une sorte de **mesure**. Ce que nous essayons de faire ici, c'est de mesurer quelque chose au sujet du comportement humain ou de l'esprit humain. Qu'est-ce que j'entends par « mesure » ?

Quelques réflexions sur la mesure psychologique

La mesure elle-même est un concept subtil, mais il s'agit essentiellement de trouver un moyen d'attribuer des numéros, ou des étiquettes, ou d'autres types de descriptions bien définies, aux « choses ». Donc, n'importe lequel des éléments suivants compterait comme une mesure psychologique :

- J'ai *33 ans*.
- Je *n'aime pas* **les anchois**.
- Mon **sexe chromosomique** est *masculin*.
- Je m'**identifie comme** un *homme*.[8]

---

[7] Allocution présidentielle au premier Congrès indien de la statistique, 1938. Source : https://en.wikiquote.org/wiki/Ronald_Fisher

[8] C'est ennuyeux. Cette section est l'une des parties les plus anciennes du livre, et elle est dépassée et plutôt embarrassante. Je l'ai écrit en 2010, date à laquelle tous ces faits *étaient* véridiques. En 2018, je n'ai plus 33 ans, mais ce n'est pas surprenant. Je ne peux pas imaginer que mes chromosomes ont changé, alors je vais supposer que mon caryotype était alors et est maintenant XY. Le genre auto-identifié, d'un autre côté… ah. Je suppose que le fait que la page de garde me désigne maintenant comme Danielle plutôt que Daniel pourrait être un indice, mais je ne m'identifie généralement pas comme un " homme " dans un

Dans la courte liste ci-dessus, la **partie en gras** est « la chose à mesurer », et la *partie en italique* est « la mesure elle-même ». En fait, nous pouvons nous étendre un peu sur ce point, en réfléchissant à l'ensemble des mesures possibles qui auraient pu survenir dans chaque cas :

- Mon **âge** (en années) aurait pu être *0, 1, 2, 3…*, etc. La limite supérieure de ce que mon âge pourrait être est un peu floue, mais dans la pratique, vous pouvez dire que l'âge le plus élevé possible est *150 ans*, puisqu'aucun humain n'a jamais vécu aussi longtemps.
- Quand on m'a demandé si j'**aimais les anchois**, j'ai peut-être répondu que *oui,* ou *non*, ou que *je n'avais pas d'opinion*, ou que *c'était parfois le cas*.
- Mon **sexe chromosomique** sera presque certainement *masculin (XY)* ou *féminin (XX),* mais il y a d'autres possibilités. Je pourrais aussi avoir le *syndrome de Klinfelter (XXY)*, qui est plus semblable à celui des hommes que des femmes. Et j'imagine qu'il y a aussi d'autres possibilités.
- Il est également très probable que je m'**identifie comme un** *homme* ou une *femme,* mais il n'est pas nécessaire que cela corresponde à mon sexe chromosomique. Je peux aussi choisir de *ne* m'identifier à *aucun des deux* ou de m'appeler explicitement *transgenre*.

Comme vous pouvez le voir, pour certaines choses (comme l'âge), il semble assez évident ce que devrait être l'ensemble des mesures possibles, alors que pour d'autres choses cela devient un peu délicat. Mais je tiens à souligner que même dans le cas de l'âge de quelqu'un, c'est beaucoup plus subtil que cela. Dans l'exemple ci-dessus, j'ai supposé qu'il était acceptable de mesurer l'âge en années. Mais si vous êtes un psychologue du développement, c'est beaucoup trop grossier, et vous mesureriez plutôt l'âge en *années et en mois* (si un enfant a 2 ans et 11 mois, cela s'écrit habituellement « 2;11 »). Si vous vous intéressez aux nouveau-nés, vous voudrez peut-être mesurer l'âge en *jours depuis la naissance*, peut-être même en *heures depuis la naissance*. En d'autres termes, la manière dont vous spécifiez les valeurs de mesure autorisées est importante.

En y regardant d'un peu plus près, vous vous rendrez peut-être compte que le concept « d'âge » n'est pas si précis que ça. En général, lorsque nous disons « âge », nous entendons implicitement « le temps écoulé depuis la naissance ». Mais ce n'est pas toujours la meilleure façon de faire. Supposons que vous vous intéressez à la façon dont les nouveau-nés contrôlent les mouvements de leurs yeux. Si vous vous intéressez à des enfants aussi

---

questionnaire sur le genre aujourd'hui, et je préfère les pronoms *" elle/il "* par défaut (c'est une longue histoire) ! En fait, j'ai réfléchi un peu à la façon dont j'allais gérer ça dans le livre. Le livre a une voix d'auteur quelque peu distincte, et j'ai l'impression que ce serait un travail assez différent si j'écrivais tout comme Danielle et mettais à jour tous les pronoms de l'ouvrage. Mais ce serait beaucoup de travail, donc j'ai laissé"Dan" commen étant nom tout au long du livre, d'autant que "Dan" est un excellent surnom pour "Danielle". Ce n'est pas très important. Je voulais seulement le mentionner pour faciliter la vie des lecteurs qui ne savent pas trop comment se référer à moi. Je n'aime toujours pas les anchois quand même:-)

jeunes, vous pourriez aussi vous préoccuper du fait que la « naissance » n'est pas le seul moment significatif dont il faut se soucier. Si Alice naît 3 semaines avant terme et Bianca 1 semaine en retard, est-il vraiment logique de dire qu'elles ont le même âge si on les rencontre « 2 heures après la naissance » ? Dans un sens, oui. Par convention sociale, nous utilisons la naissance comme point de référence pour parler de l'âge dans la vie de tous les jours, car elle définit le temps pendant lequel la personne a évolué comme une entité indépendante dans le monde. Mais d'un point de vue scientifique, ce n'est pas la seule chose qui nous intéresse. Quand nous réfléchissons à la biologie des êtres humains, il est souvent utile de nous considérer comme des organismes qui ont grandi et mûri depuis la conception, et dans cette perspective, Alice et Bianca n'ont pas du tout le même âge. Vous pourriez donc vouloir définir le concept « d'âge » de deux façons différentes : la durée depuis la conception et la durée depuis la naissance. Lorsqu'il s'agit d'adultes, cela ne changera pas grand-chose, mais lorsqu'il s'agit de nouveau-nés, cela pourrait être le cas.

Au-delà de ces questions, il y a la question de la méthodologie. Quelle « méthode de mesure » spécifique allez-vous utiliser pour connaître l'âge de quelqu'un ? Comme auparavant, il y a beaucoup de possibilités différentes :

- Vous pourriez juste demander aux gens « quel âge avez-vous ? » La méthode d'auto-déclaration est rapide, peu coûteuse et facile. Mais cela ne fonctionne qu'avec des personnes assez âgées pour comprendre la question, et certaines personnes mentent sur leur âge.
- Vous pourriez demander à une autorité (par exemple, un parent) « Quel âge a votre enfant ? » Cette méthode est rapide, et quand il s'agit d'enfants, ce n'est pas si difficile que ça puisque le parent est presque toujours là. Cela ne fonctionne pas aussi bien si vous voulez savoir « l'âge depuis la conception », car beaucoup de parents ne peuvent pas dire avec certitude quand la conception a eu lieu. Pour cela, vous pourriez avoir besoin d'une autre autorité (p. ex. un obstétricien).
- Vous pouvez rechercher des documents officiels, par exemple des certificats de naissance ou de décès. C'est une entreprise longue et parfois frustrante, mais elle a son utilité (p. ex. si la personne est maintenant morte).

Opérationnalisation : définir votre mesure

Toutes les idées discutées dans la section précédente ont trait au concept d'**opérationnalisation**. Pour être un peu plus précis sur l'idée, l'opérationnalisation est le processus par lequel nous prenons un concept significatif mais quelque peu vague et le transformons en une mesure précise. Le processus d'opérationnalisation peut impliquer plusieurs choses différentes :

- Soyez précis sur ce que vous essayez de mesurer. Par exemple, « âge » signifie-t-il « temps depuis la naissance » ou « temps depuis la conception » dans le contexte de votre recherche ?
- Déterminer la méthode que vous utiliserez pour la mesurer. Allez-vous utiliser la déclaration pour mesurer l'âge, demander à un parent ou consulter un dossier officiel ? Si vous utilisez la déclaration, comment allez-vous formuler la question ?

- Définir l'ensemble des valeurs admissibles que la mesure peut prendre. Notez que ces valeurs n'ont pas toujours besoin d'être numériques, bien qu'elles le soient souvent. Lorsque l'on mesure l'âge, les valeurs sont numériques, mais nous devons quand même réfléchir soigneusement aux nombres autorisés. Voulons-nous avoir l'âge en années, en années et en mois, en jours ou en heures ? Pour d'autres types de mesures (ex. le sexe), les valeurs ne sont pas numériques. Mais, précédemment, nous devons réfléchir aux valeurs qui sont permises. Si nous demandons aux gens de déclarer eux-mêmes leur sexe, entre quelles options leur permettons-nous de choisir? Est-il suffisant de n'autoriser que les « hommes » ou les « femmes » ? Avez-vous besoin d'une « autre » option ? Ou ne devrions-nous pas donner aux gens des options précises et les laisser plutôt répondre avec leurs propres mots ? Et si vous ouvrez l'ensemble des valeurs possibles pour inclure toutes les réponses verbales, comment interpréterez-vous leurs réponses ?

L'opérationnalisation est une affaire délicate, et il n'y a pas de « manière unique et sérieuse » d'y parvenir. La manière dont vous choisissez d'opérationnaliser le concept informel « d'âge » ou de « sexe » en une mesure formelle dépend de la raison de votre objectif avec cette mesure. Souvent, vous constaterez que les scientifiques qui travaillent dans votre domaine ont des idées assez bien arrêtées sur la façon de procéder. En d'autres termes, l'opérationnalisation doit être envisagée au cas par cas. Néanmoins, bien qu'il y ait beaucoup de questions propres à chaque projet de recherche, il y a certains aspects qui sont assez généraux.

Avant de poursuivre, j'aimerais prendre un moment pour clarifier notre terminologie et, ce faisant, introduire un autre terme. Voici quatre choses différentes qui sont étroitement liées les unes aux autres :

- **Une élaboration théorique**. C'est ce que vous essayez de mesurer, comme « l'âge », « le sexe » ou une « opinion ». Une élaboration théorique ne peut pas être observée directement, et elles sont souvent un peu vagues.
- **Une mesure**. La mesure fait référence à la méthode ou à l'outil que vous utilisez pour faire vos observations. Une question dans une enquête, une observation comportementale ou un scanner du cerveau pourraient toutes vu comme une mesure.
- **Une opérationnalisation**. Le terme « opérationnalisation » fait référence à la connexion logique entre la mesure et l'élaboration théorique, ou au processus par lequel nous essayons de dériver une mesure d'une élaboration théorique.
- **Une variable**. Enfin, un nouveau terme. Une variable est ce que nous obtenons lorsque nous appliquons notre mesure à quelque chose dans le monde. Autrement dit, les variables sont les « données » réelles que nous obtenons dans nos ensembles de données.

En pratique, même les scientifiques ont tendance à ne pas bien faire la distinction entre ces choses, mais il est utile d'essayer d'en comprendre les différences.

Échelles de mesure

Comme l'indique la section précédente, le résultat d'une mesure psychologique s'appelle une variable. Mais toutes les variables ne sont pas du même type qualitatif et il est donc utile de comprendre de quels types il s'agit. Un concept très utile pour distinguer les différents types de variables est ce qu'on appelle les **échelles de mesure**.

Échelle nominale

Une variable d'**échelle nominale** (également appelée **variable catégorielle**) est une variable dans laquelle il n'y a pas de relation particulière entre les différentes possibilités. Pour ce genre de variables, il n'est pas logique de dire que l'une d'entre elles est « plus grande » ou « meilleure » que n'importe quelle autre, et il n'est absolument pas logique de faire la moyenne. L'exemple classique en est la « couleur des yeux ». Les yeux peuvent être bleus, verts ou bruns, entre autres possibilités, mais aucun d'entre eux n'est plus « grand » qu'un autre. Par conséquent, il serait vraiment bizarre de parler d'une « couleur moyenne des yeux ». De même, le sexe est aussi nominal : l'homme n'est ni meilleur ni pire que la femme. Il n'est pas non plus logique d'essayer de parler d'un « genre moyen ». En bref, les variables de l'échelle nominale sont celles pour lesquelles la seule chose que vous pouvez dire sur les différentes possibilités est qu'elles sont différentes./p>

Regardons ça de plus près. Supposons que je fasse des recherches sur la façon dont les gens se rendent au travail et en reviennent. Je pourrais mesurer le type de transport que les gens utilisent pour se rendre au travail. Cette variable « type de transport » pourrait avoir un certain nombre de valeurs possibles, notamment : « train », « bus », « voiture », « vélo ». Pour l'instant, supposons que ces quatre possibilités soient les seules possibles. Alors imaginez que je demande à 100 personnes comment elles sont arrivées à travailler aujourd'hui, avec ce résultat :>/p>

Transport	Nombre de personnes
Train	12
Bus	30
Voiture	48
Bicyclette	10

Alors, quel est le moyen de transport moyen ? Évidemment, la réponse ici est qu'il n'y en a pas. C'est une question idiote. Vous pouvez dire que les voyages en voiture sont la méthode la plus populaire, et les voyages en train sont la méthode la moins populaire, mais c'est à peu près tout. De même, remarquez que l'ordre dans lequel j'énumère les options n'est pas très intéressant. J'aurais pu choisir d'afficher les données comme ci-dessous sans que cela ne change rien.

Transport	Nombre de personnes

(3)	Voiture	48
(1)	Train	12
(4)	Bicyclette	10
(2)	Bus	30

Échelle ordinale

Les variables de l'**échelle ordinale** ont un peu plus de structure que les variables de l'échelle nominale. Une variable d'échelle ordinale est une variable dans laquelle il existe un moyen naturel et significatif d'ordonner les différentes possibilités, mais vous ne pouvez rien faire d'autre. L'exemple habituel d'une variable ordinale est « classement dans une course ». Vous *pouvez* dire que la personne qui a terminé première a été plus rapide que celle qui a terminé deuxième, mais vous *ne* savez *pas de* combien de temps la première devance la seconde. En conséquence, nous savons que 1er> 2ème, et nous savons que 2ème> 3ème, mais la différence entre 1er et 2ème pourrait être beaucoup plus grande que la différence entre 2ème et 3ème.

Voici un exemple plus intéressant sur le plan psychologique. Supposons que je m'intéresse à l'attitude des gens face au changement climatique. Je demander pour cela à des personnes de choisir la proposition (parmi quatre propositions listées) qui correspond le mieux à leurs croyances :

1. Les températures augmentent en raison de l'activité humaine
2. Les températures augmentent, mais nous ne savons pas pourquoi.
3. Les températures augmentent, mais pas à cause des humains.
4. Les températures n'augmentent pas

Remarquez que ces quatre énoncés ont en fait un ordre naturel, du point de vue de leur accord avec l'état actuel de la science «. L'énoncé 1 y correspond exactement, l'énoncé 2 y correspond raisonnablement, l'énoncé 3 n'y correspond pas très bien et l'énoncé 4 est en forte opposition avec l'état actuel de la science. Donc, pour ce qui m'intéresse (la mesure de l'accord des gens avec la science), je peux ordonner les réponses ainsi 1>2>3>4. Puisque cet ordre existe, il serait très bizarre d'énumérer les options comme ceci…

3. Les températures augmentent, mais pas à cause des humains.
4. Les températures augmentent en raison de l'activité humaine
5. Les températures n'augmentent pas
6. Les températures augmentent mais nous ne savons pas pourquoi.

…parce qu'il semble contrevenir à la « structure » naturelle de la question. Supposons que j'ai posé ces questions à 100 personnes et que j'ai obtenu les réponses suivantes :

		Nombre de réponse
(1)	Les températures augmentent en raison de l'activité humaine	51

(2) Les températures augmentent, mais nous ne savons pas pourquoi	20
(3) Les températures augmentent, mais pas à cause des humains	10
(4) Les températures n'augmentent pas	19

En analysant ces données, il semble tout à fait raisonnable d'essayer de regrouper (1), (2) et (3) et de dire que 81 personnes sur 100 étaient disposées à être d'accord au *moins partiellement avec* la science. Et il est *également* tout à fait raisonnable de regrouper (2), (3) et (4) ensemble et de dire que 49 personnes sur 100 ont exprimé au *moins un certain désaccord* avec l'opinion scientifique dominante. Cependant, il serait tout à fait bizarre d'essayer de regrouper (1), (2) et (4) ensemble et de dire que 90 personnes sur 100 ont dit… quoi ? Il n'y a rien de sensé qui vous permette de regrouper ces réponses.

Cela dit, notez que même si nous *pouvons* utiliser l'ordre naturel de ces éléments pour construire des regroupements raisonnables, ce que nous *ne pouvons pas* faire, c'est faire la moyenne. Par exemple, dans mon exemple simple, la réponse « moyenne » à la question est de 1,97. Si vous pouvez me dire ce que cela veut dire, j'aimerais le savoir, parce que ça me semble être du charabia !

Échelle d'intervalle

Contrairement aux variables d'échelle nominale et ordinale, les variables d'échelle d'**intervalle** et de rapport sont des variables pour lesquelles la valeur numérique est réellement significative. Dans le cas des variables d'échelle d'intervalle, les *différences* entre les nombres sont interprétables, mais la variable n'a pas de valeur zéro « naturelle ». La mesure de la température en degrés Celsius est un bon exemple d'une variable d'échelle d'intervalle. Par exemple, s'il faisait 15°C hier et 18°C aujourd'hui, alors la différence de 3°C entre les deux est vraiment significative. De plus, la différence 3˝ est *exactement la même* que la différence 3°C entre7°C et 10°C. En bref, l'addition et la soustraction sont significatives pour les variables de l'échelle d'intervalle.[9]

Notez cependant que le 0°C ne signifie pas « pas de température du tout ». C'est en fait « la température à laquelle l'eau gèle », ce qui est plutôt arbitraire. Par conséquent, il devient inutile d'essayer de multiplier et de diviser les températures. Il est faux de dire que 20°C est deux fois plus chaud que 10°C, tout comme il est bizarre et dénué de sens de prétendre que 20°C est doublement aussi chaud que 10°C.

---

[9] En fait, des lecteurs ayant plus de connaissances en physique que moi m'ont dit que la température n'est pas strictement une échelle d'intervalle, dans le sens où la quantité d'énergie nécessaire pour chauffer quelque chose de 3°C dépend de sa température actuelle. Donc, dans la mesure où les physiciens y attachent de l'importance, la température n'est pas vraiment une échelle d'intervalle. Mais ça constitue quand même un bel exemple, je vais donc ignorer cette petite vérité gênante.

Prenons encore une fois un exemple plus psychologique. Supposons que je m'intéresse à la façon dont les attitudes des étudiants universitaires de première année ont changé au fil du temps. Évidemment, je vais vouloir enregistrer l'année où chaque élève a commencé. Il s'agit d'une variable d'échelle d'intervalle. Un étudiant qui a commencé en 2003 est arrivé 5 ans avant un étudiant qui a commencé en 2008. Cependant, il serait complètement idiot pour moi de diviser 2008 par 2003 et de dire que le deuxième élève a commencé « 1,0024 fois plus tard » que le premier. Cela n'a aucun sens.

Échelle des rapports

Le quatrième et dernier type de variable à prendre en considération est une variable de **l'échelle des ratios**, dans laquelle zéro signifie vraiment zéro, et il est acceptable de multiplier et de diviser. Le temps de réponse (TR) est un bon exemple psychologique d'une variable de l'échelle de rapport. Dans beaucoup de tâches, il est très courant d'enregistrer le temps que quelqu'un prend pour résoudre un problème ou répondre à une question, car c'est un indicateur de la difficulté de la tâche. Supposons qu'Alan prenne 2,3 secondes pour répondre à une question, alors que Ben en prend 3,1 secondes. Comme pour une variable d'échelle d'intervalle, l'addition et la soustraction sont toutes deux significatives ici. Ben a vraiment pris 3,1 – 2,3 = 0,8 secondes de plus qu'Alan. Cependant, notez que la multiplication et la division ont aussi du sens ici aussi : Ben a pris 3,1/2,3 = 1,35 fois plus de temps qu'Alan pour répondre à la question. Et la raison pour laquelle vous pouvez le faire, c'est que pour une variable d'échelle de rapport telle que TR « zéro seconde » signifie vraiment « aucun temps du tout ».

Variables continues et variables discrètes

Il y a un deuxième type de distinction que vous devez connaître, concernant les types de variables que vous pouvez rencontrer. C'est la distinction entre les variables continues et les variables discrètes. La différence entre les deux est la suivante :

Tableau 2-1: La relation entre les échelles de mesure et la distinction discrète/continue. Les cellules marquées d'une croix correspondent à ce qui est possible.

	continue	discrète
nominale		X
ordinale		X
intervalle	X	X
ratio	X	X

- Une **variable continue** est une **variable** dans laquelle, pour deux valeurs auxquelles vous pouvez penser, il est toujours logiquement possible d'avoir une autre valeur entre les deux.
- Une **variable discrète** est, en effet, une variable qui n'est pas continue. Pour une variable discrète, il arrive parfois qu'il n'y ait rien entre deux valeurs.

Ces définitions semblent probablement un peu abstraites, mais elles sont assez simples une fois que vous aurez vu quelques exemples. Par exemple, le temps de réponse est continu. Si Alan prend 3,1 secondes et Ben, 2,3 secondes pour répondre à une question, alors le temps de réponse de Cameron se situera entre les deux s'il a pris 3,0 secondes. Et bien sûr, il serait également possible pour David de prendre 3,031 secondes pour répondre, ce qui signifie que son TR se situerait entre celui de Cameron et celui d'Alan. Et bien qu'en pratique, il est presque impossible de mesurer le TR avec cette précision, c'est certainement possible en principe. Parce que nous pouvons toujours trouver une nouvelle valeur du TR entre deux autres, nous considérons la TR comme une mesure continue.

Les variables discrètes apparaissent lorsque cette règle est violée. Par exemple, les variables de l'échelle nominale sont toujours discrètes. Il n'y a pas un type de transport qui se situe « entre » les trains et les bicyclettes, pas d'un point de vue mathématique strict que comme lorsqu'on dit que 2,3 se situe entre 2 et 3. Le type de transport est donc discret. De même, les valeur d'une échelle ordinale sont toujours discrètes. Bien que la « 2e place » se situe entre la « 1ère place » et la « 3e place », il n'y a rien qui puisse logiquement se situer entre « 1ère place » et « 2e place ». Les variables d'échelle d'intervalle et de ratio ont ces deux caractéristiques. Comme nous l'avons vu plus haut, le temps de réponse (une variable sur une échelle de rapport) est continu. La température en degrés Celsius (une variable sur une échelle d'intervalle) est également continue. Cependant, l'année où vous êtes allé à l'école (une variable sur une échelle d'intervalle) est discrète. Il n'y a pas d'année entre 2002 et 2003. Le nombre de questions que vous obtenez correctement sur un test vrai ou faux (une variable sur une échelle de ratio) est également discret. Puisqu'une question vrai ou faux ne vous permet pas d'être « partiellement correcte », il n'y a rien entre 5/10 et 6/10. Le Tableau 2-1 résume la relation entre les échelles de mesure et la distinction discrète/continuité. Les cellules marquées d'une croix correspondent à ce qui est possible. J'insiste sur ce point, parce que (a) certains manuels se trompent, et (b) les gens disent très souvent des choses comme « variable discrète » quand ils veulent dire « variable nominale ». C'est très regrettable.

Quelques complexités

Bon, je sais que vous allez être choqué d'entendre cela, mais le monde réel est beaucoup plus confus que ne le suggère ce petit schéma de classification. Très peu de variables dans la vie réelle tombent réellement dans ces belles catégories soignées, donc vous devez faire attention à ne pas traiter les échelles de mesure comme s'il s'agissait de règles strictes et rapides. Ça ne marche pas comme ça. Il s'agit de lignes directrices visant à vous aider à réfléchir aux situations dans lesquelles vous devriez traiter différentes variables différemment. Rien de plus.

Prenons donc un exemple classique, peut-être l'exemple classique, d'un outil de mesure psychologique : l'**échelle de Likert**. L'humble échelle de Likert est l'outil de base de toute enquête. Vous en avez vous-même rempli des centaines, voire des milliers, et il y a fort à parier que vous en avez même utilisé un vous-même. Supposons que nous ayons une question d'enquête qui ressemble à ceci :

Lequel des énoncés suivants décrit le mieux votre opinion sur l'affirmation selon laquelle « tous les pirates sont incroyables « ? et les options présentées au participant sont les suivantes :

1. Fortement en désaccord
2. En désaccord
3. Ni d'accord ni en désaccord
4. D'accord
5. Tout à fait d'accord

Cet ensemble d'items est un exemple d'une échelle de Likert à 5 points, dans laquelle on demande aux participants de choisir parmi plusieurs possibilités clairement ordonnées (dans ce cas-ci 5), généralement avec un descripteur verbal donné dans chaque cas. Cependant, il n'est pas nécessaire que tous les éléments soient explicitement décrits. C'est aussi un excellent exemple d'une échelle de Likert à 5 points :

1. Fortement en désaccord
2. Tout à fait d'accord

Les échelles de Likert sont des outils très pratiques, quoique quelque peu limités. La question est de savoir de quel type de variable il s'agit. Ils sont évidemment discrets, puisque vous ne pouvez pas donner une réponse de 2,5. Ce n'est évidemment pas l'échelle nominale, puisque les articles sont commandés ; et ce n'est pas non plus l'échelle des rapports, puisqu'il n'y a pas de zéro naturel.

Mais s'agit-il d'une échelle ordinale ou d'une échelle d'intervalle ? Un argument dit que nous ne pouvons pas vraiment prouver que la différence entre « tout à fait d'accord » et « d'accord » est de la même taille que la différence entre « d'accord » et « ni d'accord ni en désaccord ». En fait, dans la vie de tous les jours, il est assez évident qu'ils ne sont pas du tout les mêmes. Cela suggère donc que nous devrions traiter les échelles de Likert comme des variables ordinales. D'autre part, dans la pratique, la plupart des participants semblent prendre l'ensemble « sur une échelle de 1 à 5 » assez au sérieux, et ils ont tendance à agir comme si les différences entre les cinq options de réponse étaient assez semblables entre elles. Par conséquent, de nombreux chercheurs traitent les données de l'échelle de Likert comme des échelles d'intervalles.[10] Ce n'est pas une échelle d'intervalles, mais dans la pratique, elle est suffisamment proche pour que nous la considérions habituellement comme une **échelle quasi-intervalle**.

Évaluer la fiabilité d'une mesure

A ce stade, nous avons réfléchi un peu à la manière d'opérationnaliser une construction théorique et de créer ainsi une mesure psychologique. Et nous avons vu qu'en appliquant des mesures psychologiques, nous nous retrouvons avec des variables, qui peuvent se présenter sous différentes formes. A ce stade, nous devrions commencer à discuter de la

[10] Ah, la psychologie…jamais une réponse facile à quoi que ce soit !

question évidente : la mesure est-elle bonne ? Nous le ferons en fonction de deux idées connexes : la *fiabilité* et la *validité*. En termes simples, la fiabilité d'une mesure vous indique avec quelle *précision* vous mesurez quelque chose, alors que la validité d'une mesure vous indique à quel point la mesure est *précise*. Dans cette section, je parlerai de fiabilité ; nous parlerons de validité dans la section 2.6.

La fiabilité est en fait un concept très simple. Il s'agit de la répétabilité ou de la cohérence de votre mesure. La mesure de mon poids à l'aide d'un « pèse-personne » est très fiable. Si je monte et descends de la balance encore et encore, ça me donnera toujours la même réponse. Mesurer mon intelligence à l'aide de « demander à ma mère » n'est pas très fiable. Certains jours, elle me dit que je suis un peu épais, et d'autres jours, elle me dit que je suis un idiot complet. Notez que ce concept de fiabilité est différent de la question de savoir si les mesures sont correctes (l'exactitude d'une mesure est liée à sa validité). Si je tiens un sac de pommes de terre lorsque je monte et descends de la balance de la salle de bain, la mesure sera toujours fiable : elle me donnera toujours la même réponse. Cependant, cette réponse très fiable ne correspond pas du tout à mon poids réel, donc c'est faux. En termes techniques, il s'agit d'une mesure *fiable mais non valable.* De même, bien que l'estimation de mon intelligence faite par ma mère ne soit pas très fiable, elle a peut-être raison. Peut-être que je ne suis tout simplement pas très intelligent, et alors que son estimation de mon intelligence fluctue énormément d'un jour à l'autre, c'est fondamentalement juste. Ce serait une mesure *peu fiable mais valable.* Bien sûr, si les estimations de ma mère ne sont pas assez fiables, il sera très difficile de déterminer laquelle de ses nombreuses affirmations sur mon intelligence est en fait la bonne. Dans une certaine mesure, donc, une mesure très peu fiable tend à finir par être invalide pour des raisons pratiques, à tel point que beaucoup de gens diraient que la fiabilité est nécessaire (mais pas suffisante) pour assurer la validité.

Ok, maintenant que nous sommes clairs sur la distinction entre fiabilité et validité, réfléchissons aux différentes façons dont nous pourrions mesurer la fiabilité :

- **Fiabilité test-retest**. Il s'agit de l'uniformité dans le temps. Si nous répétons la mesure à une date ultérieure, obtenons-nous la même réponse ?
- **Fiabilité entre évaluateurs**. Cela concerne l'uniformité entre les personnes. Si quelqu'un d'autre répète la mesure (p. ex. quelqu'un d'autre évalue mon intelligence), est-ce qu'il produira la même réponse ?
- **Fiabilité des formes parallèles**. Il s'agit de la cohérence entre les mesures théoriquement équivalentes. Si j'utilise un autre ensemble de pèse-personnes pour mesurer mon poids, est-ce que cela donne la même réponse ?
- **Fiabilité de la cohérence interne**. Si une mesure est construite à partir d'un grand nombre de parties différentes qui remplissent des fonctions similaires (p. ex. un résultat de questionnaire de personnalité est additionné à travers plusieurs questions), les parties individuelles ont tendance à donner des réponses similaires. Nous examinerons cette forme particulière de fiabilité plus loin dans le livre, à la section 15.5.

Il n'est pas nécessaire que toutes les mesures possèdent toutes les formes de fiabilité. Par exemple, l'évaluation de l'éducation peut être considérée comme une forme de mesure. L'une des matières que j'enseigne, la *science cognitive computationnelle*, a une structure

d'évaluation qui comporte un volet recherche et un volet examen (plus d'autres choses). La composante de l'examen est *destinée* à mesurer quelque chose de différent de la composante de recherche, de sorte que l'évaluation dans son ensemble a une faible cohérence interne. Cependant, l'examen comporte plusieurs questions qui visent à mesurer (approximativement) les mêmes choses, et celles-ci ont tendance à produire des résultats similaires. L'examen en lui-même a donc une consistance interne assez élevée. Ce qui est comme il se doit. Vous ne devriez exiger la fiabilité que dans les situations où vous voulez mesurer la même chose !

Le « rôle » des variables : prédicteurs et résultats

J'ai un dernier élément de terminologie que je dois vous expliquer avant de m'éloigner des variables. Normalement, lorsque nous faisons de la recherche, nous nous retrouvons avec un grand nombre de variables différentes. Ensuite, lorsque nous analysons nos données, nous essayons habituellement d'expliquer certaines des variables en fonction d'autres variables. Il est important de distinguer les deux rôles « chose qui explique » et « chose qui est expliquée ». Soyons clairs sur ce point maintenant. Tout d'abord, autant s'habituer à l'idée d'utiliser des symboles mathématiques pour décrire des variables, puisque cela va se reproduire à l'infini. Désignons la variable « à expliquer » Y, et les variables « explicative » comme X_1, X_2, etc.

Lorsque nous faisons une analyse, nous avons des noms différents pour X et Y, car ils jouent des rôles différents dans l'analyse. Les noms classiques de ces rôles sont **variable indépendante** (VI) et **variable dépendante** (VD). La VI est la variable que vous utilisez pour expliquer (c.-à-d. X) et la DV est la variable expliquée (c.-à-d. Y). La logique derrière ces noms est la suivante : s'il y a vraiment une relation entre X et Y, alors nous pouvons dire que Y dépend de X, et si nous avons conçu notre étude « correctement », alors X ne dépend de rien d'autre. Cependant, je trouve personnellement ces noms horribles. Elles sont difficiles à retenir et elles sont très trompeuses parce que (a) la VI n'est jamais « indépendante de tout le reste », et (b) s'il n'y a pas de relation, alors la DV ne dépend pas réellement de la VI. Et en fait, parce que je ne suis pas la seule personne qui pense que IV et DV ne sont que des noms affreux, il y a un certain nombre d'alternatives que je trouve plus attirantes. Les termes que j'utiliserai dans ce livre sont des **prédicteurs** et des **résultats**. L'idée ici est que ce que vous essayez de faire est d'utiliser X (les prédicteurs) pour faire des suppositions sur Y (les résultats).[11] Ce point est résumé dans le Tableau 2-2.

Tableau 2-2 : La terminologie utilisée pour distinguer les différents rôles qu'une variable peut jouer dans l'analyse d'un ensemble de données. Notez que ce livre aura tendance à éviter la terminologie classique en faveur des noms plus récents.

---

[11] L'ennui, c'est qu'il y a beaucoup de noms différents qui sont utilisés. Je ne les énumérerai pas tous - cela ne servirait à rien de le faire - si ce n'est de noter que "variable réponse" est parfois utilisée là où j'ai utilisé "résultat". Ce genre de confusion terminologique est très courant, je le crains.

rôle de la variable	nom classique	nom moderne
«A expliquer»	Variable dépendante (DV)	Résultat
«Explicative»	Variable indépendante (IV)	Prédicteur

Recherche expérimentale et non expérimentale

L'une des grandes distinctions que vous devez connaître est la distinction entre « recherche expérimentale » et « recherche non expérimentale ». Lorsque nous faisons cette distinction, nous parlons en fait du degré de contrôle que le chercheur exerce sur les personnes et les événements de l'étude.

Recherche expérimentale

La principale caractéristique de la **recherche expérimentale** est que le chercheur contrôle tous les aspects de l'étude, en particulier ce que les participants vivent pendant l'étude. En particulier, le chercheur manipule ou fait varier les variables prédicteurs (VI), mais laisse la variable résultat (VD) varier naturellement. L'idée ici est de faire varier délibérément les prédicteurs (IV) pour voir s'ils ont des effets causaux sur les résultats. De plus, afin de s'assurer qu'il n'y a aucune possibilité que quelque chose d'autre que les variables prédictrices cause les résultats, tout le reste est maintenu constant ou « équilibré » d'une autre façon, pour s'assurer qu'ils n'ont aucun effet sur les résultats. En pratique, il est presque impossible de *penser* à tout ce qui pourrait avoir une influence sur le résultat d'une expérience, et encore moins de la maintenir constante. La solution standard est **l'aléatorisation**. C'est-à-dire que nous assignons au hasard des personnes à des groupes différents, puis donnons à chaque groupe un traitement différent (c.-à-d., leur assignons des valeurs différentes des variables prédictrices). Nous parlerons plus en détail de l'aléatorisation plus tard, mais pour l'instant, il suffit de dire que ce que fait l'aléatorisation est de minimiser (mais pas d'éliminer) la possibilité qu'il y ait une différence systématique entre les groupes.

Prenons un exemple très simple, complètement irréaliste et tout à fait contraire à l'éthique. Supposons que vous vouliez savoir si le tabagisme cause le cancer du poumon. Une façon d'y parvenir serait de trouver des fumeurs et des non-fumeurs et de vérifier si les fumeurs ont un taux plus élevé de cancer du poumon. Ce *n'*est *pas* une expérience correcte, puisque le chercheur n'a pas suffisamment de contrôle sur qui est et qui n'est pas un fumeur. Et c'est vraiment important. Par exemple, il se peut que les gens qui choisissent de fumer des cigarettes aient aussi tendance à avoir une mauvaise alimentation, ou peut-être qu'ils ont tendance à travailler dans les mines d'amiante, ou bien d'autres choses. Le fait est que les groupes (fumeurs et non-fumeurs) diffèrent en fait sur beaucoup de choses, et pas *seulement* sur le tabagisme. Il se peut donc que l'incidence plus élevée de cancer du poumon chez les fumeurs soit causée par autre chose, et non par le tabagisme lui-même. En termes techniques, ces autres choses (par exemple, l'alimentation) sont appelées « facteurs de confusion «, et nous en parlerons dans un instant.

En attendant, considérons à quoi pourrait ressembler une expérience correcte. Rappelez-vous que nous craignions que les fumeurs et les non-fumeurs puissent différer à bien des égards. La solution, tant que vous n'avez pas d'éthique, est de *contrôler* qui fume et qui ne fume pas. Plus précisément, si nous divisons au hasard les jeunes non-fumeurs en deux groupes et forçons la moitié d'entre eux à devenir fumeurs, il est très peu probable que les groupes diffèrent sur un autre point que le fait que la moitié d'entre eux fument. De cette façon, si notre groupe de fumeurs a un taux de cancer plus élevé que le groupe de non-fumeurs, nous pouvons être assez confiants que (a) le tabagisme cause le cancer et (b) nous sommes des meurtriers.

Recherche non expérimentale

La recherche non expérimentale est un terme large qui couvre « toute étude dans laquelle le chercheur n'a pas autant de contrôle que dans une expérience ». Évidemment, le contrôle est quelque chose que les scientifiques aiment avoir, mais comme l'exemple précédent l'illustre, il y a beaucoup de situations dans lesquelles vous ne pouvez ou ne devriez pas essayer d'obtenir ce contrôle. Puisqu'il est tout à fait contraire à l'éthique (et presque certainement criminel) de forcer les gens à fumer pour savoir s'ils ont le cancer, c'est un bon exemple d'une situation dans laquelle vous ne devriez vraiment pas essayer d'obtenir un contrôle expérimental. Mais il y a aussi d'autres raisons. Même en laissant de côté les questions éthiques, notre « expérience du tabagisme » soulève d'autres problèmes. Par exemple, lorsque j'ai suggéré de « forcer » la moitié des gens à devenir fumeurs, je parlais de *commencer* avec un échantillon de non-fumeurs, puis de les forcer à devenir fumeurs. Bien que cela ressemble au genre de plan expérimental solide et maléfique qu'un savant fou adorerait, ce n'est peut-être pas une façon très saine d'étudier l'effet dans le monde réel. Supposons, par exemple, que le tabagisme ne cause le cancer du poumon que lorsque les gens ont une mauvaise alimentation et que les gens qui fument normalement ont tendance à avoir une mauvaise alimentation. Cependant, comme les « fumeurs » de notre expérience ne sont pas des fumeurs « naturels » (c.-à-d. que nous avons forcé les non-fumeurs à devenir des fumeurs, mais qu'ils n'ont pas adopté toutes les autres caractéristiques normales et réelles que les fumeurs pourraient avoir tendance à avoir), ils ont probablement une meilleure alimentation. Ainsi, dans cet exemple stupide, ils n'auraient pas de cancer du poumon et notre expérience échouera, parce qu'elle viole la structure du monde « naturel » (le nom technique pour ceci est un résultat « artefact »).

Une distinction qu'il convient de faire entre deux types de recherche non expérimentale est la différence entre la **recherche quasi-expérimentale** et les **études de cas**. L'exemple dont j'ai parlé plus tôt, dans lequel nous voulions examiner l'incidence du cancer du poumon chez les fumeurs et les non-fumeurs sans essayer de contrôler qui fume et qui ne fume pas, est un modèle quasi expérimental. C'est-à-dire qu'il est similaire à une expérience, mais nous ne contrôlons pas les prédicteurs (VI). Nous pouvons encore utiliser les statistiques pour analyser les résultats, mais nous devons être beaucoup plus prudents et circonspects.

L'approche alternative, les études de cas, vise à fournir une description très détaillée d'un ou de quelques cas. En général, on ne peut pas utiliser les statistiques pour analyser les résultats des études de cas et il est généralement très difficile de tirer des conclusions

générales sur « les gens en général » à partir de quelques exemples isolés. Toutefois, les études de cas sont très utiles dans certaines situations. Tout d'abord, il y a des situations où vous n'avez pas d'alternative. C'est le cas de la neuropsychologie en particulier. Parfois, vous ne pouvez tout simplement pas trouver beaucoup de personnes atteintes de lésions cérébrales dans une région précise du cerveau, alors la seule chose que vous pouvez faire est de décrire les cas que vous avez avec autant de détails et avec autant de soin que possible. Cependant, les études de cas présentent aussi de véritables avantages. Comme vous n'avez pas autant de personnes à étudier, vous avez la possibilité d'investir beaucoup de temps et d'efforts pour essayer de comprendre les facteurs spécifiques en jeu dans chaque cas. C'est très important de le faire. Par conséquent, les études de cas peuvent compléter les approches plus axées sur les statistiques que l'on trouve dans les plans expérimentaux et quasi expérimentaux. Nous ne parlerons pas beaucoup des études de cas dans ce livre, mais elles sont néanmoins des outils très précieux !

Évaluer la validité d'une étude

Plus que toute autre chose, un scientifique veut que sa recherche soit « valide ». L'idée conceptuelle derrière la **validité** est très simple. Pouvez-vous faire confiance aux résultats de votre étude ? Si ce n'est pas le cas, l'étude n'est pas valide. Cependant, bien qu'il soit facile à énoncer, dans la pratique, il est beaucoup plus difficile de vérifier la validité qu'il ne l'est de vérifier la fiabilité. Et en toute honnêteté, il n'y a pas de notion précise et clairement acceptée de ce qu'est réellement la validité. En fait, il existe de nombreux types de validité différents, qui soulèvent chacun des questions qui lui sont propres. Et toutes les formes de validité ne sont pas pertinentes pour toutes les études. Je vais présenter cinq types de validité :

- Validité interne
- Validité externe
- Validité de construction
- Validité apparente
- Validité écologique

Tout d'abord, un guide rapide sur ce qui compte ici. (1) La validité interne et externe est la plus importante, car elle est directement liée à la question fondamentale de savoir si votre étude fonctionne réellement. (2) La validité de construction demande si vous mesurez ce que vous pensez mesurer. (3) La validité apparente n'est pas très importante sauf dans la mesure où vous vous souciez des « apparences ». (4) La validité écologique est un cas particulier de validité apparente qui correspond à un type d'apparence qui peut vous intéresser beaucoup.

Validité interne

La validité interne se réfère à la mesure dans laquelle vous êtes capable de tirer les conclusions correctes sur les relations causales entre les variables. On l'appelle « interne » parce qu'il fait référence aux relations entre les choses « à l'intérieur » de l'étude. Illustrons le concept par un exemple simple. Supposons que vous souhaitiez savoir si une formation universitaire vous permet d'écrire mieux. Pour ce faire, vous formez un groupe d'étudiants

de première année, leur demandez d'écrire un essai de 1000 mots et comptez le nombre de fautes d'orthographe et de grammaire qu'ils font. Ensuite, vous trouvez des étudiants de troisième année, qui ont de toute évidence fait plus d'études universitaires que les étudiants de première année, et vous répétez l'exercice. Et supposons qu'il s'avère que les étudiants de troisième année produisent moins d'erreurs. Vous pensez pouvoir conclure qu'une formation universitaire améliore les compétences en rédaction ?

Sauf que le gros problème de cette expérience est que les étudiants de troisième année sont plus âgés et qu'ils ont plus d'expérience dans l'écriture. Il est donc difficile de savoir avec certitude quelle est la relation de cause à effet. Les personnes âgées écrivent-elles mieux ? Ou est-ce parce qu'ils ont plus d'expérience en écriture ? Ou parce qu'ils ont fait plus d'études ? Laquelle de ces raisons est la véritable *cause* de la performance supérieure de la troisième année ? L'âge ? Expérience ? L'éducation ? Vous ne pouvez ne peux pas le dire. C'est un exemple d'échec de validité interne, parce que votre étude ne distingue pas correctement les relations *causales* entre les différentes variables.

Validité externe

La validité externe se rapporte à la **généralisabilité** ou à l'**applicabilité** de vos conclusions. C'est-à-dire, dans quelle mesure vous attendez-vous à voir le même schéma de résultats dans la « vie réelle » que celui que vous avez vu dans votre étude. Pour être un peu plus précis, toute étude que vous ferez en psychologie comportera un ensemble assez précis de questions ou de tâches, se déroulera dans un environnement particulier et impliquera des participants provenant d'un sous-groupe particulier (malheureusement, ce sont souvent des étudiants d'université !). Donc, s'il s'avère que les résultats ne peuvent pas être généralisés ou ne s'appliquent pas aux personnes et aux situations au-delà de celles que vous avez étudiées, alors vous avez un manque de validité externe.

L'exemple classique de cette question est le fait qu'une très grande proportion des études en psychologie font appel à des étudiants de premier cycle en psychologie comme participants. Évidemment, cependant, les chercheurs ne se soucient pas *seulement* des étudiants en psychologie. Ils se soucient des gens en général. Par conséquent, une étude qui utilise uniquement des étudiants en psychologie comme participants comporte toujours le risque de manquer de validité externe. Autrement dit, s'il y a quelque chose de « spécial » chez les étudiants en psychologie qui les rend différents du reste de la population à certains égards, alors nous pourrions commencer à nous inquiéter d'un manque de validité externe.

Cela dit, il est absolument essentiel de réaliser qu'une étude qui n'utilise que des étudiants en psychologie n'a pas nécessairement un problème de validité externe. J'en reparlerai plus tard, mais c'est une erreur tellement courante que je vais en parler ici. La validité externe d'une étude est menacée par le choix de la population si (a) la population à partir de laquelle vous échantillonnez vos participants est très restreinte (p. ex. les étudiants en psychologie) et (b) la population restreinte que vous avez échantillonnée est systématiquement différente de la population générale à *certains égards qui est pertinente au phénomène psychologique que vous voulez étudier*. La partie en italique est la partie que beaucoup de gens oublient. Il est vrai que les étudiants de premier cycle en psychologie diffèrent de la population générale à bien des égards, de sorte qu'une étude qui utilise

uniquement des étudiants en psychologie *peut* avoir des problèmes de validité externe. Cependant, si ces différences ne sont pas très pertinentes par rapport au phénomène que vous étudiez, il n'y a pas de quoi s'inquiéter. Pour rendre cela un peu plus concret, voici deux exemples extrêmes :

- Vous voulez mesurer « les attitudes du grand public envers la psychothérapie », mais tous vos participants sont des étudiants en psychologie. Cette étude aurait presque certainement un problème de validité externe.
- Vous voulez mesurer l'efficacité d'une illusion visuelle et vos participants sont tous des étudiants en psychologie. Il est peu probable que cette étude ait un problème de validité externe.

Après avoir passé les deux derniers paragraphes à se concentrer sur le choix des participants, puisque c'est une question importante qui préoccupe le plus tout le monde, il est bon de se rappeler que la validité externe est un concept plus large. Voici également des exemples de choses qui pourraient menacer la validité externe, selon le type d'étude que vous effectuez :

- Les gens pourraient répondre à un « questionnaire psychologique » d'une manière qui ne reflète pas ce qu'ils feraient dans la vie réelle.
- Votre expérience de laboratoire sur (par exemple) « l'apprentissage humain » a une structure différente de celle des problèmes d'apprentissage auxquels les gens font face dans la vie réelle.

Construire la validité

La validité de construction consiste essentiellement à se demander si vous mesurez ce que vous voulez mesurer. Une mesure a une bonne validité de construction si elle mesure réellement le bonne construit théorique, et une mauvaise validité de construction si ce n'est pas le cas. Pour donner un exemple très simple (quoique ridicule), supposons que j'essaie de mesurer les taux avec lesquels les étudiants universitaires trichent à leurs examens. Une façon d'essayer de le mesurer est de demander aux élèves tricheurs de se lever dans l'amphithéâtre pour que je puisse les compter. Quand je fais cela avec une classe de 300 élèves, 0 personnes prétendent être des tricheurs. J'en conclus donc que la proportion de tricheurs dans ma classe est de 0%. Il est clair que c'est un peu ridicule. Mais il ne s'agit pas ici d'un exemple méthodologique très profond, mais plutôt d'expliquer ce qu'est la validité conceptuelle. Le problème avec ma mesure, c'est que pendant que j'*essaie de* mesurer « la proportion de personnes qui trichent », ce que je mesure en fait est « la proportion de personnes assez stupides pour avouer qu'elles trichent, ou assez dérangé pour prétendre le faire ». Évidemment, ce n'est pas la même chose ! Mon étude a donc mal tourné, parce que ma mesure a une très mauvaise validité conceptuelle.

Validité apparente

La validité apparente fait simplement référence au fait qu'une mesure « ressemble » ou non à ce qu'elle est censée faire. Si je conçois un test d'intelligence, et que quelqu'un le regarde et dit « non, ce test ne mesure pas l'intelligence », alors la mesure manque de

validité apparente. C'est aussi simple que ça. Évidemment, la validité apparente n'est pas très importante d'un point de vue purement scientifique. Après tout, ce qui nous importe, c'est de savoir si la mesure fait ce qu'elle est censée faire ou non, et non si elle *semble* faire ce qu'elle est censée faire. Par conséquent, nous ne nous soucions généralement pas beaucoup de la validité apparente. Cela dit, le concept de validité apparente sert trois objectifs pragmatiques utiles :

- Parfois, un scientifique expérimenté aura l'intuition qu'une mesure particulière ne fonctionnera pas. Bien que ce genre d'intuition n'ai pas de valeur probante, elle vaut souvent la peine d'y prêter attention. Parce que souvent, les gens ont des connaissances qu'ils ne peuvent pas verbaliser, cela vaut la peine de s'en préoccuper, même si vous ne pouvez pas dire tout à fait pourquoi. En d'autres termes, lorsque quelqu'un en qui vous avez confiance critique la validité apparente de votre étude, il vaut la peine de prendre le temps de réfléchir plus attentivement à votre étude pour voir si vous avez pensé aux raisons pour lesquelles elle pourrait ne pas être valide. Mais rappelez-vous, si vous ne trouvez aucune raison de vous inquiéter, alors vous ne devriez probablement pas vous inquiéter. Après tout, la validité apparente n'a pas vraiment d'importance.
- Souvent (très souvent), des personnes complètement mal informées auront aussi l'intuition que votre recherche ne vaut rien. Ils la critiqueront sur Internet ou ailleurs. En y regardant de plus près, vous remarquerez peut-être que ces critiques se concentrent en fait entièrement sur ce à quoi l'étude « ressemble », mais pas sur ces fondements. Le concept de validité apparente est utile pour expliquer doucement aux gens qu'ils ont besoin d'étayer davantage leurs arguments.
- Pour en revenir à ce dernier point, si les croyances de personnes non formées sont critiques (par exemple, c'est souvent le cas pour la recherche appliquée où l'on veut convaincre les décideurs d'une chose ou d'une autre), alors il faut se soucier de la validité apparente. Tout simplement parce que, que vous le vouliez ou non, beaucoup de gens utiliseront la validité apparente comme un substitut de la validité réelle. Si vous voulez que le gouvernement modifie une loi pour des raisons psychologiques scientifiques, peu importe la qualité de vos études. S'ils manquent de validité apparente, vous constaterez que les politiciens vous ignoreront. Bien sûr, c'est un peu injuste que la politique dépende souvent davantage de l'apparence que des faits, mais c'est ainsi.

Validité écologique

La validité écologique est une notion différente de la validité, qui est semblable à la validité externe, mais moins importante. L'idée est que, pour être valable du point de vue écologique, l'ensemble de l'étude doit se rapprocher le plus possible du scénario du monde réel qui est à l'étude. Dans un sens, la validité écologique est une sorte de validité apparente. Il s'agit surtout de savoir si l'étude « semble » correcte, mais avec un peu plus de rigueur. Pour être valable du point de vue écologique, l'étude doit avoir un aspect assez précis. L'idée sous-jacente est l'intuition qu'une étude qui est écologiquement valide est plus susceptible d'avoir une validité externe. Ce n'est pas une garantie, bien sûr. Mais ce qu'il y a de bien avec la validité écologique, c'est qu'il est beaucoup plus facile de vérifier si

une étude est valide sur le plan écologique que de vérifier si une étude est valide sur le plan externe. Un exemple simple serait les études d'identification par témoin oculaire. La plupart de ces études ont tendance à se faire dans un cadre universitaire, souvent avec un tableau assez simple de visages à regarder, plutôt qu'en ligne. Il s'écoule généralement moins de temps entre le moment où l'on voit le « criminel » et celui où l'on lui demande d'identifier le suspect dans la « file d'attente ». Le « crime » n'est pas réel, donc il n'y a aucune chance que le témoin ait peur, et il n'y a aucun policier présent, donc il n'y a pas autant de chances de se sentir sous pression. Toutes ces choses montrent que l'étude manque *définitivement* de validité écologique. Elles peuvent (mais pas nécessairement) signifier qu'elle manque également de validité externe.

Confusion, artefacts et autres menaces à la validité

Si nous examinons la question de la validité de la manière la plus générale, les deux plus grandes préoccupations que nous avons sont les *facteurs de confusion* et les *artefacts*. Ces deux termes sont définis de la manière suivante :

- **Confusion** : Un facteur de confusion est une variable supplémentaire, souvent non mesurée,[12] qui s'avère être liée à la fois aux prédicteurs et au résultat. L'existence de variables de confusion menace la validité interne de l'étude parce qu'on ne peut pas dire si c'est le prédicteur qui cause le résultat ou si la variable confusionnelle en est la cause.
- **Artefact** : Un résultat est dit « artefactuel » s'il ne tient que dans la situation spéciale que vous avez testée dans votre étude. La possibilité que votre résultat soit un artefact est une menace à votre validité externe, parce qu'elle soulève la possibilité que vous ne puissiez pas généraliser ou appliquer vos résultats à la population réelle à laquelle vous tenez.

En règle générale, les facteurs de confusion sont plus préoccupants pour les études non expérimentales, précisément parce qu'il ne s'agit pas d'expériences à proprement parler. Par définition, vous laissez beaucoup de choses non contrôlées, alors il y a beaucoup de place pour les facteurs de confusion dans votre étude. La recherche expérimentale tend à être beaucoup moins vulnérable aux facteurs de confusion. Plus vous avez de contrôle sur ce qui se passe pendant l'étude, plus vous pouvez empêcher les facteurs de confusion d'affecter les résultats. Dans le cas de la répartition aléatoire, par exemple, les variables de confusion sont réparties de façon aléatoire et uniforme entre différents groupes.

---

[12] La raison pour laquelle je dis que ce n'est pas mesuré, c'est que si vous l'*avez* mesuré, vous pouvez utiliser des astuces statistiques fantaisistes pour faire face au facteur de confusion. En raison de l'existence de ces solutions statistiques au problème des variables de confusion, nous faisons souvent référence à un facteur de confusion que nous avons mesuré et traité comme une *covariable*. Traiter des covariables est un sujet plus avancé, mais j'ai pensé le mentionner en passant, car c'est plutôt réconfortant de savoir au moins que ce genre de chose existe.

Cependant, il y a toujours des fluctuations et des revirements et lorsque nous commençons à penser aux artefacts plutôt qu'aux facteurs de confusion, la situation est souvent inversée. Dans la plupart des cas, les résultats artefactuels ont tendance à être une préoccupation pour les études expérimentales plutôt que pour les études non expérimentales. Pour s'en rendre compte, il est utile de considérer que la raison pour laquelle beaucoup d'études ne sont pas expérimentales est précisément parce que le chercheur essaie d'examiner le comportement humain dans un contexte plus écologique. En travaillant dans un contexte plus réel, vous perdez le contrôle expérimental (ce qui vous rend vulnérable aux facteurs de confusion), mais parce que vous avez tendance à étudier la psychologie humaine « dans la vie réelle », vous réduisez les chances d'obtenir un résultat artefactuel. En d'autres termes, lorsque vous sortez la psychologie de la vie réelle et l'amenez au laboratoire (ce que nous devons habituellement faire pour obtenir notre contrôle expérimental), vous courez toujours le risque d'étudier accidentellement quelque chose de différent de ce que vous voulez étudier.

Attention cependant, ce qui précède n'est donné qu'à titre indicatif. Il est absolument possible d'avoir des facteurs de confusion dans une expérience, et d'obtenir des résultats artefactuels avec des études non expérimentales. Cela peut se produire pour toutes sortes de raisons, dont la moindre n'est pas l'erreur de l'expérimentateur ou du chercheur. Dans la pratique, il est très difficile de tout prévoir à l'avance et même les très bons chercheurs font des erreurs.

Bien que d'un certain point de vue, presque toute menace à la validité peut être qualifiée de facteur de confusion ou d'artefact, ce sont des concepts assez vagues. Regardons donc de plus près certains des exemples les plus courants.

Effets de l'histoire

Les effets historiques renvoient à la possibilité que des événements particuliers susceptibles d'influer sur la mesure des résultats se produisent au cours de l'étude. Par exemple, il peut se passer quelque chose entre un prétest et un post-test ou entre le participant 23 et le participant 24. Il se peut aussi que vous considériez les résultats d'une étude plus ancienne qui était parfaitement valable à l'époque, mais que le monde a suffisamment changé depuis lors pour que les conclusions ne soient plus dignes de foi. Voici des exemples qui pourraient être considérées comme des effets de l'histoire :

Vous vous intéressez à la façon dont les gens perçoivent le risque et l'incertitude. Vous avez commencé votre collecte de données en décembre

- Mais il faut du temps pour trouver des participants et recueillir des données, alors vous trouverez encore de nouveaux sujets en février 2011. Malheureusement pour vous (et encore plus malheureusement pour les autres), les inondations du Queensland se sont produites en janvier 2011, causant des milliards de dollars de dégâts et tuant de nombreuses personnes. Comme on pouvait s'y attendre, les personnes testées en février 2011 expriment des croyances très différentes de celles des personnes testées en décembre 2010 en matière de gestion du risque. Lequel d'entre eux (le cas échéant) reflète les « vraies » croyances des participants ? Je pense que la réponse est probablement les deux. Les inondations du Queensland ont véritablement changé les

croyances du public australien, mais peut-être seulement temporairement. L'essentiel ici, c'est que « l'histoire » des personnes testées en février est très différente de celle des personnes testées en décembre.

- Vous testez les effets psychologiques d'un nouveau médicament contre l'anxiété. Il faut donc mesurer l'anxiété avant d'administrer le médicament (p. ex. par auto déclaration et en prenant des mesures physiologiques). Ensuite, vous administrez le médicament, puis vous prenez les mêmes mesures. Pendant ce temps, parce que votre laboratoire est à Los Angeles, il y a un tremblement de terre qui augmente l'anxiété des participants.

Effets de maturation

Comme dans le cas des effets historiques, les **effets de maturation** sont fondamentalement liés au changement au fil du temps. Cependant, les effets de maturation ne sont pas une réponse à des événements spécifiques. Elles sont plutôt liées à la façon dont les gens changent d'eux-mêmes au fil du temps. On vieillit, on se fatigue, on s'ennuie, etc. Voici quelques exemples d'effets de maturation :

- Lorsque vous faites de la recherche en psychologie du développement, vous devez être conscient que les enfants grandissent assez rapidement. Supposons donc que vous souhaitiez savoir si une approche éducative aide à améliorer la taille du vocabulaire chez les enfants de 3 ans. Il faut avoir à l'esprit que la taille du vocabulaire des enfants de cet âge augmente spontanément à un rythme incroyablement rapide (plusieurs mots par jour). Si vous concevez votre étude sans tenir compte de cet effet de maturation, vous ne serez pas en mesure de dire si votre approche éducative fonctionne.

- Lorsqu'on fait une très longue expérience en laboratoire (disons 3 heures), il est très probable que les gens commenceront à s'ennuyer et à se fatiguer, et que cet effet de maturation entraînera une baisse de performance indépendamment du contenu l'expérience.

Effets de tests répétés

Un type important d'effet historique est l'effet des **essais répétés**. Supposons que je veuille prendre deux mesures d'une dimension psychologique (p. ex., l'anxiété). Je pourrai chercher à savoir si la première mesure a un effet sur la seconde. En d'autres termes, s'il s'agit d'un effet historique dans lequel « l'événement » qui influence la deuxième mesure est la première mesure elle-même ! Ce n'est pas du tout rare. En voici quelques exemples :

- *Apprentissage et pratique* : par exemple, La mesure de « l'intelligence » au temps 2 peut sembler augmenter par rapport au temps 1 parce que les participants ont appris les règles générales sur la résolution les questions de type « test d'intelligence » au cours de la première séance de test.
- *Familiarité avec la situation du test* : par exemple, si les gens sont nerveux au moment 1, cela peut faire baisser la performance. Mais après avoir participer à la première

situation de test, ils pourraient être plus confiant parce qu'ils ont vu à quoi ressemble le test.

- *Changements auxiliaires causés par les tests* : par exemple, si un questionnaire d'évaluation de l'humeur est ennuyeux, l'évaluation de l'humeur au moment de la mesure 2 est plus susceptible de s'ennuyer précisément en raison de la mesure ennuyeuse effectuée au moment 1.

Biais de sélection

Le biais de sélection est un terme assez large. Supposons que vous menez une expérience avec deux groupes de participants où chaque groupe reçoit un « traitement « différent, et que vous voulez voir si les différents traitements donnent des résultats différents. Supposons toutefois que, malgré tous vos efforts, vous vous retrouvez avec un déséquilibre entre les sexes dans tous les groupes (disons, le groupe A compte 80 % de femmes et le groupe B 50 % de femmes). On pourrait croire que ça n'arrivera jamais, mais croyez-moi, c'est possible. Il s'agit d'un exemple de biais de sélection, dans lequel les personnes « sélectionnées » dans les deux groupes ont des caractéristiques différentes. Si l'une ou l'autre de ces caractéristiques s'avère pertinente (par exemple, votre traitement est plus efficace sur les femmes que sur les hommes), vous êtes dans une situation très difficile.

Attrition différentielle

Lorsqu'on pense aux effets de l'attrition, il est parfois utile de faire la distinction entre deux types différents. La première est l'**attrition homogène**, dans laquelle l'effet d'attrition est le même pour tous les groupes, traitements ou conditions. Dans l'exemple que j'ai donné ci-dessus, l'attrition serait homogène si (et seulement si) les participants facilement ennuyés abandonnent toutes les conditions de mon expérience à peu près au même rythme. En général, le principal effet de l'attrition homogène est de rendre votre échantillon non représentatif. Ainsi, la plus grande inquiétude que vous aurez est que la généralisabilité des résultats diminue. En d'autres termes, vous perdez la validité externe.

Le deuxième type d'attrition est l'**attrition hétérogène**, dans laquelle l'effet d'attrition est différent pour différents groupes. Plus souvent appelé **attrition différentielle**, il s'agit d'une sorte de biais de sélection causé par l'étude elle-même. Supposons que, pour la première fois dans l'histoire de la psychologie, j'arrive à trouver l'échantillon de personnes parfaitement équilibré et représentatif. Je commence à faire « l'expérience incroyablement longue et fastidieuse de Dani » sur mon échantillon parfait mais ensuite, parce que mon étude est incroyablement longue et fastidieuse, beaucoup de gens commencent à abandonner. Je ne peux pas l'empêcher. Les participants ont absolument le droit de cesser toute expérience, à tout moment, pour quelque raison que ce soit, et en tant que chercheurs, nous sommes moralement (et professionnellement) obligés de rappeler aux gens qu'ils ont ce droit. Supposons donc que « l'expérience incroyablement longue et fastidieuse de Dani » ait un taux de décrochage très élevé. Quelles sont les chances que ce décrochage soit aléatoire ? Réponse : zéro. Il est presque certain que les personnes qui restent sont plus consciencieuses, plus tolérantes à l'ennui, etc. que celles qui partent. Dans la mesure où (disons) la conscience professionnelle est pertinente au phénomène psychologique qui m'intéresse, cette attrition peut diminuer la validité de mes résultats.

Voici un autre exemple. Supposons que je conçoive mon expérience avec deux conditions. En condition « traitement », l'expérimentateur insulte le participant et lui remet ensuite un questionnaire destiné à mesurer son obéissance. Dans la condition « contrôle », l'expérimentateur s'engage dans un petit bavardage inutile et leur donne ensuite le questionnaire. Laissons de côté les mérites scientifiques douteux et l'éthique douteuse d'une telle étude, réfléchissons à ce qui pourrait mal tourner ici. En règle générale, quand quelqu'un m'insulte en face, j'ai tendance à devenir beaucoup moins coopératif. Il y a donc de fortes chances qu'il y ait beaucoup plus de personnes qui abandonnent le traitement que de personnes qui en sont témoins. Et cet abandon ne sera pas aléatoire. Les personnes les plus susceptibles de se désister seraient probablement celles qui accordent peu d'importance à la participation docile à l'expérience. Étant donné que les personnes les plus désobéissantes et les plus irritées ont toutes quitté le groupe de traitement mais pas le groupe témoin, nous avons introduit une confusion : les personnes qui ont répondu au questionnaire dans le groupe de traitement étaient *déjà* plus susceptibles d'être consciencieuses et obéissantes que les personnes du groupe témoin. Bref, dans cette étude, insulter les gens ne les rend pas plus obéissants. Plus les gens désobéissent, plus ils quittent l'expérience ! La validité interne de cette expérience est complètement foutue.

Biais de non-réponse

Le biais de non-réponse est étroitement lié au biais de sélection et à l'attrition différentielle. La version la plus simple du problème est la suivante. Vous envoyez un sondage par la poste à 1000 personnes, mais seulement 300 d'entre elles y répondent. Les 300 personnes qui ont répondu ne constituent certainement pas un sous-échantillon aléatoire. Les personnes qui répondent aux enquêtes sont systématiquement différentes de celles qui n'y répondent pas. Cela pose un problème lorsque l'on tente de généraliser à partir des 300 personnes qui ont répondu à l'ensemble de la population, puisque l'on dispose maintenant d'un échantillon manifestement non aléatoire. La question du biais de non-réponse est toutefois plus générale. Parmi (disons) les 300 personnes qui ont répondu au sondage, vous constaterez peut-être que tout le monde ne répond pas à toutes les questions. Si (disons) 80 personnes ont choisi de ne pas répondre à l'une de vos questions, est-ce que cela pose problème ? Comme toujours, la réponse est peut-être. Si la question à laquelle on n'a pas répondu se trouvait à la dernière page du questionnaire, et que ces 80 questionnaires ont été retournés avec la dernière page manquante, il y a de fortes chances que les données manquantes ne soient pas un problème ; probablement que les pages sont simplement tombées. Cependant, si la question à laquelle 80 personnes n'ont pas répondu était la question personnelle la plus conflictuelle ou la plus invasive du questionnaire, alors vous avez presque certainement un problème. Il s'agit ici essentiellement de ce qu'on appelle le problème des **données manquantes**. Si les données manquantes ont été « perdues » au hasard, ce n'est pas un gros problème. Si elles manquent systématiquement, il peut s'agir d'un gros problème.

Régression à la moyenne

La régression à la moyenne fait référence à toute situation où vous sélectionnez des données en fonction d'une valeur extrême sur une mesure donnée. Comme la variable a une

variation naturelle, cela signifie presque certainement que lorsque vous prenez une mesure ultérieure, la mesure ultérieure sera moins extrême que la première, purement par hasard.

En voici un exemple. Supposons que je m'intéresse à la question de savoir si une formation en psychologie a un effet négatif sur les enfants très intelligents. Pour ce faire, je trouve les 20 étudiants en psychologie I qui ont les meilleures notes au secondaire et je regarde comment ils réussissent à l'université. Il s'avère qu'ils réussissent beaucoup mieux que la moyenne, mais ils ne sont pas en tête de la classe à l'université, même s'ils ont terminé premiers de leurs cours au secondaire. Qu'est-ce qu'il se passe ? La première pensée naturelle est que cela doit signifier que les cours de psychologie doivent avoir un effet négatif sur ces étudiants. Cependant, bien que cela puisse très bien être l'explication, il est plus probable que ce que vous voyez est un exemple de « régression vers la moyenne ». Pour voir comment cela fonctionne, prenons un moment pour réfléchir à ce qui est nécessaire pour obtenir la meilleure note dans une classe, que ce soit à l'école secondaire ou à l'université. Quand vous avez une grande classe, il y aura *beaucoup* de gens très intelligents inscrits. Pour obtenir la meilleure note, il faut être très intelligent, travailler très dur et avoir un peu de chance. L'examen doit poser les bonnes questions en fonction de vos compétences idiosyncrasiques, et vous devez éviter de faire des erreurs stupides (nous le faisons tous parfois) en y répondant. Et c'est ça, alors que l'intelligence et le travail acharné sont transférables d'une classe à l'autre, la chance ne l'est pas. Les gens qui ont eu de la chance à l'école secondaire ne seront pas les mêmes que ceux qui ont eu de la chance à l'université. C'est la définition même de la « chance ». La conséquence en est que lorsque vous sélectionnez des personnes aux valeurs extrêmes d'une mesure (les 20 meilleurs élèves), vous choisissez pour le travail acharné, la compétence et la chance. Mais comme la chance ne se transfère pas à la deuxième mesure (seulement la compétence et le travail), on s'attend à ce que toutes ces personnes diminuent un peu quand on les mesure une deuxième fois (à l'université). Leurs scores reviennent donc un peu en arrière, vers tout le monde. C'est une régression vers la moyenne.

La régression vers la moyenne est étonnamment courante. Par exemple, si deux personnes très grandes ont des enfants, leurs enfants auront tendance à être plus grands que la moyenne, mais pas aussi grands que les parents. C'est l'inverse qui se produit chez les parents très petits. Deux parents très petits auront tendance à avoir des enfants petits, mais néanmoins ces enfants auront tendance à être plus grands que les parents. Elle peut aussi être extrêmement subtile. Par exemple, des études ont montré que les gens apprennent mieux de la rétroaction négative que de la rétroaction positive. Cependant, la façon dont les gens ont essayé de montrer cela était de donner aux gens un renforcement positif chaque fois qu'ils faisaient bien, et un renforcement négatif chaque fois qu'ils faisaient mal. On peut observer qu'après le renforcement positif, les gens avaient tendance à faire moins bien, alors qu'après le renforcement négatif, ils avaient tendance à faire mieux. Nous avons ici un biais de sélection! Quand les gens réussissent très bien, on observe des valeurs « élevées », et il faut donc *s'attendre*, en raison de la régression vers la moyenne, à ce que la performance lors du prochain essai soit moins bonne, peu importe si le renforcement est donné ou non. De même, après un mauvais essai, les gens auront tendance à s'améliorer tous seuls. La supériorité apparente de la rétroaction négative est un artefact causé par la régression vers la moyenne (voir Kahneman et Tversky, 1973, pour une discussion).

Biais de l'expérimentateur

Les biais de l'expérimentateur** peuvent prendre de multiples formes. L'idée de base est que l'expérimentateur, malgré les meilleures intentions, peut finir par influencer accidentellement les résultats de l'expérience en communiquant subtilement la « bonne réponse » ou le « comportement souhaité » aux participants. Généralement, cela se produit parce que l'expérimentateur a des connaissances particulières que le participant ne possède pas, par exemple la bonne réponse aux questions posées ou la connaissance du modèle de performance attendu pour l'état dans lequel se trouve le participant. L'exemple classique est l'étude de cas de « Clever Hans », qui remonte à 1907 (Pfungst 1911; Hothersall 2004). Clever Hans était un cheval qui était apparemment capable de lire, de compter et de réaliser d'autres exploits caractéristiques des humains ressemblant à de l'intelligence. Après que Clever Hans soit devenu célèbre, les psychologues ont commencé à examiner son comportement de plus près. Il s'est avéré que, comme on pouvait s'y attendre, Hans ne savait pas faire de maths. Au contraire, Hans répondait aux observateurs humains qui l'entouraient, parce que les humains savaient compter et que le cheval avait appris à changer de comportement quand les gens changeaient le leur.

La solution générale au problème du biais de l'expérimentateur est de s'engager dans des études en double aveugle, où ni l'expérimentateur ni le participant ne savent dans quel état se trouve le participant ou quel est le comportement souhaité. C'est une très bonne solution au problème, mais il est important de reconnaître que ce n'est pas tout à fait idéal et difficile à réaliser parfaitement. Par exemple, la façon évidente dont je pourrais essayer de construire une étude en double aveugle, c'est qu'un de mes étudiants au doctorat (un qui ne connaît rien à l'expérience) mène l'étude. J'ai l'impression que ça devrait suffire. La seule personne (moi) qui connaît tous les détails (ex. les bonnes réponses aux questions, les affectations des participants aux conditions) n'a aucune interaction avec les participants, et la personne qui parle aux gens (l'étudiant en doctorat) ne sait rien. Sauf qu'en réalité que la dernière partie est très peu susceptible d'être vraie. Pour que l'étudiant au doctorat puisse mener l'étude efficacement, il doit avoir été informé par moi, le chercheur. Et il se trouve que l'étudiant au doctorat me connaît aussi et connaît un peu mes croyances générales sur les gens et la psychologie (par exemple, j'ai tendance à penser que les humains sont beaucoup plus intelligents que les psychologues ne le croient). Par conséquent, il est presque impossible pour l'expérimentateur d'éviter d'en savoir un peu plus sur les attentes que j'ai. Et même un peu de connaissance peut avoir un effet. Supposons que l'expérimentateur communique accidentellement le fait que l'on s'attend à ce que les participants réussissent bien dans cette tâche. En vertu de ce qu'on appelle « l'effet Pygmalion », si vous attendez de grandes choses des gens, ils auront tendance à être à la hauteur de vos attentes. Mais si vous vous attendez à ce qu'ils échouent, ils le feront aussi. En d'autres termes, les attentes deviennent une prophétie qui se réalise d'elle-même.

Effets de la demande et réactivité

Lorsqu'on parle de biais de l'expérimentateur, on craint que les connaissances ou les désirs de l'expérimentateur soient communiqués aux participants et qu'ils puissent changer le comportement des gens (Rosenthal, 1966). Cependant, même si vous parvenez à empêcher cela, il est presque impossible d'empêcher les gens de savoir qu'ils font partie d'une étude

psychologique. Et le simple fait de savoir que quelqu'un vous regarde ou étudie peut avoir un effet assez important sur votre comportement. C'est ce qu'on appelle généralement les **effets de réactivité** ou **de demande**. L'idée de base est illustré par l'effet Hawthorne : les gens modifient leur performance en raison de l'attention que l'étude leur porte. L'effet tire son nom d'une étude qui a eu lieu dans l'usine « Hawthorne Works » près de Chicago (voir Adair 1984). Cette étude, datant des années 1920, portait sur les effets de l'éclairage des usines sur la productivité des travailleurs. Ce qui est important, c'est que les travailleurs ont changé de comportement parce qu'ils savaient qu'ils faisaient l'objet d'une étude, plutôt que d'un quelconque effet de l'éclairage de l'usine.

Pour mieux comprendre la façon dont le simple fait de participer à une étude peut changer le comportement des gens, il est utile de penser comme un psychologue social et d'examiner les *rôles* que les gens pourraient *adopter* pendant une expérience, mais qu'ils pourraient *ne pas adopter* si les événements correspondants survenaient dans le monde réel :

- Le *bon participant* essaie d'être trop utile au chercheur. Il cherche à comprendre les hypothèses de l'expérimentateur et à les confirmer.
- Le *participant négatif* fait exactement le contraire du bon participant. Il ou elle cherche à briser ou à invalider l'étude ou l'hypothèse d'une manière ou d'une autre.
- Le *participant fidèle* est anormalement obéissant. Il ou elle cherche à suivre parfaitement les instructions, quoi qu'il se soit passé dans un contexte plus réaliste.
- L'*appréhension du participant.* Il devient nerveux à l'idée d'être testé ou étudié, à tel point que son comportement devient très anormal ou trop désirable sur le plan social.

Effets placebo

L'**effet placebo** est un type spécifique d'effet de demande qui nous inquiète beaucoup. Il s'agit de la situation où le simple fait d'être traité entraîne une amélioration des résultats. L'exemple classique vient des essais cliniques. Si vous donnez aux gens un médicament chimiquement inerte et que vous leur dites que c'est un remède contre une maladie, ils auront tendance à aller mieux plus vite que les gens qui ne sont pas traités du tout. En d'autres termes, c'est la croyance des gens qu'ils sont traités qui cause l'amélioration des résultats, et non le médicament.

Situation, mesure et effets sur la sous-population

A certains égards, ces termes sont des fourre-tout pour désigner « toutes les autres menaces à la validité externe ». Ils font référence au fait que le choix de la sous-population à partir de laquelle vous recrutez vos participants, le lieu, le moment et la manière dont vous menez votre étude (y compris qui collecte les données) et les outils que vous utilisez pour effectuer vos mesures pourraient tous influencer les résultats. Plus précisément, on craint que ces facteurs n'influencent les résultats d'une manière telle qu'ils ne se généralisent pas à un plus grand nombre de personnes, de lieux et de mesures.

Fraude, tromperie et auto-illusion

Il est difficile d'amener un homme à comprendre quelque chose, quand son salaire dépend de son incompréhension.
- Upton Sinclair

Il y a une dernière chose que je pense devoir mentionner. En lisant ce que les manuels ont souvent à dire sur l'évaluation de la validité d'une étude, je n'ai pas pu m'empêcher de remarquer qu'ils semblent supposer que le chercheur est honnête. Je trouve ça hilarant. Bien que la grande majorité des scientifiques soient honnêtes, du moins d'après mon expérience, certains ne le sont pas.[13] De plus, comme je l'ai mentionné plus tôt, les scientifiques ne sont pas à l'abri des préjugés. Il est facile pour un chercheur de se tromper en fin de compte et d'avoir des croyances erronées, ce qui peut l'amener à mener des recherches subtilement imparfaites, puis à cacher ces défauts lorsqu'il les écrit. Il faut donc tenir compte non seulement de la possibilité (probablement peu probable) de fraude pure et simple, mais aussi de la possibilité (probablement assez courante) que la recherche soit involontairement « orientée ». J'ai ouvert quelques manuels standard et je n'ai pas trouvé beaucoup de discussion sur ce problème, alors voici ma propre tentative pour énumérer quelques façons dont ces questions peuvent se poser :

- **Fabrication de données**. Parfois, les gens se contentent d'inventer les données. Cela se fait parfois avec de « bonnes » intentions. Par exemple, le chercheur croit que les données fabriquées reflètent la vérité et qu'elles peuvent en fait refléter des versions « légèrement nettoyées » de données réelles. Dans d'autres cas, la fraude est délibérée et malveillante. Cyril Burt (un psychologue qui aurait fabriqué certaines de ses données), Andrew Wakefield (accusé d'avoir fabriqué ses données reliant le vaccin ROR à l'autisme) et Hwang Woo-suk (qui a falsifié beaucoup de ses données sur les cellules souches) sont des exemples très médiatisés où la fabrication de données a été alléguée ou présentée.

- **Des canulars**. Les canulars ont beaucoup de similitudes avec la fabrication de données, mais ils diffèrent quant à l'usage auquel ils sont destinés. Un canular est souvent une blague, et beaucoup d'entre eux sont destinés à être (éventuellement) découverts. Souvent, le but d'un canular est de discréditer quelqu'un ou un domaine. Il y a eu pas mal de canulars scientifiques bien connus au fil des ans (p. ex., l'homme de Piltdown) et certains étaient des tentatives délibérées de discréditer certains domaines de recherche (p. ex., l'affaire Sokal).

- **Représentation erronée des données**. Bien que la fraude fasse les manchettes, il est beaucoup plus courant, d'après mon expérience, de voir des données déformées.

---

[13] Certains diront que si vous n'êtes pas honnête, vous n'êtes pas un vrai scientifique. Ce qui est vrai en partie, je suppose, mais c'est malhonnête (regardez l'erreur "No true Scotsman"). Le fait est qu'il y a beaucoup de gens qui sont ostensiblement employés en tant que scientifiques, et dont le travail a tous les attributs de la science, mais qui sont carrément frauduleux. Prétendre qu'ils n'existent pas en disant qu'ils ne sont pas des scientifiques, c'est un raisonnement confus.

Quand je dis cela, je ne parle pas des journaux qui se trompent (ce qu'ils font, presque toujours). Je fais allusion au fait que, souvent, les données ne disent pas ce que les chercheurs pensent qu'ils disent. Je pense que, presque toujours, ce n'est pas le résultat d'une malhonnêteté délibérée, mais plutôt d'un manque de sophistication dans l'analyse des données. Repensez, par exemple, au paradoxe de Simpson dont j'ai parlé au début de ce livre. Il est très courant de voir des gens présenter des données « agrégées » d'une sorte ou d'une autre et parfois, lorsque vous creusez plus profondément et que vous trouvez les données brutes, vous constatez que les données agrégées racontent une histoire différente des données désagrégées. Par ailleurs, il se peut que vous découvriez qu'un aspect des données est caché parce qu'il raconte une histoire gênante (p. ex. le chercheur peut choisir de ne pas faire référence à une variable particulière). Il y a beaucoup de variantes à ce sujet, dont beaucoup sont très difficiles à détecter.

- **Erreur de conception**. D'accord, celle-ci est subtile. Le problème ici, c'est essentiellement qu'un chercheur conçoit une étude qui comporte des lacunes et que ces lacunes ne sont jamais signalées dans la revue. Les données qui sont rapportées sont tout à fait réelles et correctement analysées, mais elles sont produites par une étude qui est en fait très mal faite. Le chercheur veut vraiment trouver un effet particulier et c'est pourquoi l'étude est conçue de manière à ce qu'il soit « facile » d'observer (artefactuellement) cet effet. Une façon sournoise de le faire, au cas où vous auriez envie de vous lancer dans un peu de fraude, est de concevoir une expérience dans laquelle il est évident pour les participants de faire ce qu'ils sont « censés » faire, puis de laisser la réactivité faire son effet magique pour vous. Si vous le souhaitez, vous pouvez ajouter tous les pièges de l'expérimentation en double aveugle, mais cela ne changera rien puisque le matériel d'étude lui-même dit subtilement aux gens ce que vous voulez qu'ils fassent. Lorsque vous rédigez les résultats, la fraude n'est pas évidente pour le lecteur. Ce qui est évident pour le participant lorsqu'il est dans le contexte expérimental ne l'est pas toujours pour la personne qui lit l'article. Bien sûr, la façon dont je l'ai décrit donne l'impression que c'est toujours de la fraude. Il y a probablement des cas où cela est fait délibérément, mais d'après mon expérience, il est plus probable que cela relève d'une mauvaise conception non intentionnelle. Le chercheur y *croit* et il se trouve que l'étude finit par présenter un défaut qui s'efface comme par magie lorsque l'étude est rédigée en vue de sa publication.

- **Exploration de données et hypothèses post hoc.** Une autre façon dont les auteurs d'une étude peuvent plus ou moins déformer les données est de s'engager dans ce qu'on appelle le « data mining » (voir Gelman and Loken 2014 pour une discussion plus large à ce sujet dans le cadre du « jardin aux sentiers qui bifurquent » en analyse statistique). Comme nous le verrons plus loin, si vous continuez à essayer d'analyser vos données de différentes manières, vous finirez par trouver quelque chose qui « ressemble » à un effet réel mais ne l'est pas. C'est ce qu'on appelle le « data mining ». Auparavant, c'était assez rare parce que l'analyse des données prenait des semaines, mais maintenant que tout le monde a un logiciel statistique très puissant sur son ordinateur, c'est devenu très courant. L'exploration de données en soi n'est pas « fausse », mais plus vous en faites, plus le risque que vous prenez est grand. Ce qui ne

va pas, et je soupçonne que c'est très courant, c'est l'exploration de données *non reconnue.* C'est-à-dire que le chercheur effectue toutes les analyses possibles connues de l'humanité, trouve celle qui fonctionne et prétend ensuite que c'est la seule analyse qu'il ait jamais faite. Pire encore, ils « inventent » souvent une hypothèse après avoir examiné les données pour masquer le data mining. Pour être clair. Il n'y a pas de mal à changer de croyance après avoir regardé les données et à réanalyser vos données à l'aide de vos nouvelles hypothèses « post hoc ». Ce qui ne va pas (et je soupçonne que c'est courant), c'est de ne pas reconnaître que vous l'avez fait. Si vous reconnaissez que vous l'avez fait, d'autres chercheurs pourront tenir compte de votre comportement. Si vous ne le faites pas, ils ne peuvent pas. Cela rend votre comportement trompeur.

- **Biais de publication et autocensure.** Enfin, un biais omniprésent est le fait de « ne pas rapporter « les résultats négatifs. C'est presque impossible à prévenir. Les revues ne publient pas tous les articles qui leur sont soumis. Ils préfèrent publier des articles qui trouvent « quelque chose ». Donc, si 20 personnes font une expérience pour savoir si la lecture de *Finnegans Wake* cause de la folie chez les humains, et que 19 d'entre elles découvrent que ce n'est pas le cas, laquelle d'entre elles sera publiée selon vous ? Évidemment, c'est la seule étude qui a trouvé que *Finnegans Wake* cause la folie. Il s'agit d'un[14] exemple de *biais de publication*. Comme personne n'a jamais publié les 19 études qui n'ont pas trouvé d'effet, un lecteur naïf ne saurait jamais qu'elles existent. Pire encore, la plupart des chercheurs « internalisent » ce biais et finissent par *autocensurer* leurs recherches. Sachant que les résultats négatifs ne seront pas acceptés pour publication, ils n'essaient même pas de les rapporter. Comme le dit un de mes amis « pour chaque expérience publiée, vous avez aussi 10 échecs ». Et elle a raison. Le piège, c'est que si certaines (peut-être la plupart) de ces études sont des échecs pour des raisons sans intérêt (par exemple, vous avez fait une erreur), d'autres peuvent être de véritables résultats « nuls » que vous devriez reconnaître lorsque vous rédigez la « bonne » expérience. Et dire quoi est ce qui est souvent difficile à faire. Un bon point de départ est un article de Ioannidis (2005) intitulé « Why most published research findings are false ». Je suggère également de jeter un coup d'œil aux travaux de Kühberger, Fritz et Scherndl (2014) qui ont montré statistiquement que cela se produit effectivement en psychologie.

Il y a probablement beaucoup d'autres questions de ce genre auxquelles il faut penser, mais cela fera l'affaire pour commencer. Ce que je veux vraiment souligner, c'est la vérité aveuglante et évidente que la science du monde réel est menée par de vrais humains, et que seuls les plus crédules d'entre nous supposent automatiquement que tout le monde est honnête et impartial. Les scientifiques actuels ne sont généralement pas *si naïfs que ça,* mais pour une raison quelconque, le monde aime faire semblant de l'être, et les manuels scolaires que nous écrivons habituellement semblent renforcer ce stéréotype.

---

[14] De toute évidence, l'effet réel est que seuls les fous essaieraient même de lire Finnegans *Wake*.

Résumé

Ce chapitre n'a pas vraiment pour but de fournir une discussion exhaustive des méthodes de recherche en psychologie. Il faudrait un autre volume aussi long que celui-ci pour rendre justice au sujet. Cependant, dans la vie réelle, les statistiques et la conception des études sont si étroitement liées qu'il est très important de discuter de certains des sujets clés. Dans ce chapitre, j'ai abordé brièvement les sujets suivants :

- *Introduction à la mesure psychologique* (Section 2.1). Que signifie opérationnaliser une construction théorique ? Que signifie avoir des variables et faire des mesures ?
- *Échelles de mesure et types de variables* (section 2.2). Rappelez-vous qu'il y a ici *deux* distinctions différentes. Il y a la différence entre les données discrètes et continues, et il y a la différence entre les quatre différents types d'échelle (nominale, ordinale, intervalle et ratio).
- *Fiabilité d'une mesure* (section 2.3). Si je mesure deux fois la même chose, dois-je m'attendre à voir le même résultat ? Seulement si ma mesure est fiable. Mais qu'est-ce que cela signifie de parler de faire la « même chose » ? C'est pourquoi nous avons différents types de fiabilité. Ne l'oubliez pas.
- *Terminologie : prédicteurs et résultats* (Section 2.4). Quels rôles les variables jouent-elles dans une analyse ? Pouvez-vous vous rappeler la différence entre les prédicteurs et les résultats ? Variables dépendantes et indépendantes ? Etc.
- *Plans de recherche expérimentaux et non expérimentaux* (section 2.5). Qu'est-ce qui fait qu'une expérience est une expérience ? S'agit-il d'une belle blouse blanche ou est-ce que cela a quelque chose à voir avec le contrôle des variables par les chercheurs ?
- *Validité et menaces* (section 2.6). Votre étude mesure-t-elle ce que vous voulez qu'elle fasse ? Comment les choses peuvent-elles mal tourner ? Et est-ce mon imagination, ou était-ce une très longue liste de façons possibles dont les choses peuvent mal tourner ?

Tout cela devrait vous indiquer clairement que la conception de l'étude est un élément essentiel de la méthodologie de recherche. J'ai construit ce chapitre à partir du petit livre classique de Campbell et al (1963), mais il y a bien sûr un grand nombre de manuels sur la conception de la recherche. Passez quelques minutes avec votre moteur de recherche préféré et vous en trouverez des dizaines.

Démarrer avec Jamovi

Les robots sont agréables à utiliser. -Roger Zelazny[15]

---

[15] Source : *Dismal Light* (1968).

Dans ce chapitre, je vais discuter de la façon de commencer à Jamovi. Je vais parler brièvement de la façon de télécharger et d'installer Jamovi, mais la plus grande partie du chapitre sera axée sur la façon de vous aider à trouver votre chemin dans l'interface graphique Jamovi. Notre but dans ce chapitre n'est pas d'apprendre des concepts statistiques : nous essayons simplement d'apprendre les bases du fonctionnement de Jamovi et de nous familiariser avec le système. Pour ce faire, nous allons passer un peu de temps à examiner les ensembles de données et les variables. Ce faisant, vous aurez une petite idée de ce que c'est que de travailler en Jamovi.

Cependant, avant d'entrer dans les détails, il vaut la peine de parler un peu des raisons pour lesquelles vous voudrez peut-être utiliser Jamovi. Étant donné que vous lisez ceci, vous avez probablement vos propres raisons. Cependant, si ces raisons sont « parce que c'est ce qu'utilise mon cours de statistiques », il serait peut-être utile d'expliquer un peu pourquoi votre professeur a choisi d'utiliser Jamovi pour la classe. Bien sûr, je ne sais pas vraiment pourquoi d'*autres* personnes choisissent Jamovi alors mais je peux dire pourquoi je l'utilise.

- C'est évident, mais cela vaut la peine d'être dit : faire ses statistiques sur un ordinateur est plus rapide, plus facile et plus puissant que faire des statistiques à la main. Les ordinateurs excellent dans les tâches répétitives et sans réflexion, et beaucoup de calculs statistiques sont à la fois sans réflexion et répétitifs. Pour la plupart des gens, la seule raison d'effectuer des calculs statistiques au crayon et sur papier est l'apprentissage. Dans mon cours, je suggère à l'occasion de faire quelques calculs de cette façon, mais la seule valeur réelle est d'ordre pédagogique. Cela vous aide à vous faire une idée des statistiques et à faire quelques calculs vous-même, donc cela vaut la peine de le faire une fois pour toutes. Mais une seule fois !

- Faire des statistiques sur un tableur conventionnel (p. ex. Microsoft Excel) est généralement une mauvaise idée à long terme. Bien que de nombreuses personnes se sentent probablement plus familières avec elles, les feuilles de calcul sont très limitées en termes d'analyses possibles. Si vous prenez l'habitude d'essayer d'analyser vos données réelles à l'aide de feuilles de calcul, vous vous retrouvez dans un piège.

- Éviter les logiciels propriétaires est une très bonne idée. Il y a beaucoup de forfaits commerciaux que vous pouvez acheter, certains que j'aime et d'autres que je n'aime pas. Ils sont généralement très brillants dans leur apparence et généralement très puissants (beaucoup plus puissants que les feuilles de calcul). Cependant, ils sont aussi très coûteux. Habituellement, la compagnie vend des « versions étudiantes » (versions limitées) à très bas prix, et ensuite ils vendent des « versions éducatives » complètes à un prix exorbitant. Ils vendront également des licences commerciales à des prix incroyablement élevés. Le modèle d'affaires ici est de vous aspirer pendant vos études et de vous laisser ensuite dépendre de leurs outils lorsque vous sortez dans le monde réel. Il est difficile de les blâmer d'avoir essayé, mais personnellement, je ne suis pas favorable à dépenser des milliers de dollars si je peux l'éviter. Et vous pouvez l'éviter. Si vous utilisez des logiciels comme Jamovi qui sont open source et gratuits, vous ne serez jamais pris au piège et n'aurez jamais à payer des frais de licence exorbitants.

- Quelque chose que vous n'apprécierez peut-être pas maintenant, mais que vous adorerez plus tard si vous faites quoi que ce soit impliquant une analyse de données, est le fait que Jamovi est fondamentalement une interface sophistiquée pour le langage de programmation statistique libre R. Lorsque vous téléchargez et installez R, vous obtenez tous les « paquets » de base et ceux-ci sont très puissants à eux seuls. Cependant, parce que R est ouvert et largement utilisé, il est devenu une sorte d'outil standard dans les statistiques et beaucoup de gens écrivent leurs propres fonctions qui étendent les possibilités du système. Ceux-ci sont également disponibles gratuitement. L'une des conséquences de cette situation, et je l'ai remarqué, c'est que si vous regardez les manuels récents d'analyse avancée des données, *beaucoup* d'entre eux utilisent R.

Ce sont les principales raisons pour lesquelles j'utilise Jamovi. Mais ce n'est pas sans défaut. C'est relativement nouveau[16] et il n'y a donc pas beaucoup de manuels et d'autres ressources en appui, et il comporte quelques bizarreries avec lesquelles nous sommes tous coincés, mais dans l'ensemble je pense que les forces l'emportent sur les faiblesses ; plus que toute autre option que j'ai rencontrée jusqu'ici.

Installer Jamovi

Bien, cessons l'argumentaire de vente. Commençons tout de suite. Comme tout logiciel, Jamovi doit être installé sur un « ordinateur », qui est une boîte magique qui fait des choses cool et rase gratis ou quelque chose du genre ; je confonds peut-être les ordinateurs avec les campagnes de marketing pour iPad. Quoi qu'il en soit, Jamovi est distribué gratuitement en ligne et vous pouvez le télécharger à partir de la page d'accueil Jamovi suivante :

https://www.Jamovi.org/

En haut de la page, sous la rubrique « Download », vous trouverez des liens séparés pour les utilisateurs Windows, Mac et Linux. Si vous suivez le lien correspondant, vous verrez que les instructions en ligne sont assez explicites. Au moment d'écrire ces lignes, la version actuelle de Jamovi est la 0.9[17], mais ils publient habituellement des mises à jour tous les quelques mois, vous aurez donc probablement une version plus récente.[18]

---

[16] Au moment d'écrire ces lignes, en août 2018.

[17] NdT. En aout 2019, la version courante est la 1.0.4, preuve que le logiciel évolue rapidement

[18] Bien que Jamovi soit fréquemment mis à jour, cela ne fait généralement pas beaucoup de différence pour le genre de travail que nous ferons dans ce livre. En fait, pendant l'écriture du livre, j'ai fait plusieurs mises à jour à plusieurs reprises et cela n'a pas fait beaucoup de différence par rapport à ce qui se trouve dans ce livre.

Démarrer Jamovi

Quel que soit le système d'exploitation que vous utilisez, il est temps d'ouvrir Jamovi et de commencer. Lors du premier démarrage de Jamovi, vous verrez apparaître une interface utilisateur qui ressemble à celle de la Figure 3-1.

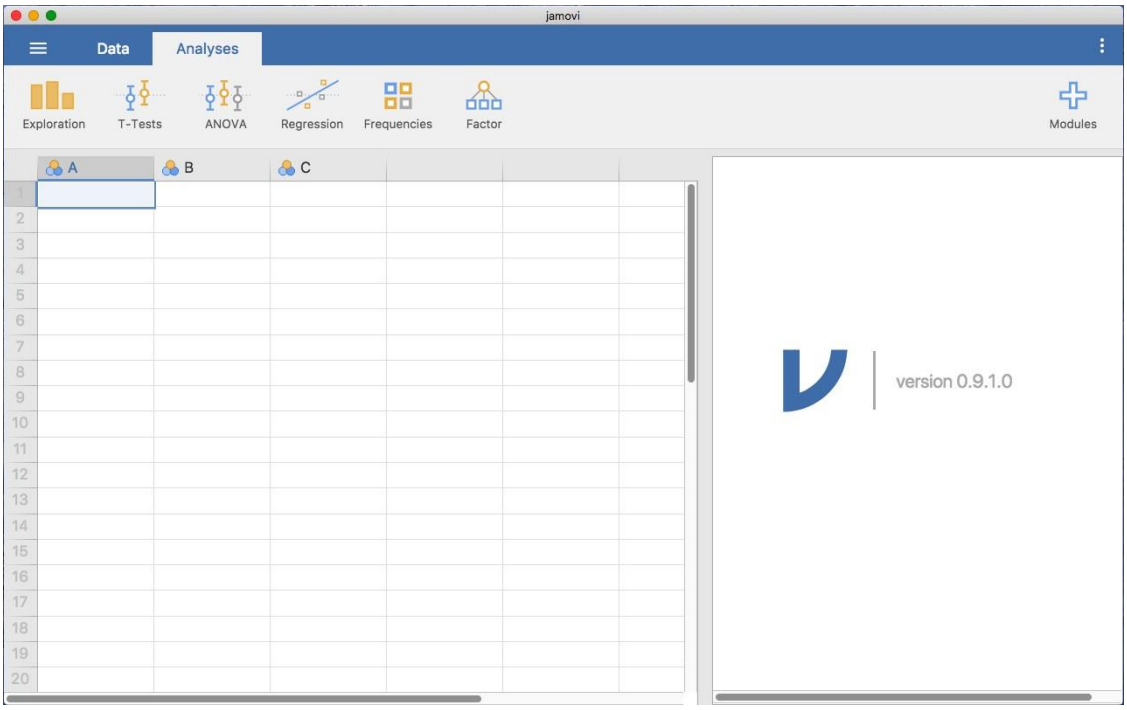


Figure 3-1 : Jamovi ressemble à ceci lorsque vous le démarrez.

A gauche se trouve la vue de la feuille de calcul, et à droite, les résultats des tests statistiques. Au milieu se trouve une barre qui sépare ces deux zones et qui peut être déplacée vers la gauche ou vers la droite pour changer leur taille.

Il est possible de simplement commencer à taper des valeurs dans le tableur Jamovi comme vous le feriez dans n'importe quel autre logiciel de feuille de calcul. Une autre façon de faire est d'importer des ensembles de données existants au format de fichier CSV (.csv) dans Jamovi. De plus, vous pouvez facilement importer des fichiers SPSS, SAS, Stata et JASP directement dans Jamovi. Pour ouvrir un fichier, sélectionnez l'onglet Fichier (trois lignes horizontales signifient cet onglet) dans le coin supérieur gauche, sélectionnez « Open », puis choisissez parmi les fichiers énumérés dans « Browse » selon que vous voulez ouvrir un exemple ou un fichier stocké sur votre ordinateur.

Analyses

Les analyses peuvent être sélectionnées à partir du ruban d'analyse ou du menu situé en haut. La sélection d'une analyse présentera un « options panel » pour cette analyse particulière, vous permettant d'assigner différentes variables à différentes parties de l'analyse et de sélectionner différentes options. En même temps, les résultats de l'analyse

apparaîtront dans le panneau de droite « Résultats » et seront mis à jour en temps réel au fur et à mesure que vous apporterez des changements aux options.

Lorsque l'analyse est correctement configurée, vous pouvez fermer les options d'analyse en cliquant sur la flèche en haut à droite du panneau optionnel. Si vous souhaitez revenir à ces options, vous pouvez cliquer sur les résultats obtenus. De cette façon, vous pouvez revenir à n'importe quelle analyse que vous (ou disons, un collègue) avez créée précédemment.

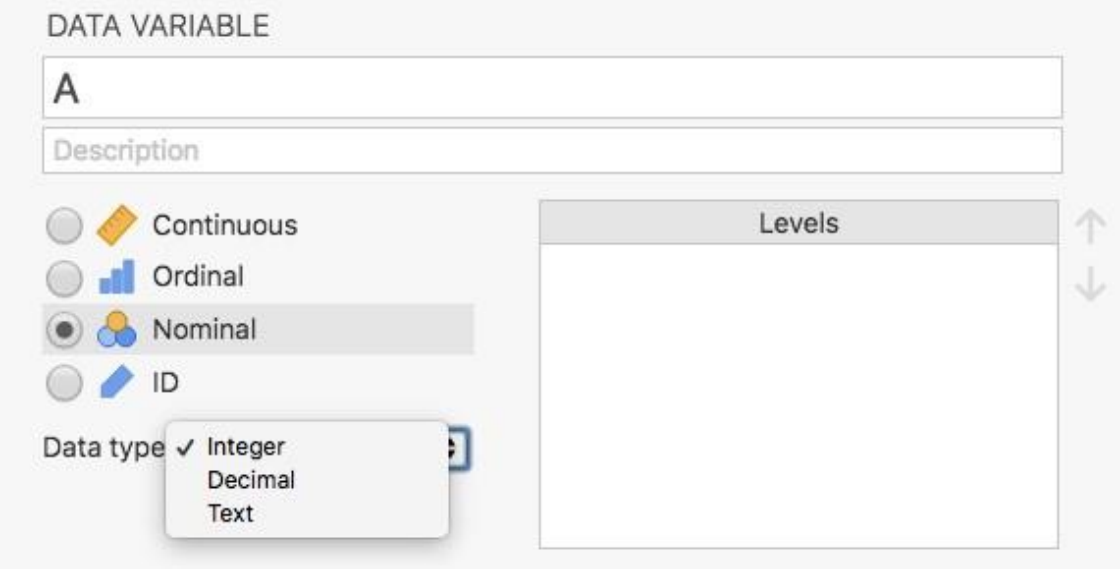
Si vous décidez que vous n'avez plus besoin d'une analyse particulière, vous pouvez la supprimer avec le menu contextuel des résultats. En cliquant avec le bouton droit de la souris sur les résultats de l'analyse, un menu s'affiche et en sélectionnant « Analyses » puis « Remove », l'analyse peut être supprimée. Mais on en reparlera plus tard. Tout d'abord, jetons un coup d'œil plus détaillé à la vue de la feuille de calcul.

La feuille de calcul

Dans Jamovi, les données sont représentées dans une feuille de calcul, chaque colonne représentant une « variable » et chaque ligne un « cas » ou un « participant ».

Variables

Les variables les plus couramment utilisées dans Jamovi sont les « Variables de données », ces variables contiennent simplement des données soit chargées depuis un fichier de données, soit « saisies » par l'utilisateur. Les variables de données peuvent être l'un des trois niveaux de mesure :s



Ces niveaux sont désignés par le symbole dans l'en-tête de la colonne de la variable.<:p>

Le type de variable *ID* est unique à Jamovi. Il est destiné aux variables qui contiennent des identificateurs que vous ne voudriez presque jamais analyser. Par exemple, un nom de personne ou un ID de participant. La spécification d'un type de variable ID peut améliorer les performances lors de l'interaction avec de très grands ensembles de données.

Les variables *nominales* sont pour les variables catégorielles qui sont des étiquettes de texte, par exemple, une colonne intitulée *Sexe* avec les valeurs *Homme* et *Femme* serait nominale. Tout comme le nom d'une personne. Les valeurs des variables nominales peuvent aussi avoir une valeur numérique. Ces variables sont le plus souvent utilisées lors de l'importation de données qui codent les valeurs avec des nombres plutôt qu'avec du texte. Par exemple, une colonne d'un ensemble de données peut contenir les valeurs 1 pour les hommes et 2 pour les femmes. Il est possible d'ajouter des étiquettes « lisibles par l'homme » à ces valeurs avec l'éditeur de variables (plus d'informations à ce sujet ultérieurement).

Les variables *ordinales* sont comme les variables nominales, sauf que les valeurs ont un ordre spécifique. Un exemple est une échelle de Likert où 3 signifie » tout à fait d'accord » et -3 signifie « tout à fait en désaccord ».

Les variables *continues* sont des variables qui existent sur une échelle continue. Il peut s'agir par exemple de la taille ou du poids. C'est ce que l'on appelle aussi « l'intervalle » ou « l'échelle de rapport ».

En outre, vous pouvez également spécifier différents types de données : les variables ont un type de données « Text », « Integer » (entier) ou « Décimal ».

Lorsque vous commencez avec une feuille de calcul vierge et que vous tapez des valeurs dans le type de variable, elles changent automatiquement en fonction des données que vous saisissez. C'est un bon moyen pour avoir une idée de quels types de variables vont avec quels types de données. De même, lors de l'ouverture d'un fichier de données, Jamovi essaiera de deviner le type de variable à partir des données de chaque colonne. Dans les deux cas, cette approche automatique peut ne pas être correcte et il peut être nécessaire de spécifier manuellement le type de variable avec l'éditeur de variable.

L'éditeur de variables peut être ouvert en sélectionnant « Setup » dans l'onglet de données ou en double-cliquant sur l'en-tête de colonne variable. L'éditeur de variables vous permet de modifier le nom de la variable et, pour les variables de données, le type de variable, l'ordre des niveaux et l'étiquette affichée pour chaque niveau. Les modifications peuvent être appliquées en cliquant sur le bouton en haut à droite. L'éditeur de variables peut être désactivé en cliquant sur la flèche « Hide ».

De nouvelles variables peuvent être insérées ou ajoutées à l'ensemble de données à l'aide du bouton « Add » du ruban de données. Le bouton « Add » permet également d'ajouter des variables calculées.

Variables calculées

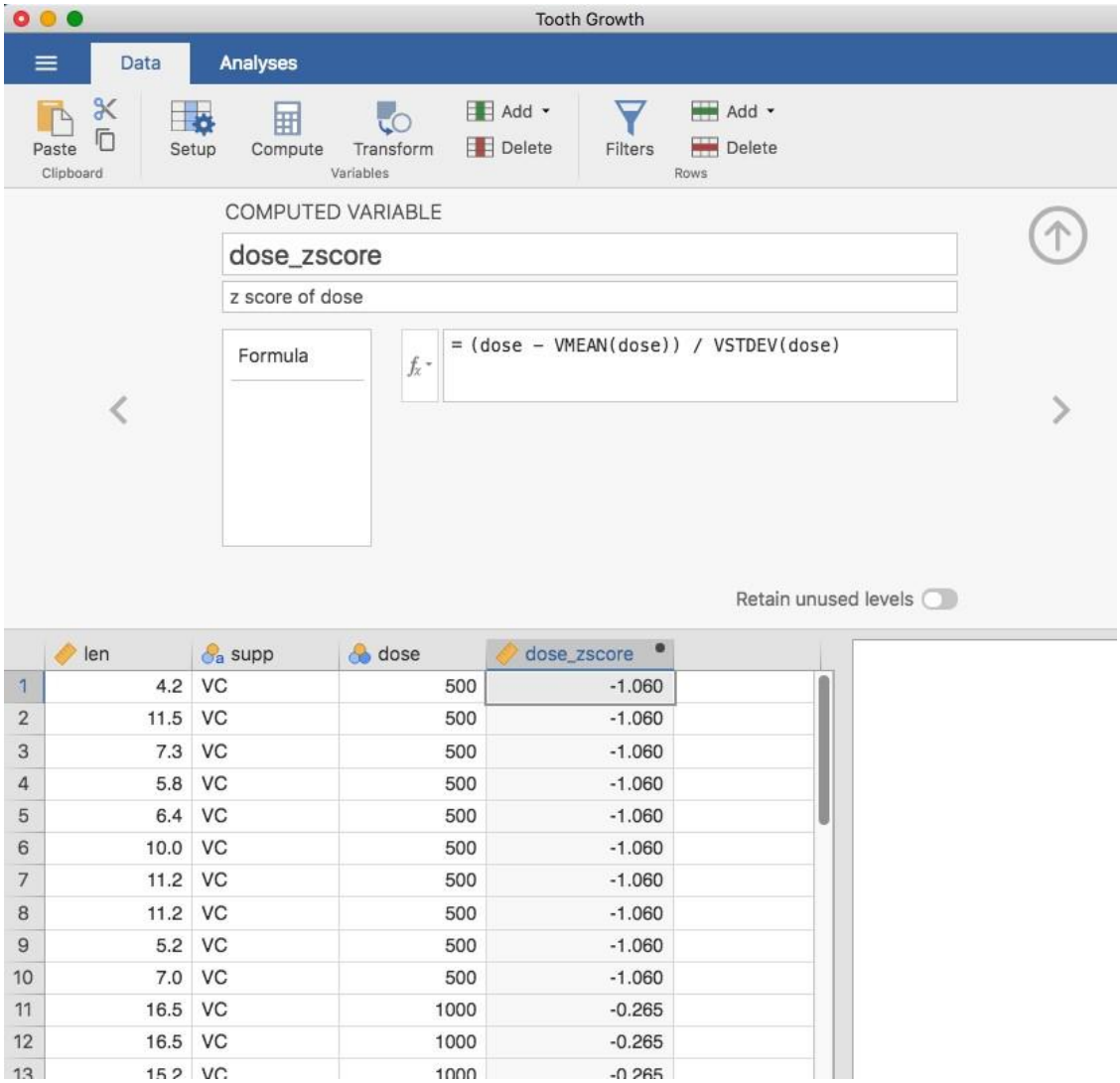
Les variables calculées sont celles qui prennent leur valeur en effectuant un calcul sur d'autres variables. Les variables calculées peuvent être utilisées à diverses fins, y compris les transformations logarithmiques, les z-scores, les scores sommatifs, les scores négatifs et les moyennes.

Les variables calculées peuvent être ajoutées à l'ensemble de données à l'aide du bouton « ajouter » disponible dans l'onglet Données. Vous obtiendrez une boîte de formule dans

laquelle vous pouvez spécifier la formule. Les opérateurs arithmétiques habituels sont disponibles. Voici quelques exemples de formules :

A + B LOG10(len) MOYEN(A, B) (dose - VMEAN(dose)) / VSTDEV(dose)

Dans l'ordre, il s'agit de la somme de A et B, d'une transformation logarithmique (base 10) de len, de la moyenne de A et B et du z-score de la dose variable. La Figure 3-2 ci-dessous montre l'écran Jamovi pour la nouvelle variable calculée comme le score z de la dose (à partir de l'exemple de l'ensemble de données « Tooth Growth »).



	len	supp	dose	dose_zscore
1	4.2	VC	500	-1.060
2	11.5	VC	500	-1.060
3	7.3	VC	500	-1.060
4	5.8	VC	500	-1.060
5	6.4	VC	500	-1.060
6	10.0	VC	500	-1.060
7	11.2	VC	500	-1.060
8	11.2	VC	500	-1.060
9	5.2	VC	500	-1.060
10	7.0	VC	500	-1.060
11	16.5	VC	1000	-0.265
12	16.5	VC	1000	-0.265
13	15.2	VC	1000	-0.265

Figure 3-2: Variable nouvellement calculée, le z-score de « dose »

Fonctions V

Plusieurs fonctions sont déjà disponibles dans Jamovi et dans le menu déroulant *fx*. Un certain nombre de fonctions apparaissent par paires, l'une préfixée d'un V et l'autre non. Les fonctions V effectuent leur calcul sur l'ensemble d'une variable, alors que les fonctions

non V effectuent leur calcul ligne par ligne. Par exemple, MEAN(A, B) produira la moyenne de A et B pour chaque rangée. Où VMEAN(A) donne la moyenne de toutes les valeurs de A.

Copier et coller

Jamovi produit de beaux tableaux formatés selon les normes de l'American Psychological Association (APA) et des graphiques attrayants. Il est souvent utile de pouvoir les copier-coller, par exemple dans un document Word ou dans un courriel à un collègue. Pour copier les résultats, cliquez avec le bouton droit de la souris sur l'objet qui vous intéresse et sélectionnez dans le menu exactement ce que vous voulez copier. Le menu vous permet de choisir de ne copier que l'image ou l'analyse complète. Sélectionner « copier » permet de copier le contenu dans le presse-papiers et de le coller dans d'autres programmes de la manière habituelle. Vous pourrez vous entrainer à cela plus tard quand nous ferons des analyses.

Mode syntaxe

Jamovi propose également un « R Syntax Mode ». Dans ce mode, Jamovi produit un code R équivalent pour chaque analyse. Pour passer en mode syntaxique, sélectionnez le menu Application en haut à droite de Jamovi (un bouton avec trois points verticaux) et cochez la case « Syntax mode ». Vous pouvez désactiver le mode syntaxique en cliquant une seconde fois dessus.

En mode syntaxique, les analyses continuent de fonctionner comme auparavant, mais elles produisent maintenant une syntaxe R et une « sortie ascii » comme dans une session R. Comme tous les objets de résultats dans Jamovi, vous pouvez faire un clic droit sur ces éléments (y compris la syntaxe R) et les copier-coller, par exemple dans une session R. Actuellement, la syntaxe R fournie n'inclut pas l'étape d'importation de données et doit donc être effectuée manuellement dans R. Il existe de nombreuses ressources expliquant comment importer des données dans R et si vous êtes intéressé, nous vous recommandons d'y jeter un coup d'œil ; il suffit de faire une recherche sur le web.

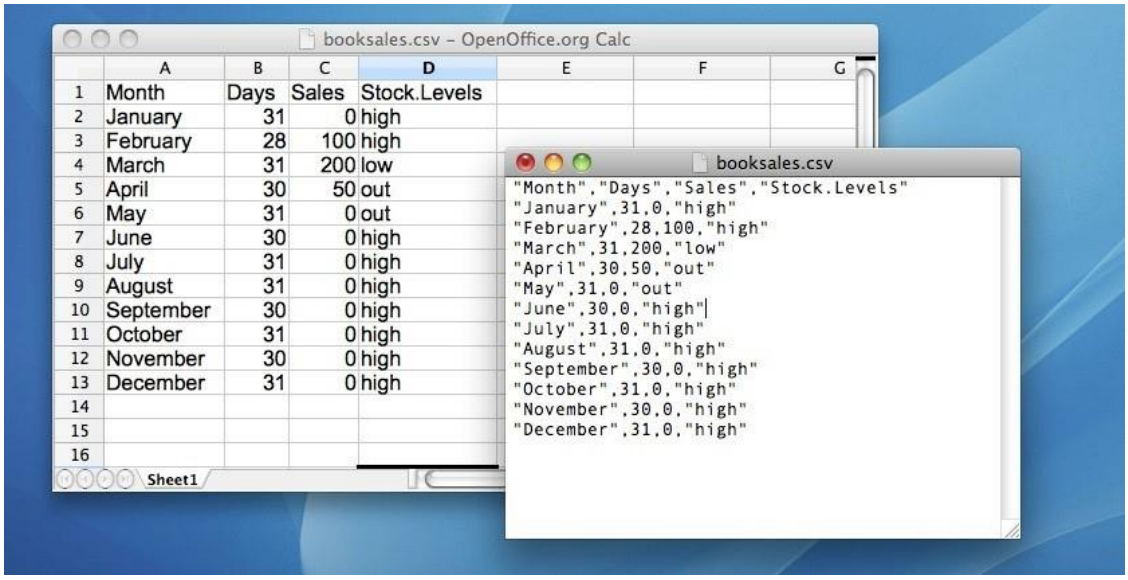
Chargement des données dans Jamovi

Il existe plusieurs types de fichiers différents qui sont susceptibles de nous intéresser dans le cadre de l'analyse des données. Il y en a deux en particulier qui sont particulièrement importants du point de vue de ce livre :

- *Jamovi* sont ceux avec une extension de fichier.omv. C'est le type de fichier standard que Jamovi utilise pour stocker les données, les variables et les analyses.
- *Les fichiers de valeurs séparées par des virgules (csv)* sont ceux dont l'extension de fichier est.csv. Il ne s'agit que d'anciens fichiers texte ordinaires et ils peuvent être ouverts avec de nombreux logiciels différents. Il est très courant de stocker des données dans des fichiers csv, précisément parce qu'ils sont si simples.

Il existe également plusieurs autres types de fichiers de données que vous pouvez importer dans Jamovi. Par exemple, vous pouvez ouvrir des feuilles de calcul Microsoft Excel (fichiers.xls) ou des fichiers de données qui ont été enregistrés dans les formats de fichier

natifs d'autres logiciels de statistiques, tels que SPSS ou SAS. Quel que soit le format de fichier que vous utilisez, une bonne pratique consiste à créer un dossier ou des dossiers spécifiques pour vos ensembles de données et analyses Jamovi et de vous assurer de les sauvegarder régulièrement.



	A	B	C	D	E	F	G
1	Month	Days	Sales	Stock.Levels			
2	January	31	0	high			
3	February	28	100	high			
4	March	31	200	low			
5	April	30	50	out			
6	May	31	0	out			
7	June	30	0	high			
8	July	31	0	high			
9	August	31	0	high			
10	September	30	0	high			
11	October	31	0	high			
12	November	30	0	high			
13	December	31	0	high			
14							
15							
16							

Figure 3-3 : Le fichier de données booksales.csv. Sur la gauche, j'ai ouvert le fichier à l'aide d'un tableur (OpenOffice), qui montre que le fichier est essentiellement un tableau. A droite, le même fichier est ouvert dans un éditeur de texte standard (le programme TextEdit sur un Mac), qui montre comment le fichier est formaté. Les entrées du tableau sont entourées de guillemets et séparées par des virgules.

Importation de données à partir de fichiers csv

Un format de données assez communément utilisé est l'humble fichier « comma separated value », aussi appelé fichier csv, et portant généralement l'extension de fichier.csv. Les fichiers csv sont simplement des fichiers texte démodés et ce qu'ils stockent n'est en fait qu'un tableau de données. Ceci est illustré dans la Figure 3-3, qui montre un fichier appelé booksales.csv que j'ai créé. Comme vous pouvez le voir, chaque ligne représente les données sur les ventes de livres pour un mois. La première ligne ne contient pas de données réelles, mais les noms des variables.

Il est facile d'ouvrir des fichiers csv dans Jamovi. Dans le menu en haut à gauche (le bouton avec trois lignes parallèles) choisissez « Open » et naviguez jusqu'à l'endroit où vous avez stocké le fichier csv sur votre ordinateur. Si vous êtes sur un Mac, il ressemblera à la fenêtre habituelle du Finder que vous utilisez pour choisir un fichier ; sur Windows, il ressemblera à une fenêtre d'explorateur. Un exemple de ce à quoi cela ressemble sur un Mac est illustré à la Figure 3-4. Je suppose que vous êtes familier avec votre propre ordinateur, donc vous ne devriez pas avoir de problème à trouver le fichier csv que vous voulez importer ! Trouvez celui que vous voulez, puis cliquez sur le bouton « Open ».

Il y a quelques points que vous pouvez vérifier pour vous assurer que les données sont importées correctement :

- Entêtes. La première ligne du fichier contient-elle les noms de chaque variable - une ligne d'entête - ? Le fichier booksales.csv a un entête, donc c'est un oui.
- Séparateur. Quel caractère est utilisé pour séparer les différentes entrées ? Dans la plupart des fichiers csv, il s'agit d'une virgule (après tout, cela s'appelle un « comma separated value »).

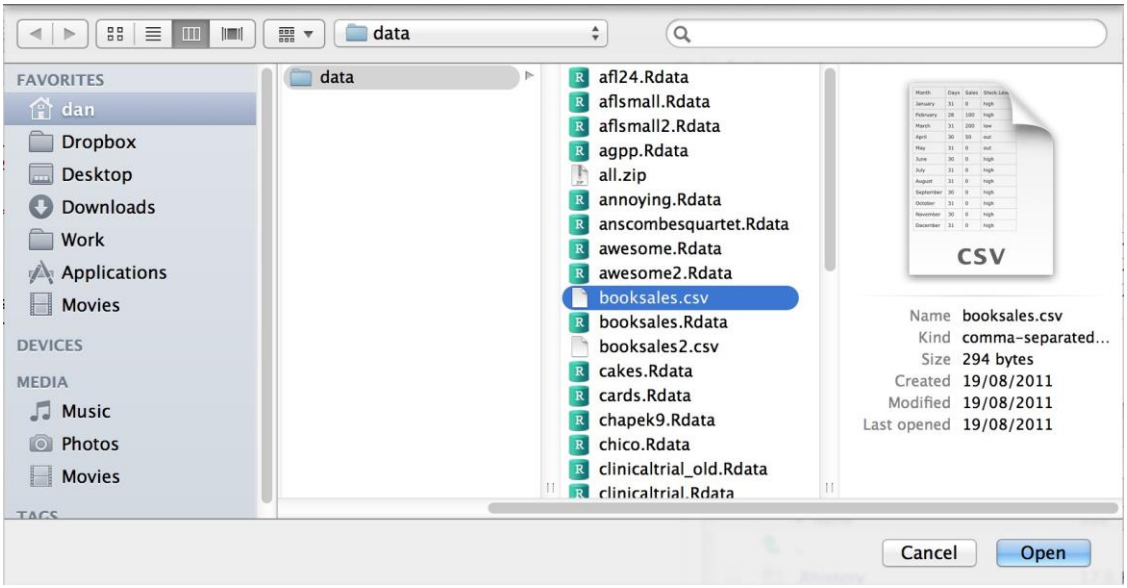


Figure 3-4: Une boîte de dialogue sur un Mac vous demandant de sélectionner le fichier csv que Jamovi devrait essayer d'importer. Les utilisateurs de Mac le reconnaîtront immédiatement, c'est la manière habituelle dont un Mac vous demande de trouver un fichier. Les utilisateurs de Windows ne verront pas cela, au lieu de cela, ils verront la fenêtre habituelle de l'explorateur que Windows vous donne toujours quand il veut que vous sélectionniez un fichier.

- Décimale. Quel caractère est utilisé pour spécifier le point décimal ? Dans les pays anglophones, il s'agit presque toujours d'un point (c.-à-d., .). Ce n'est pas universellement vrai, cependant, beaucoup de pays européens utilisent une virgule.
- Guillemet. Quel caractère est utilisé pour désigner un bloc de texte ? Il s'agit généralement d'un guillemet double (“). C'est pour le fichier booksales.csv.

Importation de fichiers de données inhabituels

Tout au long de ce livre, j'ai supposé que vos données sont stockées dans un fichier Jamovi.omv ou dans un fichier csv formaté « correctement ». Cependant, dans la vraie vie, ce n'est pas une hypothèse très plausible, je ferais donc mieux de parler des autres possibilités que vous pourriez rencontrer.

Chargement de données à partir de fichiers texte

La première chose que je dois souligner est que si vos données sont sauvegardées dans un fichier texte mais ne sont pas *tout à fait* dans le bon format csv alors il y a de forte chance que Jamovi soit en mesure de l'ouvrir. Vous devriez juste essayer et voir si ça marche. Parfois, vous aurez cependant besoin de changer une partie de la mise en forme. Ceux que j'ai souvent eu besoin de changer sont :

- entête. La plupart du temps, lorsque vous stockez des données dans un fichier csv, la première ligne contient en fait les noms des colonnes et non les données. Si ce n'est pas le cas, une bonne pratique consiste à ouvrir le fichier csv dans un tableur comme Open Office et d'ajouter la ligne d'en-tête manuellement.

- sep. Comme l'indique le nom « comma separated value », les valeurs d'une ligne d'un fichier csv sont généralement séparées par des virgules. Ce n'est cependant pas universel. En Europe, la virgule décimale s'écrit généralement comme, au lieu de ., par conséquent, il serait un peu gênant de l'utiliser comme séparateur. Il n'est donc pas inhabituel d'utiliser ; au lieu de ; comme séparateur. D'autres fois, c'est un caractère TAB (tabulation) qui est utilisé.

*guillemet. Il est conventionnel dans les fichiers csv d'inclure un caractère de citation (quote) pour les données textuelles. Comme vous pouvez le voir en regardant le fichier booksalesales.csv, il s'agit généralement d'un caractère double quote ("), mais parfois il n'y a pas de caractère de citation du tout, ou vous pouvez voir un seul quote ' utilisé à la place.

- saut. Il est en fait très courant de recevoir des fichiers csv dans lesquels les premières lignes n'ont rien à voir avec les données réelles. Au lieu de cela, ils fournissent un résumé lisible par l'homme de l'origine des données, ou peut-être qu'ils incluent des informations techniques qui n'ont pas de rapport avec les données.

- valeurs manquantes. Souvent, vous obtiendrez des données avec des valeurs manquantes. Pour une raison ou une autre, certaines entrées du tableau sont manquantes. Le fichier de données doit inclure une valeur « spéciale » pour indiquer que l'entrée est manquante. Par défaut, Jamovi suppose que cette valeur est NA[19], à la fois pour les données numériques et textuelles, vous devez donc vous assurer que, si nécessaire, toutes les valeurs manquantes dans le fichier csv sont remplacées par NA (ou tout autre valeur, selon votre choix) avant d'ouvrir / importer le fichier dans

---

[19] Vous pouvez changer la valeur par défaut pour les valeurs manquantes dans Jamovi à partir du menu en haut à droite (trois points verticaux), mais cela ne fonctionne que lors de l'importation des fichiers de données dans Jamovi. La valeur manquante par défaut dans l'ensemble de données ne doit pas être un nombre valide associé à l'une des variables, par exemple, vous pouvez utiliser -9999 car il est peu probable qu'il s'agisse d'une valeur valide. NdT. Dans la version que j'ai utilisée, la valeur par defaut est NA. J'ai donc corrigé le texte original qui au lieu de NA indique une valeur par défaut 99.

Jamovi. Une fois que vous avez ouvert / importé le fichier dans Jamovi, toutes les valeurs manquantes sont converties en cellules vides dans le tableur Jamovi.

Chargement des données à partir de SPSS (et d'autres progiciels statistiques)

Les commandes listées ci-dessus sont les principales dont nous aurons besoin pour les fichiers de données de ce livre. Mais dans la vie réelle, nous avons beaucoup plus de possibilités. Par exemple, vous pouvez vouloir lire des fichiers de données à partir d'autres programmes de statistiques. Comme SPSS est probablement le logiciel de statistiques le plus utilisé en psychologie, il est à noter que Jamovi peut également importer des fichiers de données SPSS (extension de fichier.sav). Suivez simplement les instructions ci-dessus pour savoir comment ouvrir un fichier csv, mais cette fois-ci naviguez jusqu'au fichier.sav que vous voulez importer. Pour les fichiers SPSS, Jamovi considérera toutes les valeurs comme manquantes si elles sont considérées comme des fichiers « system missing » dans SPSS. La valeur « Default Missings » ne semble pas fonctionner comme prévu lors de l'importation de fichiers SPSS, donc soyez en conscient - vous pourriez avoir besoin d'une autre étape : importer le fichier SPSS dans Jamovi, puis l'exporter comme un fichier csv avant de rouvrir dans Jamovi.[Je sais que c'est un peu compliqué, mais ça marche et j'espère que cela sera corrigé dans une version ultérieure de Jamovi.].

C'est à peu près tout, du moins en ce qui concerne SPSS. En ce qui concerne les autres logiciels statistiques, Jamovi peut également ouvrir / importer directement des fichiers SAS et STATA.

Chargement de fichiers Excel

Les fichiers Excel posent un autre problème. Malgré des années à râler après les gens qui m'envoient des données codées dans un format de données propriétaire, je reçois beaucoup de fichiers Excel. Pour manipuler les fichiers Excel, il faut les ouvrir d'abord dans Excel ou dans un autre tableur qui peut manipuler des fichiers Excel, puis exporter les données dans un fichier csv avant d'ouvrir / importer le fichier csv dans Jamovi.

Changement de données d'un niveau à l'autre

Parfois, vous souhaitez modifier le niveau d'une variable. Cela peut arriver pour toutes sortes de raisons. Parfois, lorsque vous importez des données à partir de fichiers, elles peuvent vous parvenir dans le mauvais format. Les nombres sont parfois importés sous forme de valeurs nominales textuelles. Les dates peuvent être importées sous forme de texte. Les valeurs de l'ID participant peuvent parfois être lues en continu : les valeurs nominales peuvent parfois être lues comme si elles étaient ordinales ou continues. Il y a de fortes chances que vous souhaitiez parfois convertir une variable d'un niveau de mesure à un autre. Ou, pour employer le terme correct, vous voulez **contraindre** la variable d'une classe à une autre.

Dans la section 3.3, nous avons vu comment spécifier différents niveaux de variables, et si vous voulez changer le niveau de mesure d'une variable, vous pouvez le faire dans la vue de données Jamovi pour cette variable. Cliquez simplement sur la case à cocher correspondant au niveau de mesure souhaité - continu, ordinal ou nominal.

Installer les modules d'extension dans Jamovi

Une fonctionnalité vraiment géniale de Jamovi est la possibilité d'installer des modules complémentaires à partir de la bibliothèque Jamovi. Ces modules complémentaires ont été développés par la communauté Jamovi, c'est-à-dire par les utilisateurs et développeurs de Jamovi qui ont créé des modules complémentaires spéciaux qui font d'autres analyses, généralement plus avancées, qui vont au-delà des capacités du programme Jamovi de base.

Pour installer les modules complémentaires, il suffit de cliquer sur le grand « + » en haut à droite de la fenêtre Jamovi, de sélectionner « Jamovi-library » puis de parcourir les différents modules complémentaires disponibles. Choisissez celui (ceux) que vous voulez, puis installez-les, comme dans la Figure 3-5. C'est aussi simple que ça. Les modules nouvellement installés sont alors accessibles depuis la barre de boutons « Analyses ». Essayez-les… Parmi les modules complémentaires utiles à installer vous trouverez « scatr » (ajouté sous « « Descriptives ») et « R j ».

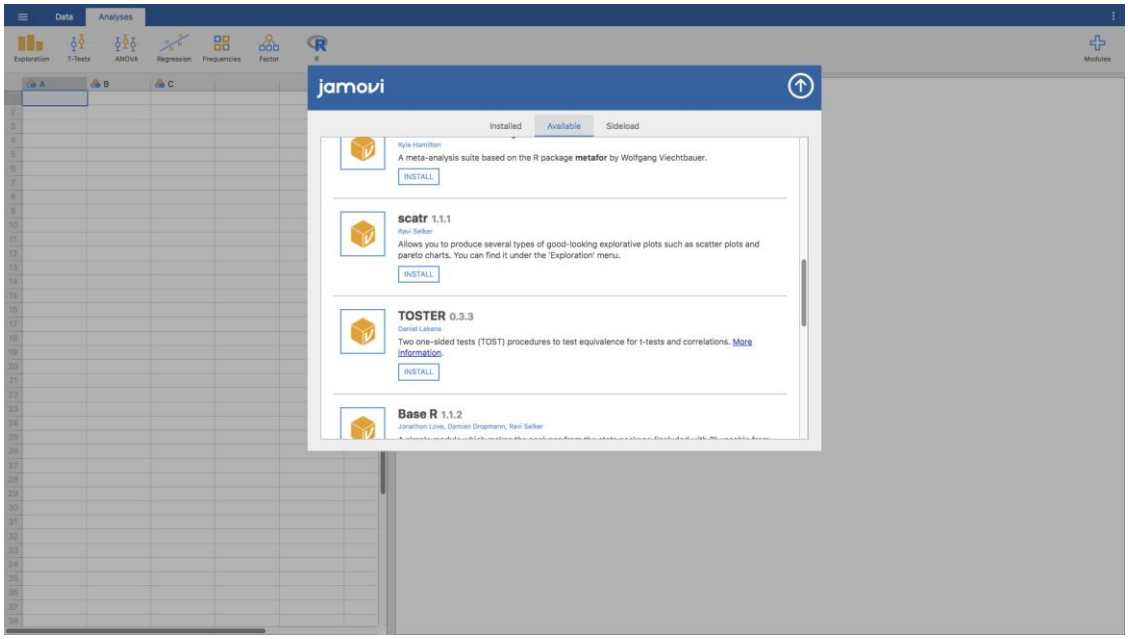


Figure 3-5: Installation des modules d'extension dans Jamovi

Quitter Jamovi

Il y a une dernière chose que je devrais aborder dans ce chapitre : comment quitter Jamovi. Ce n'est pas difficile, fermez simplement le programme de la même manière que n'importe quel autre programme. Cependant, ce que vous voudrez peut-être faire avant de cesser de quitter, c'est sauvegarder votre travail ! Il y a deux parties pour cela : la sauvegarde de toute modification apportée à l'ensemble de données et la sauvegarde des analyses que vous avez exécutées.

Il est recommandé de sauvegarder toute modification apportée à l'ensemble de données sous la forme d'un *nouvel* ensemble de données. De cette façon, vous pouvez toujours revenir aux données d'origine. Pour enregistrer tout changement dans Jamovi, sélectionnez

« Export »… « Data » dans le menu principal Jamovi (bouton avec trois barres horizontales en haut à gauche) et créez un nouveau nom de fichier pour l'ensemble de données modifié.

Alternativement, vous pouvez sauvegarder les données modifiées et les analyses que vous avez effectuées en les sauvegardant dans un fichier Jamovi. Pour ce faire, à partir du menu principal Jamovi sélectionnez « Save as » et tapez un nom de fichier pour ce fichier Jamovi (.omv). N'oubliez pas d'enregistrer le fichier dans un endroit où vous pourrez le retrouver plus tard. Je crée habituellement un nouveau dossier pour des ensembles de données et des analyses spécifiques.

Résumé

Chaque livre qui essaie d'enseigner un nouveau logiciel statistique aux novices doit couvrir à peu près les mêmes sujets, et dans le même ordre. La nôtre ne fait pas exception, et dans la grande tradition qui consiste à faire comme tout le monde l'a fait, ce chapitre a couvert les sujets suivants :

- Section 3.1.. Nous avons téléchargé et installé Jamovi, et l'avons démarré.
- Section 3.2. Nous nous sommes très brièvement orientés vers la partie de Jamovi où les analyses sont faites et les résultats apparaissent, mais nous avons ensuite reporté cela à plus tard dans le livre.
- Section 3.3. Nous avons passé plus de temps à examiner la partie tableur de Jamovi, à considérer différents types de variables, et comment calculer de nouvelles variables.
- Section 3.4. Nous avons aussi vu comment charger des fichiers de données dans Jamovi.
- Section 3.5. Ensuite, nous avons trouvé comment ouvrir d'autres fichiers de données, à partir de différents types de fichiers.
- Section 3.6. Et j'ai vu que parfois nous avons besoin de contraindre les données d'un type à l'autre.
- Section 3.7. L'installation de modules additionnels de la communauté Jamovi étend vraiment les capacités de Jamovi.
- Section 3.8. Enfin, nous avons examiné les bonnes pratiques en termes de sauvegarde de votre ensemble de données et d'analyses lorsque vous avez terminé et êtes sur le point de quitter Jamovi.

Nous ne sommes toujours pas parvenus à quelque chose qui ressemble à une analyse de données. Peut-être que le prochain Chapitre nous rapprochera un peu plus !

Statistiques descriptives

Chaque fois que vous obtenez un nouveau jeu de données à examiner, l'une des premières tâches que vous avez à faire est de trouver des moyens de résumer les données d'une manière synthétique et facile à comprendre. C'est à cela que servent les **statistiques descriptives** (par opposition aux statistiques inférentielles). En fait, pour beaucoup de gens, le terme « statistiques » est synonyme de statistiques descriptives. C'est ce sujet que nous allons aborder dans ce chapitre, mais avant d'entrer dans les détails, prenons un moment pour comprendre pourquoi nous avons besoin de statistiques descriptives. Pour ce faire, ouvrons le fichier aflsmall margins et voyons quelles variables sont stockées dans le fichier.

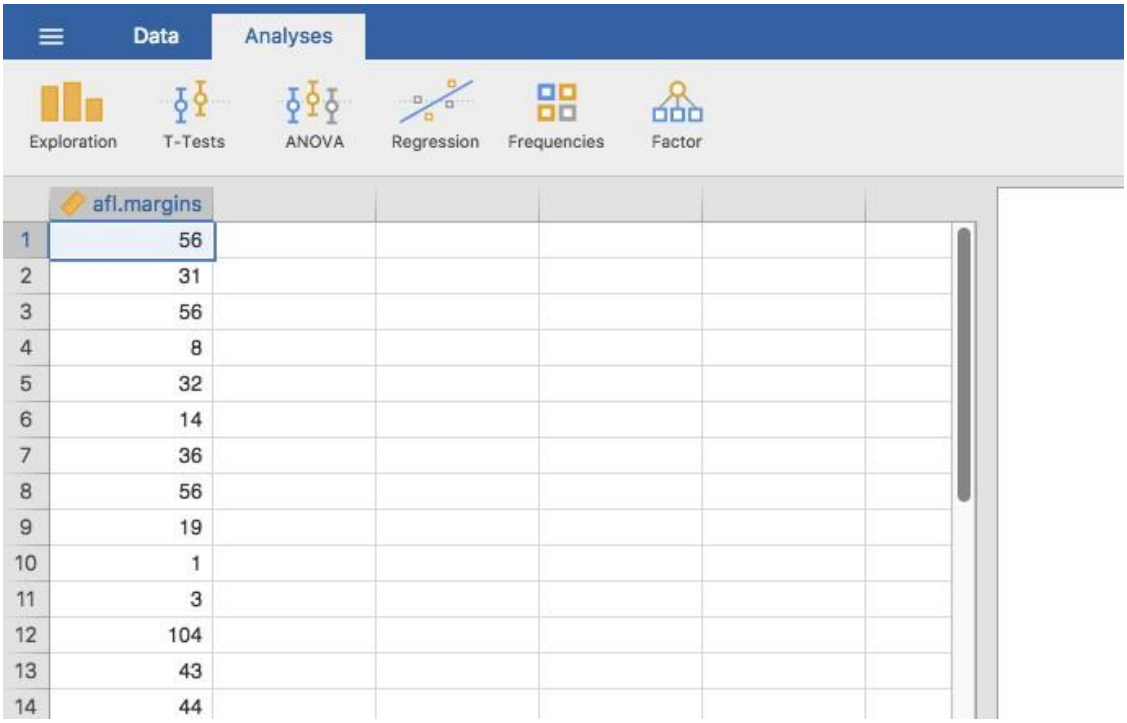


Figure 4-1 Une capture d'écran de Jamovi montrant les variables stockées dans le fichier aflsmal_margins.csv

En fait, il n'y a qu'une seule variable ici, les afl.margins Nous allons nous concentrer un peu sur cette variable dans ce chapitre et je vais préciser de quoi il s'agit. Contrairement à la plupart des ensembles de données de ce livre, il s'agit en fait de données réelles, relatives à la Ligue australienne de football (AFL).[20] La variable afl.margins contient la marge gagnante (nombre de points) pour les 176 matchs joués à domicile et à l'extérieur durant la saison 2010.

---

[20] Note pour les non-Australiens : l'AFL est une compétition de football aux règles australiennes. Vous n'avez pas besoin de connaître les règles australiennes pour suivre cette section.

Ce résultat ne permet pas de se faire une idée de ce que les données disent réellement. Le simple fait de « regarder les données » n'est pas une façon très efficace de comprendre les données. Afin d'avoir une idée de ce que les données nous disent réellement, nous devons calculer quelques statistiques descriptives (ce chapitre) et dessiner quelques belles images (chapitre 5). Puisque les statistiques descriptives sont le plus facile des deux sujets, je vais commencer par celles-ci, cependant je vais vous montrer un histogramme des données afl.margins puisqu'il devrait vous aider à avoir une idée de ce à quoi ressemblent les données que nous essayons de décrire, (voir Figure 4-2). Nous parlerons plus en détail de la façon de dessiner des histogrammes dans la section 5.1. Pour l'instant, il suffit de regarder l'histogramme et de noter qu'il fournit une représentation assez interprétable des données des marges afl.

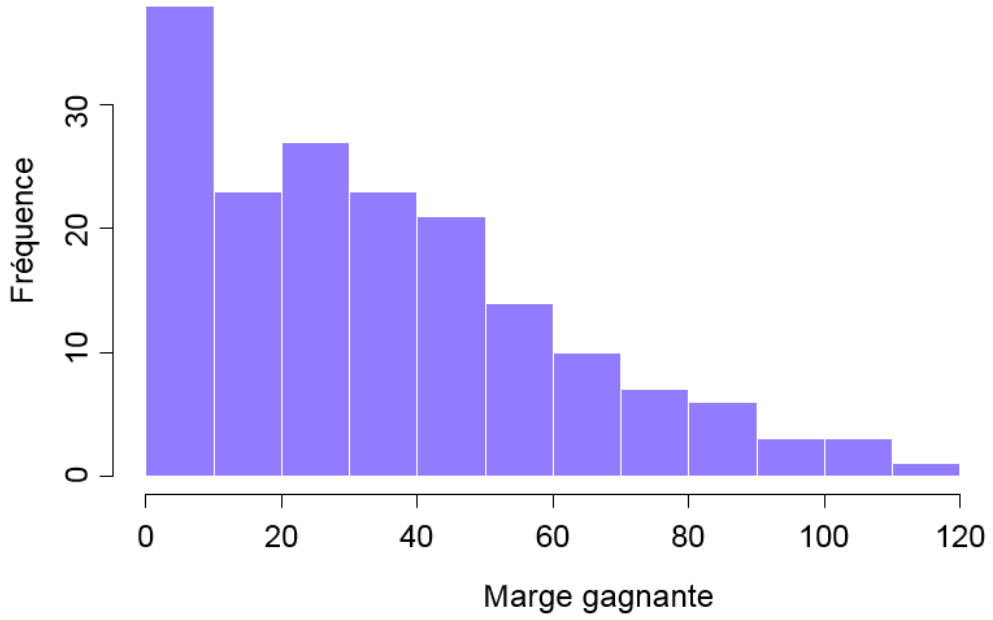


Figure 4-2 : Un histogramme des données de la marge gagnante AFL 2010 (la variable afl.margins). Comme vous pouvez vous y attendre, plus la marge gagnante est grande, moins vous avez tendance à la voir fréquemment.

Mesures de la tendance centrale

Dessiner des graphiques des données, comme je l'ai fait dans la Figure 4-2, est une excellente façon de donner un aperçu de ce que les données tentent de vous dire. Il est souvent extrêmement utile d'essayer de condenser les données en quelques statistiques « sommaires » simples. Dans la plupart des situations, la première chose que vous voudrez calculer est une mesure de la **tendance centrale**. En d'autres termes, vous aimeriez savoir où se situe la « moyenne » ou le « milieu » de vos données. Les trois mesures les plus couramment utilisées sont la moyenne, la médiane et le mode. J'expliquerai chacun d'eux à tour de rôle, puis je discuterai de l'utilité de chacun d'entre eux.

La moyenne

La **moyenne** d'un ensemble d'observations n'est qu'une moyenne normale et classique. Additionnez toutes les valeurs, puis divisez-les par le nombre total de valeurs. Les cinq premières marges gagnantes de l'AFL étaient 56, 31, 56, 8 et 32, de sorte que la moyenne de ces observations est juste :

$$\frac{56 + 31 + 56 + 8 + 32}{5} = \frac{183}{5} = 36{,}60$$

Bien sûr, cette définition de la moyenne n'est nouvelle pour personne. Les valeurs moyennes (c.-à-d. les moyennes) sont utilisées si souvent dans la vie de tous les jours que cela en fait d'une notion assez familière. Cependant, comme le concept de moyenne est quelque chose que tout le monde comprend déjà, je vais m'en servir comme excuse pour commencer à introduire une partie de la notation mathématique que les statisticiens utilisent pour décrire ce calcul, et parler de la façon dont les calculs seraient effectués dans Jamovi.

La première notation à introduire est N, que nous utiliserons pour faire référence au nombre d'observations que nous faisons la moyenne (dans ce cas $N = 5$). Ensuite, nous devons apposer une étiquette sur les observations elles-mêmes. Il est traditionnel d'utiliser X pour cela, et d'utiliser des indices pour indiquer de quelle observation il s'agit. C'est-à-dire, nous utiliserons X_1 pour faire référence à la première observation, X_2 pour faire référence à la deuxième observation, et ainsi de suite jusqu'à X_N pour la dernière. Ou, pour dire la même chose d'une manière un peu plus abstraite, nous utilisons X_i pour faire référence à la *i-ème* observation. Juste pour être sûr d'être clair sur la notation, le tableau suivant énumère les 5 observations de la variable afl.margins, ainsi que le symbole mathématique utilisé pour s'y référer et la valeur réelle à laquelle l'observation correspond :

Bien, maintenant essayons d'écrire une formule pour la moyenne. Par tradition, nous utilisons $\overline{X}$ comme notation de la moyenne. Le calcul de la moyenne pourrait donc être exprimé à l'aide de la formule suivante :

$$\overline{X} = \frac{X_1 + X_2 + X_2 + \cdots X_{N-1} + X_N}{N}$$

Cette formule est tout à fait correcte, mais elle est terriblement longue, c'est pourquoi nous utilisons le **symbole de sommation** $\sum$ pour la raccourcir[21]. Si je veux additionner les cinq

[21] Le choix d'utiliser $\sum$ pour indiquer la sommation n'est pas arbitraire. C'est la lettre grecque en majuscules sigma, qui est l'analogue de la lettre S dans cet alphabet. De même, il y a un symbole équivalent utilisé pour désigner la multiplication de beaucoup de nombres, parce que les multiplications sont aussi appelées « produits » nous utilisons le symbole Π pour cela (le grec en majuscule pi, qui est l'analogue de la lettre P).

premières observations, je pourrais écrire la somme de façon longue, $X_1 + X_2 + X_3 + X_4 + X_5$ ou je pourrais utiliser le symbole de somme pour l'abréger comme ceci :

$$\sum_{i=1}^{5} X_i$$

Prise littéralement, cela pourrait se lire comme « la somme, prise sur toutes les *i* valeurs de 1 à 5, de la valeur *Xi* ». Mais au fond, cela signifie qu'il faut « additionner les cinq premières observations ». Dans tous les cas, nous pouvons utiliser cette notation pour écrire la formule de la moyenne, qui ressemble à ceci :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

En toute honnêteté, je ne peux pas imaginer que toute cette notation mathématique aide à clarifier le concept de moyenne. En fait, c'est juste une façon d'écrire la même chose que ce que j'ai dit avec des mots : additionnez toutes les valeurs, puis divisez-les par le nombre total d'éléments. Cependant, ce n'est pas vraiment la raison pour laquelle je suis entré dans tous ces détails. Mon but était d'essayer de m'assurer que tous ceux qui lisent ce livre soient clairs sur la notation que nous utiliserons tout au long du livre : $\overline{X}$ pour la moyenne, $\sum$ pour l'idée de sommation, X_i pour la i-ème observation et N pour le nombre total d'observations. Nous allons réutiliser ces symboles un peu, alors il est important que vous les compreniez assez bien pour pouvoir « lire » les équations et voir qu'il s'agit simplement de dire « additionnez beaucoup de choses et divisez-les par une autre chose ».

Calcul de la moyenne avec Jamovi

Bien, c'est des maths. Alors, comment pouvons-nous obtenir que la boîte informatique magique fasse le travail pour nous ? Lorsque le nombre d'observations commence à augmenter, il est beaucoup plus facile d'effectuer ce genre de calcul à l'aide d'un ordinateur. Pour calculer la moyenne en utilisant toutes les données, nous pouvons utiliser Jamovi. La première étape consiste à cliquer sur le bouton « Exploration », puis sur « Descriptives ». Ensuite, vous pouvez mettre en surbrillance la variable afl.margins et cliquer sur la flèche vers la droite pour la déplacer dans la boîte « Variables ». Après avoir fait cela, un tableau apparaît sur le côté droit de l'écran contenant les informations par défaut « Descriptives » ; voir Figure 4-3.

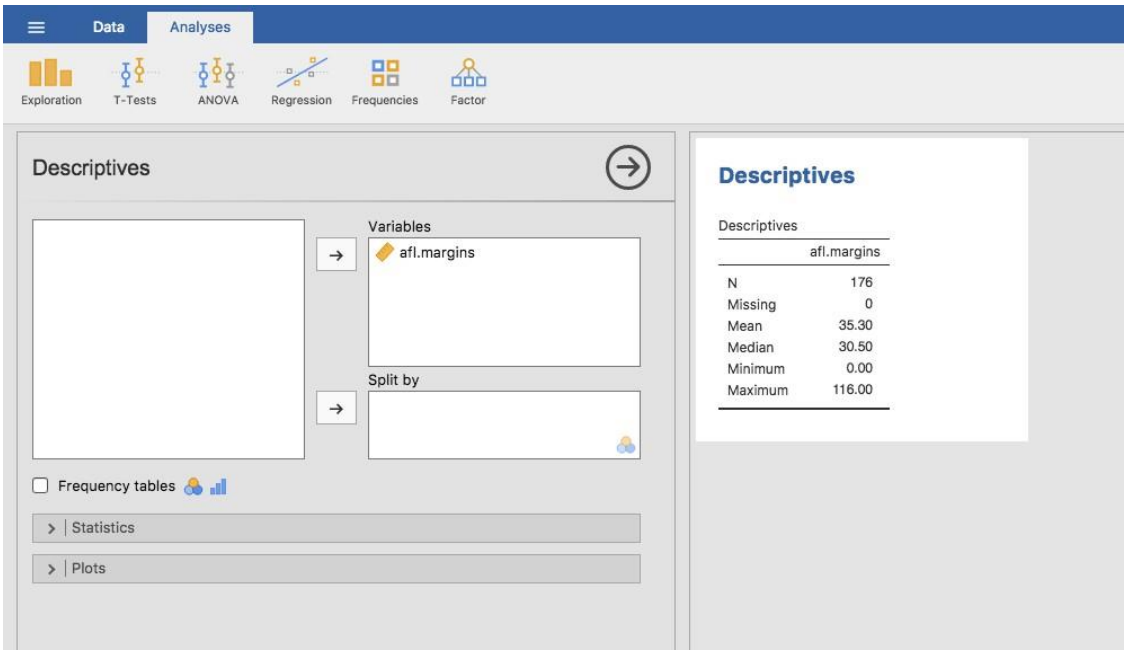


Figure 4-3 Descriptifs par défaut des données de la marge gagnante AFL 2010 (la variable afl.margins).

Comme vous pouvez le voir à la Figure 4-3, la valeur moyenne de la variable afl.margins est de 35,30. Les autres renseignements présentés comprennent le nombre total d'observations (N=176), le nombre de valeurs manquantes (aucune) et les valeurs médiane, minimale et maximale pour la variable.

La médiane

La deuxième mesure de la tendance centrale que les gens utilisent beaucoup est la **médiane**, et elle est encore plus facile à décrire que la moyenne. La médiane d'un ensemble d'observations n'est que la valeur centrale. Comme auparavant, imaginons que nous n'étions intéressés que par les 5 premières marges gagnantes de l'AFL : 56, 31, 56, 8 et 32. Pour déterminer la médiane, nous trions ces nombres par ordre croissant :

8,31,**32**,56,56

En inspectant les données, il est évident que la valeur médiane de ces 5 observations est de 32 puisque c'est celle du milieu dans la liste triée (je l'ai mise en gras pour la mettre en évidence). C'est facile. Mais que faire si nous sommes intéressés par les 6 premiers jeux plutôt que par les 5 premiers ? Puisque le sixième match de la saison avait une marge gagnante de 14 points, notre liste triée est maintenant la suivante :

8,14,**31**,**32**,56,*56

et il y a *deux* nombres intermédiaires, 31 et 32. La médiane est définie comme la moyenne de ces deux nombres, qui est bien sûr de 31,5. Comme auparavant, c'est très fastidieux de le faire à la main quand on a beaucoup de chiffres. Dans la vraie vie, bien sûr, personne ne calcule réellement la médiane en triant les données et en cherchant ensuite la valeur

moyenne. Dans la vraie vie, nous utilisons un ordinateur pour faire le travail pénible pour nous, et Jamovi nous a fourni une valeur médiane de 30,50 pour la variable afl.margins (Figure 4-3).

Moyenne ou médiane ? Quelle est la différence ?

Savoir calculer les moyennes et les médianes n'est qu'une partie de l'histoire. Vous devez également comprendre ce que chacun dit au sujet des données, et ce que cela implique au moment où vous devez utiliser chacune d'elles. C'est ce qu'illustre la Figure 4-4. La moyenne est un peu comme le « centre de gravité » de l'ensemble de données, alors que la médiane est la « valeur centrale » des données. Ce que ceci implique, pour ce qui est de savoir lequel vous devriez utiliser, dépend un peu du type de données que vous possédez et de ce que vous essayez de faire. A titre indicatif :

- Si vos données sont dans une échelle nominale, vous ne devriez probablement pas utiliser la moyenne ou la médiane. La moyenne et la médiane reposent toutes deux sur l'idée que les nombres attribués aux valeurs sont significatifs. Si la numérotation est arbitraire, il est probablement préférable d'utiliser le mode (Section 4.1.6) à la place.

- Si vos données sont sur une échelle ordinale, vous êtes plus susceptible de privilégier la médiane que la moyenne. La médiane n'utilise que les informations de classement de vos données (c'est-à-dire les chiffres les plus grands) mais ne dépend pas des nombres précis en cause. C'est exactement la situation qui s'applique lorsque vos données sont à l'échelle ordinale. La moyenne, par contre, utilise les valeurs numériques précises attribuées aux observations, donc elle n'est pas vraiment appropriée pour les données ordinales.

- Pour les données de l'échelle d'intervalles et de rapport, l'une ou l'autre est généralement acceptable. Le choix de celui que vous choisissez dépend un peu de ce que vous essayez de faire. La moyenne a l'avantage d'utiliser toute l'information contenue dans les données (ce qui est utile lorsque vous n'avez pas beaucoup de données). Mais elle est très sensible aux valeurs extrêmes et marginales.

Développons un peu cette dernière partie. L'une des conséquences est qu'il existe des différences systématiques entre la moyenne et la médiane lorsque l'histogramme est asymétrique (asymétrique ; voir section 4.3). C'est ce qu'illustre la Figure 4-4 Notez que la médiane (côté droit) est située plus près du « corps » de l'histogramme, alors que la moyenne (côté gauche) est traînée vers la « queue » (où se trouvent les valeurs extrêmes). Pour donner un exemple concret, supposons que Robert (revenu de 50 000 $), Kate (revenu de 60 000 $) et Jeanne (revenu de 65 000 $) sont assis à une table. Le revenu moyen à la table est de 58 333 $ et le revenu médian est de 60 000 $. Puis Bill s'assoit avec eux (revenu de 100 000 000 $). Le revenu moyen est maintenant passé à 25 043 750 $, mais la médiane n'est que de 62 500 $. Si vous voulez examiner le revenu global à la table, la moyenne pourrait être la bonne réponse. Mais si vous vous intéressez à ce qui est considéré comme un revenu typique à la table, la médiane serait un meilleur choix ici.

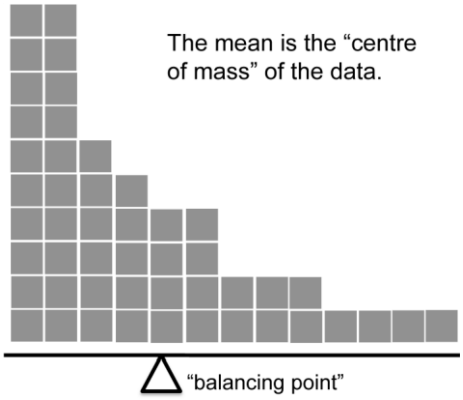


Figure 4-4 : Une illustration de la différence entre l'interprétation de la moyenne et celle de la médiane. La moyenne est essentiellement le « centre de gravité » de l'ensemble de données. Si vous imaginez que l'histogramme des données est un objet solide, alors le point sur lequel vous pouvez l'équilibrer (comme sur une bascule) est la moyenne. Par contre, la médiane est l'observation du milieu, la moitié des observations étant plus petites et la moitié plus grandes.

Un exemple concret

Pour essayer de comprendre pourquoi vous devez prêter attention aux différences entre la moyenne et la médiane, considérons un exemple réel. Comme j'ai tendance à me moquer des journalistes pour leurs faibles connaissances scientifiques et statistiques, je dois rendre à César ce qui est à César. Voici un excellent article sur le site d'ABC news[22] du 24 septembre 2010 :

Au cours des deux dernières semaines, des cadres supérieurs de la Commonwealth Bank ont parcouru le monde avec une présentation montrant que les prix des maisons en Australie et les principaux ratios prix-revenus se comparent avantageusement à ceux de pays similaires. « En fait, l'accessibilité à la propriété a dérapé au cours des cinq ou six dernières années », a déclaré Craig James, économiste en chef de la division commerciale de la banque, CommSec.

C'est probablement une énorme surprise pour quiconque a un prêt hypothécaire, ou qui veut un prêt hypothécaire, ou qui paie un loyer, ou qui n'est pas complètement inconscient de ce qui se passe sur le marché australien du logement depuis plusieurs années. Retournons à l'article :

L'ABC a mené sa guerre contre ce qu'elle croit être des prophètes de malheur avec des graphiques, des chiffres et des comparaisons internationales. Dans sa présentation, la banque rejette les arguments selon lesquels le logement en Australie est relativement cher par rapport aux revenus. Il indique que le ratio du prix des maisons par rapport au revenu des ménages, qui est de 5,6 dans les grandes villes et de 4,3 à l'échelle nationale, est

[22] www.abc.net.au/news/stories/2010/09/24/3021480.htm

comparable à celui de nombreux autres pays développés. Il est dit que San Francisco et New York ont des ratios de 7, Auckland est à 6,7 et Vancouver à 9,3.

Encore une excellente nouvelle ! Sauf que l'article poursuit en disant que :

De nombreux analystes disent que cela a conduit la banque à utiliser des chiffres trompeurs et des comparaisons. Si vous allez à la page 4 de l'exposé de l'ABC et que vous lisez l'information de la source au bas du graphique et du tableau, vous remarquerez qu'il y a une autre source sur la comparaison internationale - Demographia. Toutefois, si la Commonwealth Bank avait également utilisé l'analyse de Demographia sur le ratio prix/revenu des maisons en Australie, elle aurait obtenu un chiffre plus proche de 9 plutôt que 5,6 ou 4,3.

C'est, euh, un écart assez sérieux. Un groupe de personnes dit 9, un autre dit 4-5. Devrions-nous simplement couper en deux la différence et dire que la vérité se situe quelque part entre les deux ? Absolument pas ! C'est une situation où il y a une bonne et une mauvaise réponse. La démographie est correcte, et la Banque du Commonwealth a tort. Comme le souligne l'article :

[Un] problème évident avec les chiffres des prix intérieurs de la Banque du Commonwealth par rapport au revenu est qu'elle compare les revenus moyens aux prix médians des maisons (contrairement aux chiffres démographiques qui comparent les revenus médians aux prix médians). La médiane est le point central, ce qui signifie que la moyenne est généralement plus élevée lorsqu'il s'agit des revenus et des prix des actifs, car elle inclut les revenus des personnes les plus riches de l'Australie. En d'autres termes, les chiffres de la Commonwealth Bank comptent le salaire de plusieurs millions de dollars de Ralph Norris du côté des revenus, mais pas sa maison (sans doute) très chère dans les chiffres du prix de l'immobilier, ce qui sous-estime le ratio prix/revenu des maisons pour les Australiens à revenu moyen.

Je n'aurais pas pu mieux dire. La façon dont Demographia a calculé le ratio est la bonne. La façon dont la Banque l'a fait est incorrecte. Quant à savoir pourquoi une organisation extrêmement sophistiquée sur le plan quantitatif, comme une grande banque, a commis une erreur aussi élémentaire, eh bien… Je ne peux pas le dire avec certitude puisque je n'ai aucune idée précise de ce qu'ils pensent. Mais l'article lui-même mentionne les faits suivants, qui peuvent ou non être pertinents :

En tant que premier prêteur immobilier australien, la Banque du Commonwealth a l'un des intérêts les plus importants dans la hausse des prix de l'immobilier. Elle possède en effet une grande partie des logements australiens en garantie de ses prêts immobiliers ainsi que de nombreux prêts aux petites entreprises.

Mon Dieu, mon Dieu.

Mode

Le mode d'un échantillon est très simple. C'est la valeur qui s'observe le plus fréquemment. Nous pouvons illustrer le mode en utilisant une variable AFL différente : qui a joué le plus de finales ? Ouvrez le fichier des petits finalistes aflsmall et jetez un coup d'œil à la variable afl.finalists, voir Figure 4-5. Cette variable contient les noms des 400 équipes qui ont participé aux 200 matches de la phase finale disputés entre 1987 et 2010.

Ce que nous *pourrions* faire, c'est lire l'ensemble des 400 inscriptions et compter le nombre d'occasions où chaque nom d'équipe apparaît dans notre liste de finalistes, produisant ainsi un **tableau de fréquence**. Cependant, ce serait stupide et ennuyeux : exactement le genre de tâche pour laquelle les ordinateurs sont très doués. Alors utilisons Jamovi pour faire ça

pour nous. Sous « Exploration » - « Descriptives », cliquez sur la petite case à cocher intitulée « Frequency table » et vous devriez obtenir quelque chose comme Figure 4-6.

Maintenant que nous avons notre tableau de fréquence, nous pouvons le regarder et constater qu'au cours des 24 années pour lesquelles nous disposons de données, Geelong a participé à plus de finales que toute autre équipe. Ainsi, le mode des données des données afl.finalistes est « Geelong ». On constate que Geelong (39 finales) a disputé plus de finales que toute autre équipe au cours de la période 1987-2010. Il convient également de noter que dans le tableau des statistiques descriptives, aucun résultat n'est calculé pour la moyenne, la médiane, le minimum ou le maximum. C'est parce que la variable afl.finalists est une variable nominale et que cela n'a pas de sens de calculer ces valeurs.

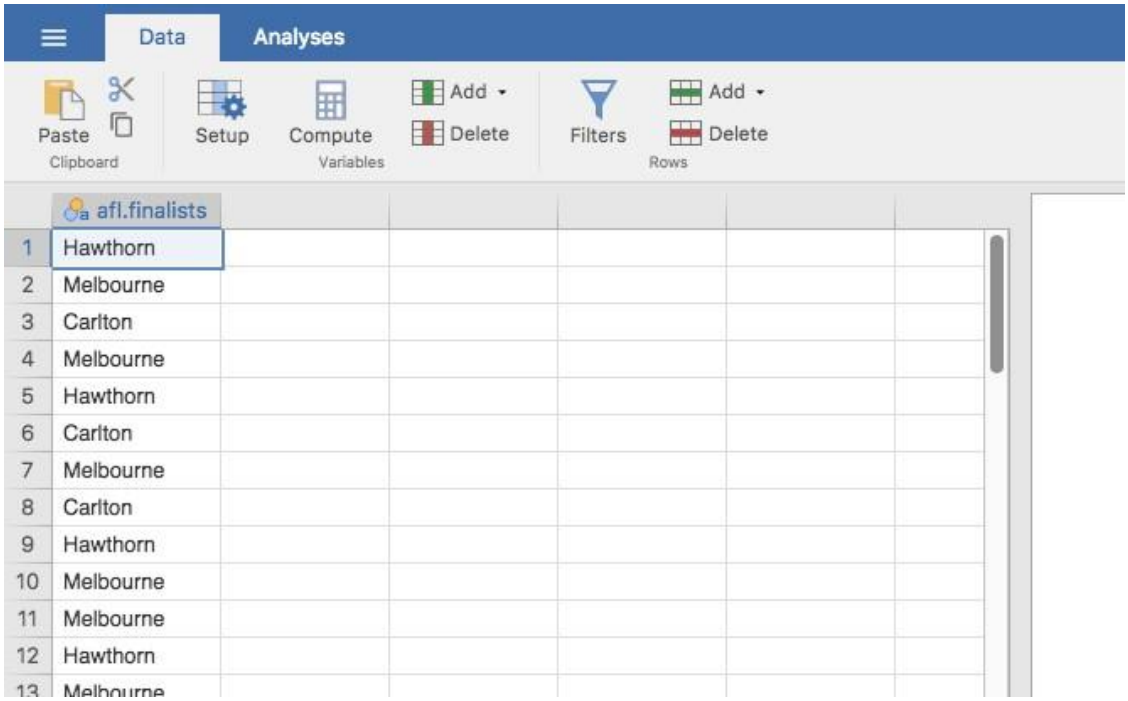


Figure 4-5 Une capture d'écran de Jamovi montrant la variable stockée dans le fichier aflsmall_finalists.csv

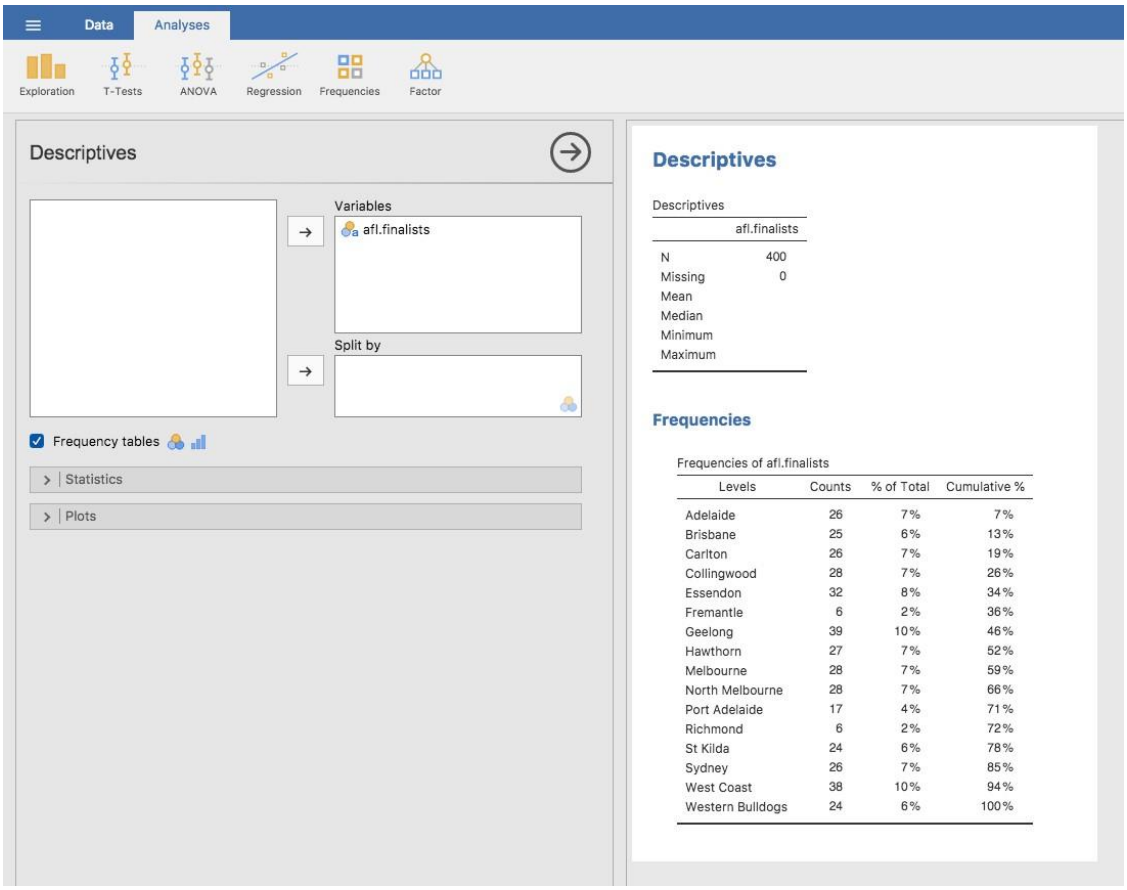


Figure 4-6 : Une capture d'écran de Jamovi montrant la table de fréquence pour la variable afl.finalists

Une dernière remarque concernant le mode. Bien que le mode soit le plus souvent calculé lorsque vous disposez de données nominales, parce que les moyennes et les médianes sont inutiles pour ce genre de variables, il y a des situations dans lesquelles vous voulez vraiment connaître le mode d'une variable ordinale, d'intervalle ou de rapport. Par exemple, revenons à notre variable afl.margins. Cette variable est clairement une échelle de ratio (si vous ne comprenez pas bien, il peut être utile de relire la section 2.2) et, dans la plupart des cas, la moyenne ou la médiane est la mesure de la tendance centrale que vous voulez. Mais considérez ce scénario : un de vos amis propose un pari et il choisit un match de football au hasard. Sans savoir qui joue, vous devez deviner la marge de gain *exacte*. Si vous devinez correctement, vous gagnez 50 $. Si vous ne le faite pas, vous perdez 1 $, il n'y a pas de prix de consolation pour avoir « presque » obtenu la bonne réponse. Vous devez deviner exactement la bonne marge. Pour ce pari, la moyenne et la médiane vous sont complètement inutiles. C'est le mode sur lequel vous devriez parier. Pour calculer le mode de la variable afl.margins dans Jamovi, retournez à cet ensemble de données et sur l'écran « Exploration » -« Descriptives » vous verrez que vous pouvez développer la section marquée « Statistics». Cliquez sur la case à cocher « Mode » et vous verrez la valeur modale apparaitre dans le tableau « Descriptives », comme dans la Figure 4-7. Les données de 2010 suggèrent donc que vous devriez miser sur une marge de 3 points.

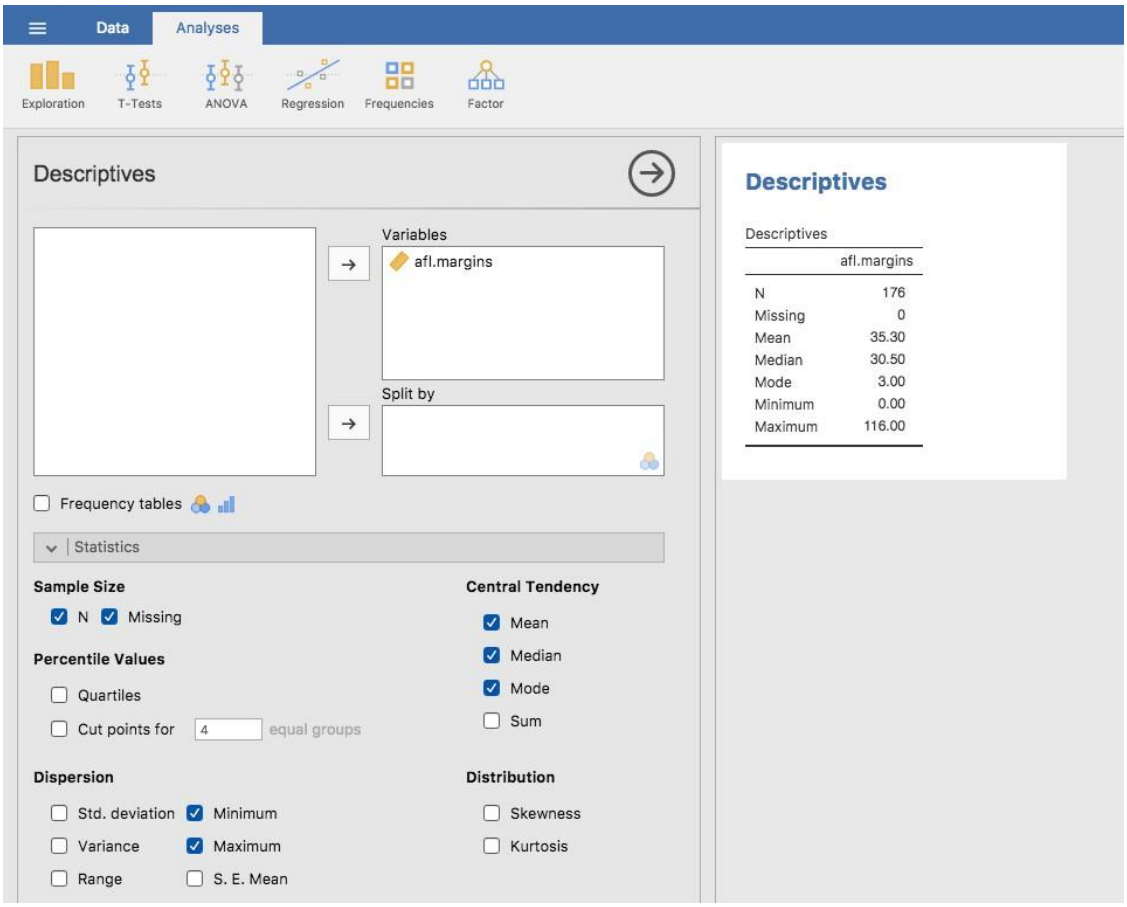


Figure 4-7 : Une capture d'écran de Jamovi montrant la valeur modale de la variable afl.margins

Mesures de la variabilité

Les statistiques dont nous avons discuté jusqu'à présent portent toutes sur la *tendance centrale*. C'est-à-dire qu'ils parlent tous des valeurs qui sont « au milieu » ou « populaires » dans les données. Cependant, la tendance centrale n'est pas le seul type de statistique sommaire que nous voulons calculer. La deuxième chose que nous voulons vraiment, c'est une mesure de la **variabilité des données**. En d'autres termes, comment les données sont-elles « étalées » ? A quelle distance de la moyenne ou de la médiane les valeurs observées ont-elles tendance à être ? Pour l'instant, supposons que les données sont des échelles d'intervalles ou de ratios, et nous continuerons à utiliser les données de marges afl. Nous utiliserons ces données pour discuter de différentes mesures de propagation, chacune ayant des forces et des faiblesses différentes.

L'étendue

L'étendue d'une variable est très simple. C'est la plus grande valeur moins la plus petite valeur. Pour les données de marges gagnantes de l'AFL, la valeur maximale est de 116 et la valeur minimale est de 0. Bien que la plage soit le moyen le plus simple de quantifier la notion de « variabilité », c'est l'une des plus mauvaises. Rappelons-nous, dans notre

discussion sur la moyenne, que nous voulons que notre mesure sommaire soit robuste. Si l'ensemble de données contient une ou deux valeurs extrêmement mauvaises, nous aimerions que nos statistiques ne soient pas indûment influencées par ces cas. Par exemple, dans une variable contenant des valeurs aberrantes très extrêmes

-100,2,3,4,5,6,7,8,9,10

Il est clair que l'étendue n'est pas robuste. Cette variable a une étendue de 110, mais si l'on supprimait la valeur aberrante, nous n'aurions qu'une étendue de 8.

Écart interquartile

L'**écart interquartile** (IQR) est comme l'écart, mais au lieu de la différence entre la plus grande et la plus petite valeur, on prend la différence entre le 25e et le 75e percentile. Si vous ne savez pas déjà ce qu'est un **percentile**, le 10e percentile d'un ensemble de données est le plus petit nombre *x de* sorte que 10 % des données sont inférieures à *x*. En fait, nous avons déjà trouvé l'idée. La médiane d'un ensemble de données est son 50e centile ! Dans Jamovi, vous pouvez facilement spécifier les 25e, 50e et 75e percentiles en cochant la case « Quartiles » dans l'écran « Exploration » « Descriptives » - « Statistics ».

Il n'est donc pas surprenant que, dans la Figure 4-8, le 50e percentile soit le même que la valeur médiane. Et, en notant que 50,50 - 12,75 = 37,75, nous pouvons voir que l'écart interquartile des marges gagnantes de l'AFL 2010 est de 37,75. Bien que l'écart soit évident à interpréter, il est un peu moins évident d'interpréter l'IQR. La façon la plus simple de la penser est la suivante : l'intervalle interquartile est l'intervalle couvert par la « moitié centrale » des données.

Descriptives

Descriptives

	afl.margins
N	176
Missing	0
Mean	35.30
Median	30.50
Mode	3.00
Minimum	0.00
Maximum	116.00
25th percentile	12.75
50th percentile	30.50
75th percentile	50.50

Figure 4-8 : Une capture d'écran de Jamovi montrant les quartiles de la variable afl. Margins

Autrement dit, un quart des données se situent sous le 25e percentile et un quart des données se situent au-dessus du 75e percentile, laissant la « moitié médiane » des données entre les deux. Et l'IQR est la plage couverte par cette moitié médiane.

Écart moyen absolu

Les deux mesures que nous avons examinées jusqu'ici, l'intervalle et l'intervalle interquartile, reposent toutes deux sur l'idée que nous pouvons mesurer la dispersion des données en examinant les percentiles des données. Cependant, ce n'est pas la seule façon de penser le problème. Une autre approche consiste à sélectionner un point de référence significatif (habituellement la moyenne ou la médiane), puis à signaler les écarts « types » par rapport à ce point de référence. Qu'entend-on par écart « typique » ? Habituellement, il s'agit de la valeur moyenne ou médiane de ces écarts. Dans la pratique, cela conduit à deux mesures différentes : « l'écart absolu moyen » (par rapport à la moyenne) et « l'écart absolu médian » (par rapport à la médiane). D'après ce que j'ai lu, la mesure fondée sur la médiane semble être utilisée dans les statistiques et semble être la meilleure des deux. Mais pour être honnête, je ne pense pas l'avoir vu beaucoup utilisé en psychologie. La mesure basée sur la moyenne apparaît cependant parfois en psychologie. Dans cette section, je parlerai de la première, et je reviendrai sur la deuxième plus tard.

Puisque le paragraphe précédent peut sembler un peu abstrait, examinons un peu plus posément l'**écart absolu moyen** par rapport à la moyenne. Ce qui est utile à propos de cette mesure est que le nom vous dit exactement comment la calculer. Pensons à nos données de marges gagnantes AFL, et une fois de plus, nous allons commencer par prétendre qu'il n'y a que 5 données au total, avec des marges gagnantes de 56, 31, 56, 8 et 32. Puisque nos calculs reposent sur un examen de l'écart par rapport à un point de référence (dans ce cas la moyenne), la première chose que nous devons calculer est la moyenne, $\overline{X}$. Pour ces cinq observations, notre moyenne est $\overline{X} = 36.6$. L'étape suivante consiste à convertir chacune de nos observations X_i en un score de déviation. Pour ce faire, nous calculons la différence entre l'observation X_i et la moyenne $\overline{X}$. C'est-à-dire que le score d'écart est défini comme étant $X_i - \overline{X}$. Pour la première observation de notre échantillon, cela correspond à 56 - 36.6 = 19.4. Bien, c'est assez simple. L'étape suivante du processus consiste à convertir ces écarts en écarts absolus, et nous le faisons en convertissant toute valeur négative en valeur positive. Mathématiquement, nous désignerions la valeur absolue de -3 comme |- 3|, et donc nous disons que |- 3|= 3. Nous utilisons la valeur absolue ici parce que nous ne nous soucions pas vraiment de savoir si la valeur est supérieure à la moyenne ou inférieure à la moyenne, nous voulons simplement savoir si elle est *proche* de la moyenne. Pour rendre ce processus aussi évident que possible, le tableau ci-dessous montre ces calculs pour les cinq observations :

En français	quel jeu	valeur	écart à la moyenne	écart absolu
Notation	i	X_i	$X_i - \overline{X}$	$
	1	56	19,4	19,4
	2	31	-5,6	5,6
	3	56	19,4	19,4

4	8	-28,6	28,6
5	32	-4,6	4,6

Maintenant que nous avons calculé le score d'écart absolu pour chaque observation sur l'ensemble de données, tout ce que nous avons à faire pour calculer la moyenne de ces scores. C'est ce qu'on va faire :

$$\frac{19{,}4 + 5{,}6 + 19{,}4 + 28{,}6 + 4{,}6}{5} = 15{,}52$$

Et c'est fini. L'écart absolu moyen pour ces cinq notes est de 15,52.

Cependant, bien que nos calculs pour ce petit exemple soient terminés, il nous reste quelques points à aborder. D'abord, nous devrions vraiment essayer d'écrire une formule mathématique appropriée. Mais pour ce faire, j'ai besoin d'une notation mathématique pour me référer à l'écart absolu moyen. Irritant, « écart moyen absolu » et « écart médian absolu » ont le même acronyme (en anglais MAD), ce qui conduit à un source d'ambiguïté, donc j'ai ferais mieux de trouver quelque chose de différent pour l'écart moyen absolu. Soupir. Ce que je vais faire, c'est d'utiliser AAD à la place, abréviation anglaise de déviation absolue moyenne (average absolute deviation). Maintenant que nous avons une notation non ambiguë, voici la formule qui formalise ce que nous venons de calculer :

$$\text{AAD}(X) = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \overline{X}\right|$$

Variance

Bien que la mesure de l'écart absolu moyen ait son utilité, ce n'est pas la meilleure mesure de la variabilité à utiliser. D'un point de vue purement mathématique, il y a de bonnes raisons de préférer les écarts au carré aux écarts absolus. Si nous faisons cela, nous obtenons une mesure appelée **variance**, qui a très bonnes propriétés statistiques que je vais ignorer,[23] et un énorme défaut psychologique dont je vais faire toute une histoire dans un instant. La variance d'un ensemble de données X est parfois écrite $Var(X)$, mais elle est plus communément appelée s^2 (la raison en sera bientôt plus claire). La formule que nous utilisons pour calculer la variance d'un ensemble d'observations est la suivante :

---

[23] Bien, je vais mentionner très brièvement celle que je trouve la plus cool, avec une définition très particulière du mot « cool ». Les variances sont *additives*. Voici ce que ça veut dire. Supposons que j'ai deux variables X et Y, dont les variances sont respectivement *Var(X)* et *Var(Y)*. Imaginez maintenant que je veuille définir une nouvelle variable Z qui est la somme des deux, Z = X + Y. Il s'avère que la variance de Z est égale à Var(X) + Var(Y). C'est une propriété très utile et qui n'est pas vraie des autres mesures dont je présente dans cette section.

$$\text{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

Comme vous pouvez le voir, c'est essentiellement la même formule que celle que nous avons utilisée pour calculer l'écart absolu moyen, sauf qu'au lieu d'utiliser « écarts absolus », nous utilisons « écarts carrés ». C'est pour cette raison que la variance est parfois appelée « écart quadratique moyen ».

Maintenant que nous avons l'idée de base, voyons un exemple concret. Encore une fois, utilisons les cinq premiers résultats de l'AFL comme données. Si nous suivons la même approche que la dernière fois, nous obtenons le tableau suivant :

En français	quel jeu	valeur	écart à la moyenne	écart au carré
Maths	i	X_i	$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$
	1	56	19,4	376,36
	2	31	-5,6	31,36
	3	56	19,4	376,36
	4	8	-28,6	817,96
	5	32	-4,6	21,16

Cette dernière colonne contient tous nos écarts quadratiques, il ne nous reste plus qu'à faire la moyenne. Si nous faisons cela à la main, c'est-à-dire à l'aide d'une calculatrice, nous obtenons une variance de 324,64. Passionnant, n'est-ce pas ? Pour le moment, ignorons la question brûlante que vous vous posez probablement tous (c'est-à-dire ce que signifie une variance de 324,64) et parlons plutôt un peu plus de la façon de faire les calculs dans Jamovi, car cela mettra en évidence quelque chose de très bizarre. Démarrez une nouvelle session Jamovi en cliquant sur le bouton du menu principal (trois lignes horizontales dans le coin supérieur gauche et en sélectionnant « Nouveau ». Saisissez maintenant les cinq premières valeurs de l'ensemble de données des marges afl. dans la colonne A (56, 31, 56, 8, 32). Changez le type de variable en « Continu » et sous « Descriptives » cliquez sur la case à cocher « Variance », vous obtenez les mêmes valeurs de variance que celles que nous avons calculées à la main (324,64). Non, attendez, vous obtenez une réponse complètement *différente* (405,80) - voir Figure 4-9. C'est bizarre. Jamovi est bogué ? C'est une faute de frappe ? Suis-je un idiot ?

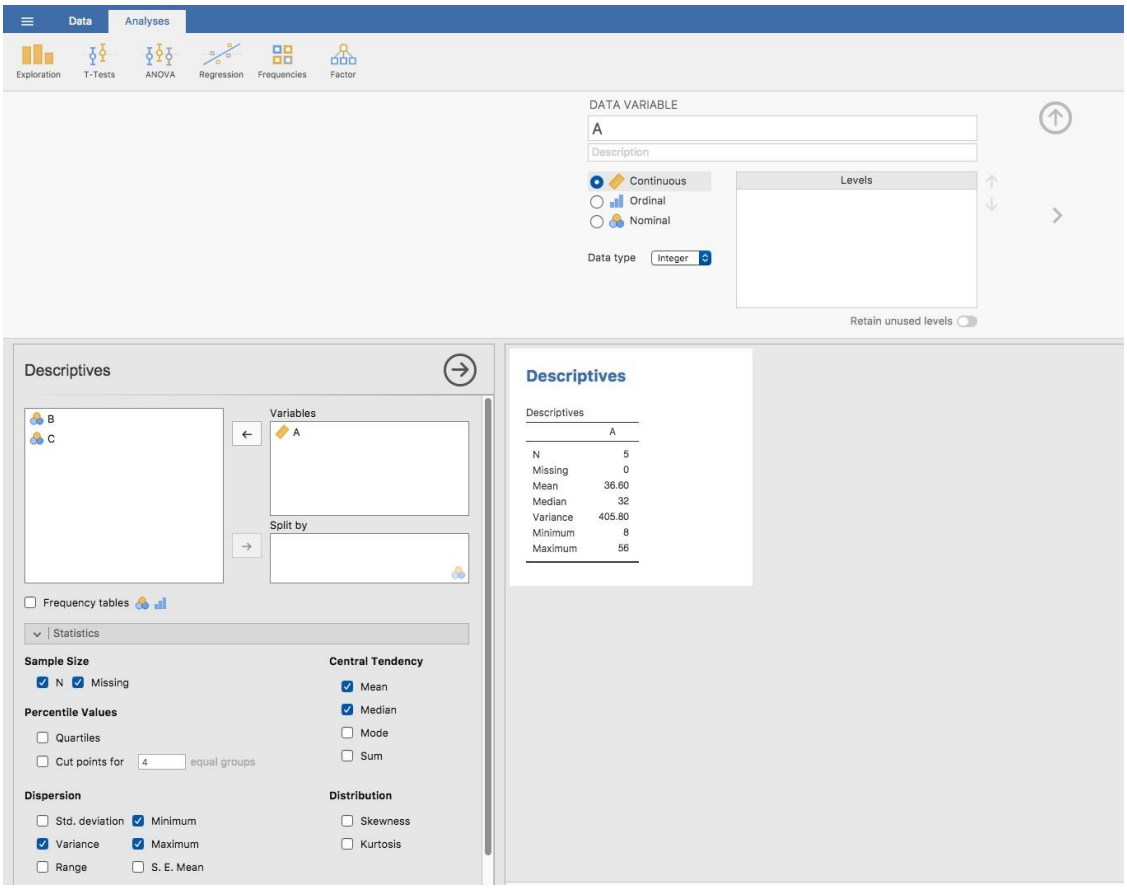


Figure 4-9 : Une capture d'écran de Jamovi montrant la variance pour les 5 premières valeurs de la variable marges afl.

Il se trouve que la réponse est non.[24] Ce n'est pas une faute de frappe, et Jamovi ne fait pas d'erreur. En fait, il est très simple d'expliquer ce que Jamovi fait ici, mais un peu plus difficile d'expliquer *pourquoi* Jamovi le fait. Commençons donc par le « quoi ». Ce que fait Jamovi, c'est évaluer une formule légèrement différente de celle que je vous ai montrée ci-dessus. Au lieu de faire la moyenne des écarts au carré, ce qui vous oblige à diviser par le nombre de points de données *N*, Jamovi a choisi de diviser par *N* - 1. En d'autres termes, la formule que Jamovi utilise est la suivante :

$$\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Voilà donc le *quoi*. La vraie question est *pourquoi* Jamovi divise par *N* - 1 et non par *N*. Après tout, la variance est supposée être l'écart quadratique *moyen*, non ? Ne devrions-nous donc pas diviser par *N*, le nombre réel d'observations dans l'échantillon ? Eh bien, oui, on devrait. Cependant, comme nous le verrons au chapitre 8, il existe une distinction subtile entre

[24] À l'exception peut-être de la troisième question.

« décrire un échantillon » et « faire des suppositions sur la population d'où provient l'échantillon ». Jusqu'à présent, c'était une distinction sans différence. Que vous décriviez un échantillon ou que vous tiriez des conclusions sur la population, la moyenne est calculée exactement de la même façon. Ce n'est pas le cas pour la variance, ni pour l'écart-type, ni pour de nombreuses autres mesures. Ce que je vous ai décrit au départ (c.-à-d. prendre la moyenne réelle, et donc diviser par *N*) suppose que vous avez littéralement l'intention de calculer la variance de l'échantillon. La plupart du temps, cependant, vous n'êtes pas très intéressé par l'échantillon en *soi*. L'échantillon existe plutôt pour vous dire quelque chose sur le monde. Si c'est le cas, vous commencez en fait à vous éloigner du calcul d'une « statistique d'échantillon » pour vous diriger vers l'idée d'estimer un « paramètre de population ». Cependant, je m'avance un peu. Pour l'instant, prenons pour acquis que Jamovi sait ce qu'il fait, et nous reviendrons sur cette question plus tard lorsque nous parlerons d'estimation au chapitre 8.

Bien, une dernière chose. Jusqu'à présent, cette section s'est lue un peu comme un roman policier. Je vous ai montré comment calculer la variance, décrit la chose étrange « $N - 1$ » que fait Jamovi et fait allusion à la raison pour laquelle il le fait, mais je n'ai pas mentionné la chose la plus importante. Comment *interprétez-vous* la variance ? Les statistiques descriptives sont censées décrire les choses, après tout, et pour le moment, la variance n'est qu'un chiffre en charabia. Malheureusement, la raison pour laquelle je ne vous ai pas donné l'interprétation humaine de la variance est qu'il n'y en a pas vraiment. C'est le problème le plus grave de la variance. Bien qu'il possède d'élégantes propriétés mathématiques qui suggèrent qu'il s'agit vraiment d'une quantité fondamentale pour exprimer la variation, il est complètement inutile si vous voulez communiquer avec un humain réel. Les écarts sont totalement ininterprétables par rapport à la variable d'origine ! Tous les chiffres ont été mis au carré et ils ne veulent plus rien dire. C'est un énorme problème. Par exemple, dans le tableau que j'ai présenté tout à l'heure, la marge dans le jeu 1 était « 376,36 points - plus élevée au carré que la marge moyenne ». C'est *exactement* aussi stupide que ça en a l'air, et lorsque nous calculons une variance de 324,64, nous sommes dans la même situation. J'ai regardé beaucoup de matchs de foot, et à aucun moment personne n'a jamais fait référence à des « points carrés ». Ce *n'est pas* une véritable unité de mesure, et puisque la variance est exprimée en termes de cette unité en charabia, elle est totalement dénuée de sens pour un humain.

Écart-type

Supposons que vous aimiez l'idée d'utiliser la variance à cause de ces belles propriétés mathématiques dont je n'ai pas parlé, mais comme vous êtes un humain et non un robot, vous aimeriez avoir une mesure qui est exprimée dans les mêmes unités que les données elles-mêmes (c'est-à-dire des points et non des points carrés). Que devriez-vous faire ? La solution au problème est évidente ! Prenons la racine carrée de la variance, connue sous le nom d'**écart-type**, également appelée «Racine de l'écart quadratique moyen », ou REQM[25]. Cela résout notre problème de façon assez nette. Alors que personne n'a pas la moindre

[25] NdT. En anglais « root mean squared deviation" ou RMSD

idée de ce que signifie réellement « une variance de 324,68 points », il est beaucoup plus facile de comprendre « un écart-type de 18,01 points » puisqu'il est exprimé dans les unités originales. Il est traditionnel de désigner l'écart-type d'un échantillon de données par *s*, bien que « sd » et « std dev » soient également utilisés à l'occasion.

Comme l'écart-type est égal à la racine carrée de la variance, vous ne serez probablement pas surpris de voir que la formule est :

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

et dans Jamovi il y a une case à cocher pour « Standard deviation » juste au-dessus de la case à cocher pour la « Variance ». En sélectionnant cette option, on obtient une valeur de 26,07 pour l'écart-type.

Cependant, comme vous l'avez peut-être noté dans notre discussion sur la variance, ce que Jamovi calcule réellement est légèrement différent de la formule donnée ci-dessus. Tout comme nous l'avons vu avec la variance, ce que Jamovi calcule est une version qui se divise par $N - 1$ plutôt que N.

Pour des raisons qui auront un sens lorsque nous reviendrons sur ce sujet au chapitre 8, je me référerai à cette nouvelle quantité comme $\hat{\sigma}$ (lire : « sigma chapeau »), et la formule pour cela est :

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

L'interprétation des écarts-types est légèrement plus complexe. Puisque l'écart-type est dérivé de la variance, et que la variance est une quantité qui a peu ou pas de sens pour nous les humains, l'écart-type n'a pas une interprétation simple. Par conséquent, la plupart d'entre nous ne se fient qu'à une simple règle empirique. En général, vous devriez vous attendre à ce que 68 % des données se situent à l'intérieur d'un écart-type de la moyenne, 95 % des données se situent à l'intérieur de deux écarts-types de la moyenne et 99,7 % des données se situent à l'intérieur de trois écarts-types de la moyenne. Cette règle a tendance à bien fonctionner la plupart du temps, mais elle n'est pas exacte. Il est en fait calculé en *supposant que* l'histogramme est symétrique et « en forme de cloche ».[26] Comme vous pouvez le voir en regardant l'histogramme des marges gagnantes AFL de la Figure 4-2, ce n'est pas vrai pour nos données ! Malgré cela, la règle est à peu près correcte. Il s'avère que

---

[26] A strictement parler, l'hypothèse est que les données sont *normalement* distribuées, ce qui est un concept important dont nous parlerons plus en détail au chapitre 7 et sur lequel nous reviendrons encore et encore plus loin dans le livre.

65,3 % des données sur les marges AFL se situent à l'intérieur d'un écart-type de la moyenne. Ceci est illustré visuellement à la Figure 4-10.

Quelle mesure utiliser ?

Nous avons discuté d'un certain nombre de mesures de l'écart : la fourchette, l'IQR, l'écart moyen absolu, la variance et l'écart type ; et nous avons fait allusion à leurs forces et à leurs faiblesses. En voici un bref résumé :

- *Etendue*. Vous donne la pleine expansion des données. Elle est très vulnérable aux valeurs aberrantes et en conséquence n'est pas souvent utilisée à moins d'avoir de bonnes raisons de vous soucier des valeurs extrêmes dans les données.
- *Écart interquartile (IQR)*. Indique où se trouve la « moitié centrale » des données. Il est assez robuste et complet bien la médiane. On s'en sert beaucoup.
- *Déviation moyenne absolue*. Indique la distance « en moyenne » entre les observations et la moyenne. Il est très interprétable, mais comporte quelques problèmes mineurs (qui ne sont pas abordés ici) qui le rendent moins attrayant pour les statisticiens que l'écart-type. Utilisé parfois,mais pas souvent.

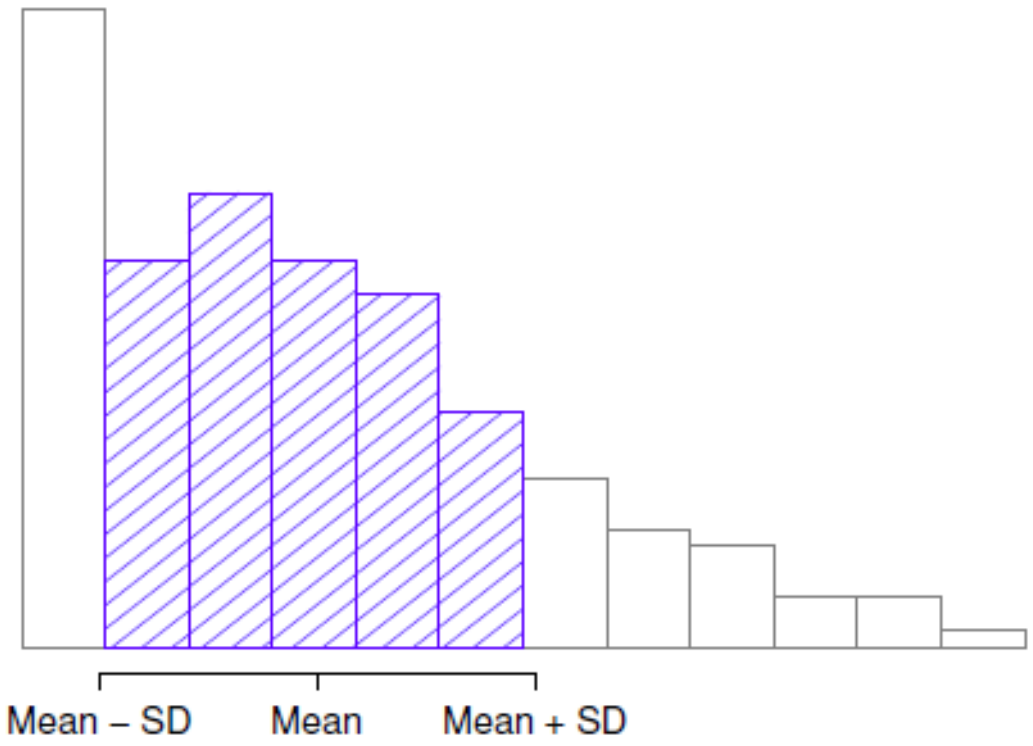


Figure 4-10 : Illustration de l'écart-type à partir des données des marges gagnantes de l'AFL. Les barres ombragées de l'histogramme indiquent la proportion des données qui se situe à l'intérieur d'un écart-type de la moyenne. Dans ce cas, 65,3 % de l'ensemble des

données se situent dans cette fourchette, ce qui est assez conforme à la « règle d'environ 68 % « dont il est question dans le texte principal.

- *Variance*. Vous indique l'écart quadratique moyen par rapport à la moyenne. C'est mathématiquement élégant et c'est probablement la « bonne » façon de décrire la variation autour de la moyenne, mais c'est complètement ininterprétable parce qu'il n'utilise pas les mêmes unités que les données. Presque jamais utilisé sauf comme un outil mathématique, mais il est enfoui « sous le capot » d'un très grand nombre d'outils statistiques.
- *Écart-type*. Il s'agit de la racine carrée de la variance. C'est assez élégant mathématiquement et c'est exprimé dans les mêmes unités que les données, de sorte qu'on peut assez bien l'interpréter. Dans les situations où la moyenne est la mesure de la tendance centrale, c'est la valeur par défaut. C'est de loin la mesure de variation la plus populaire.

En résumé, l'IQR et l'écart-type sont aisément les deux mesures les plus couramment utilisées pour rendre compte de la variabilité des données. Mais il y a des situations où les autres sont utilisées. Je les ai tous décrits dans ce livre parce qu'il y a de fortes chances que vous en rencontriez la plupart quelque part.

Asymétrie et aplatissement

Il y a deux autres statistiques descriptives que vous verrez parfois rapportées dans la littérature psychologique : l'asymétrie (skew) et l'aplatissement (kurtosis). Dans la pratique, ni l'une ni l'autre n'est utilisée aussi fréquemment que les mesures de la tendance centrale et de la variabilité dont nous avons parlé. L'asymétrie est assez importante, donc vous verrez qu'il en est fait mention assez souvent, mais je n'ai jamais vu d'applatissement rapporté dans un article scientifique à ce jour.

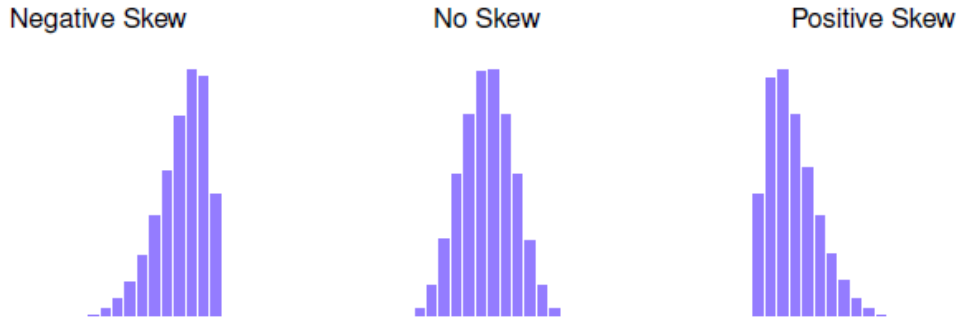


Figure 4-11 : Une illustration de l'asymétrie (skew). Sur la gauche nous avons un ensemble de données négatives (skewness = -.93), au milieu nous avons un ensemble de données sans asymétrie (enfin, presque pas : skewness =.-.006), et sur la droite nous avons un ensemble de données positivement asymétriques (skewness = .93).

Puisque c'est le plus intéressant des deux, commençons par parler de l'**asymétrie**. L'asymétrie est fondamentalement une mesure de l'asymétrie et la façon la plus facile de l'expliquer est de dessiner quelques images. Comme l'illustre la Figure 4-11, si les données ont tendance à avoir beaucoup de valeurs extrêmement petites (c.-à-d. que la queue inférieure est « plus longue « que la queue supérieure) et pas autant de valeurs extrêmement grandes (panneau de gauche), nous disons que les données sont *faussées de façon négative*. Par contre, s'il y a des valeurs plus importantes que des valeurs extrêmement faibles (panneau de droite), nous disons que les données sont *positivement faussées*. C'est l'idée qualitative derrière l'asymétrie. S'il y a relativement plus de valeurs qui sont beaucoup plus grandes que la moyenne, la distribution est positivement inclinée ou inclinée vers la droite, avec une queue qui s'étend vers la droite. Le biais négatif ou gauche est le contraire. Une distribution symétrique a une asymétrie de 0, la valeur d'asymétrie pour une distribution positivement asymétrique est positive, et une valeur négative pour une distribution négativement asymétrique.

Une formule pour l'asymétrie d'un ensemble de données est la suivante

$$\text{skewness}(X) = \frac{1}{N\,\hat{\sigma}^3} \sum_{i=1}^{N} (X_i - \bar{X})^3$$

où N est le nombre d'observations, $\bar{X}$ est la moyenne de l'échantillon et $\hat{\sigma}$ est l'écart-type (la version « divisé par N – 1 »).

Peut-être pour vous aider, vous pourriez utiliser Jamovi pour calculer l'asymétrie : c'est une case à cocher dans les options « Statistiques » sous « Exploration » - « Descriptifs ». Pour la variable afl.margins, l'asymétrie est de 0,780. Si vous divisez l'estimation de l'asymétrie par l'erreur standard d'asymétrie, vous obtenez une indication de l'asymétrie des données. Surtout dans les petits échantillons (N<50), une règle empirique suggère qu'une valeur de 2 ou moins peut signifier que les données ne sont pas très asymétriques, et une valeur de plus de 2 qu'il y a suffisamment de biais dans les données pour éventuellement limiter leur utilisation dans certaines analyses statistiques. Il n'y a cependant pas d'accord clair sur cette interprétation. Malgré tout cela indique que les données sur les marges gagnantes de l'AFL sont quelque peu asymétriques (0,780 / 0,183 = 4,262).

La dernière mesure à laquelle on se réfère parfois, mais très rarement dans la pratique, est **l'aplatissement** d'un ensemble de données. En termes simples, l'aplatissement est une mesure de « l'acuité » d'un ensemble de données, comme l'illustre la Figure 4-12. Par convention, on dit que la « courbe normale » (lignes noires) a une kurtosis nulle, de sorte que le point d'un ensemble de données est évalué par rapport à cette courbe.

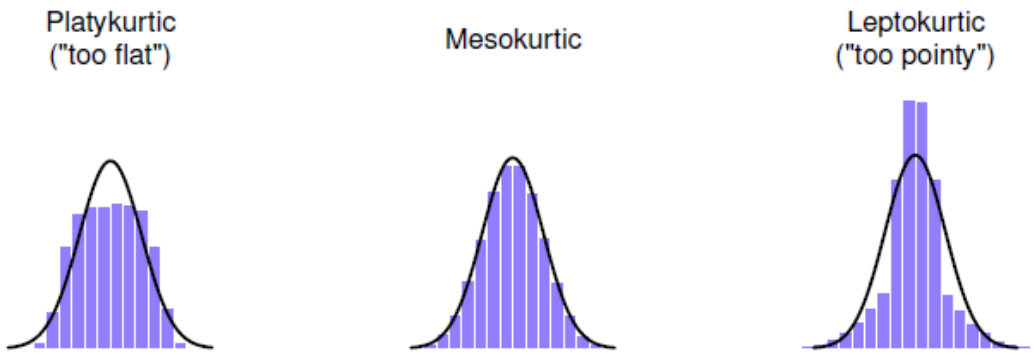


Figure 4-12 : Une illustration de l'aplatissement. A gauche, nous avons un ensemble de données « platykurtique » (kurtosis = -.95) qui signifie que l'ensemble de données est « trop plat ». Au milieu, nous avons un ensemble de données « mésokurtiques » (kurtosis est presque égal à 0), ce qui signifie que la précision des données est à peu près correcte. Enfin, à droite, nous avons un ensemble de données « leptokurtiques » (kurtosis = 2.12) indiquant que l'ensemble de données est « trop pointu ». Notez que l'aplatissement est mesuré par rapport à une courbe normale (ligne noire).

Dans cette figure, les données à gauche ne sont pas assez pointues, donc le kurtosis est négatif et nous appelons les données *platykurtiques*. Les données à droite sont trop pointues, donc l'aplatissement est positif et nous disons que les données sont *leptokurtiques*. Mais les données au milieu sont juste assez pointues, donc nous disons qu'il est *mésokurtique* et a un kurtosis égal à zéro. Ce point est résumé dans le tableau ci-dessous :

Terme informel	Nom technique	Valeur du kurtosis
trop plat	platykurtique	négatif
Juste assez pointu	mésokurtique	zero
trop pointu	leptokurtique	positif

L'équation de l'aplatissement est assez semblable de conception aux formules que nous avons déjà vues pour la variance et l'asymétrie. Sauf que lorsque la variance impliquait des écarts au carré et que l'asymétrie impliquait des écarts au cube, l'aplatissement impliquait d'augmenter les écarts à la puissance quatre[27] :

$$\text{kurtosis}(X) = \frac{1}{N\hat{\sigma}^4}\sum_{i=1}^{N}(X_i - \bar{X})^4 - 3$$

[27] La partie « -3 » est quelque chose que les statisticiens collent pour s'assurer que la courbe normale a un kurtosis zéro. Cela semble un peu stupide, il suffit de coller un « -3 » à la fin de la formule, mais il y a de bonnes raisons mathématiques de le faire.

Je sais, ce n'est pas très intéressant pour moi non plus.

Plus précisément, Jamovi a une case à cocher pour l'aplatissement juste en dessous de la case à cocher pour l'asymétrie, ce qui donne une valeur pour l'aplatissement de 0,101 avec une erreur standard de 0,364. Cela signifie que les données de marges gagnantes AFL sont juste assez pointues.

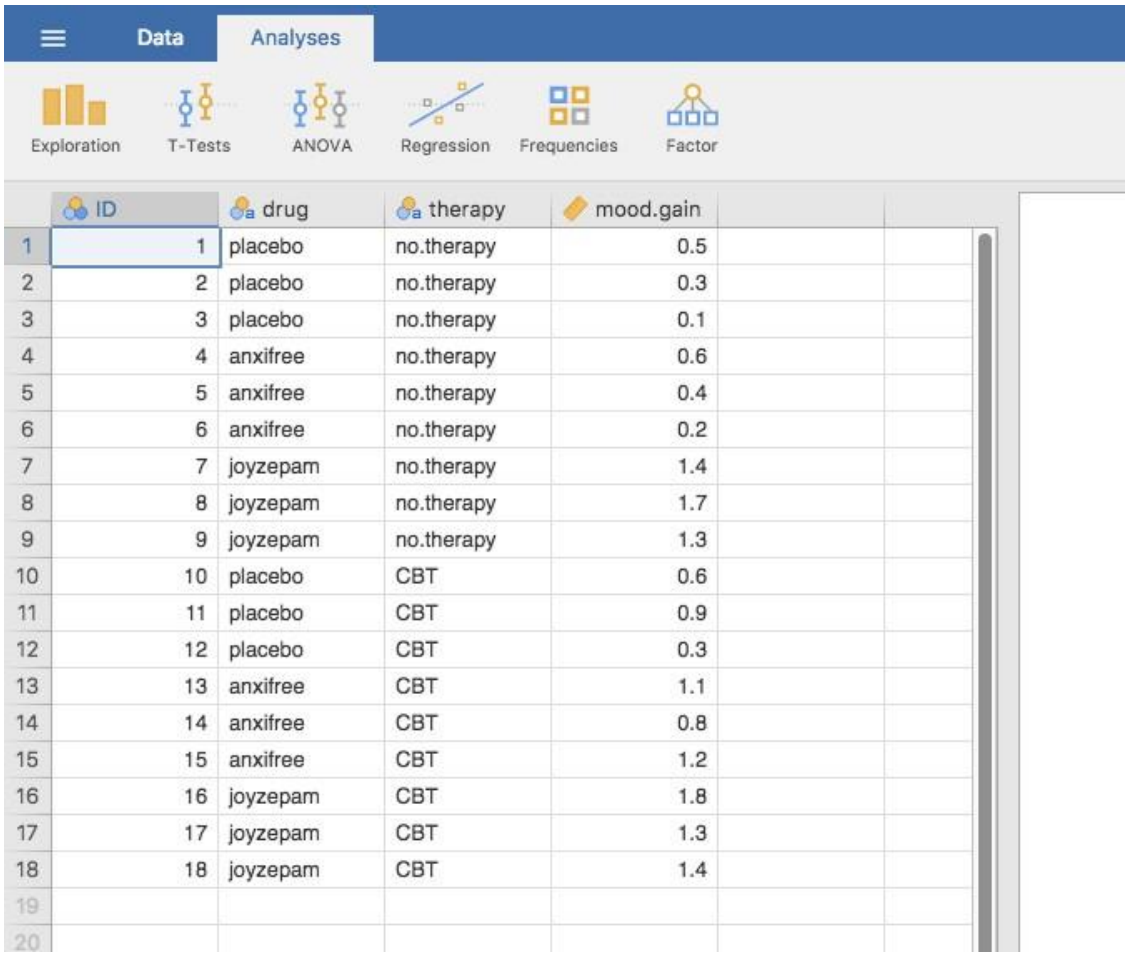
Statistiques descriptives distinctes pour chaque groupe

Il est très fréquent que vous ayez besoin d'examiner des statistiques descriptives ventilées par variable de regroupement. C'est assez facile à faire avec Jamovi. Par exemple, supposons que je veux examiner les statistiques descriptives de certaines données de clin.trial, ventilées séparément par type de thérapie. Il s'agit d'un nouvel ensemble de données que vous n'avez jamais vu auparavant. Les données sont stockées dans le fichier clinicaltrial.csv et nous les utiliserons beaucoup au chapitre 13 (vous trouverez une description complète des données au début de ce chapitre). Chargeons-le et voyons ce qu'on a :

Évidemment, il y avait trois médicaments : un placebo, quelque chose appelé « anxifree » et quelque chose appelé « joyzepam », et il y avait 6 personnes qui recevaient chaque médicament. Neuf personnes ont été traitées par thérapie cognitivo-comportementale (TCC) et neuf personnes n'ont reçu aucun traitement psychologique. Et nous pouvons voir en regardant les « Descriptives » de la variable mood.gain que la plupart des gens ont montré un gain d'humeur (moyenne = 0.88), bien que sans savoir quelle est l'échelle ici, il est difficile d'en dire beaucoup plus que cela. Mais ce n'est pas si mal. Dans l'ensemble, j'ai l'impression d'avoir appris quelque chose avec cela.

Nous pouvons également examiner d'autres statistiques descriptives, et cette fois-ci séparément pour chaque type de thérapie. Dans Jamovi, cochez « Std deviation », « Skewness » et « Kurtosis » dans les options « Statistics ». En même temps, faite glisser la variable de therapy dans la case « Split by »[28], et vous devriez obtenir quelque chose comme Figure 4-14

[28] NdT « Split by » signifie « divisé par »



	ID	drug	therapy	mood.gain
1	1	placebo	no.therapy	0.5
2	2	placebo	no.therapy	0.3
3	3	placebo	no.therapy	0.1
4	4	anxifree	no.therapy	0.6
5	5	anxifree	no.therapy	0.4
6	6	anxifree	no.therapy	0.2
7	7	joyzepam	no.therapy	1.4
8	8	joyzepam	no.therapy	1.7
9	9	joyzepam	no.therapy	1.3
10	10	placebo	CBT	0.6
11	11	placebo	CBT	0.9
12	12	placebo	CBT	0.3
13	13	anxifree	CBT	1.1
14	14	anxifree	CBT	0.8
15	15	anxifree	CBT	1.2
16	16	joyzepam	CBT	1.8
17	17	joyzepam	CBT	1.3
18	18	joyzepam	CBT	1.4
19				
20				

Figure 4-13 : Une capture d'écran de Jamovi montrant les variables stockées dans le fichier clinicaltrial.csv

Qu'arrive-t-il si vous avez plusieurs variables de regroupement ? Supposons que vous souhaitiez examiner séparément le gain moyen de l'humeur pour toutes les combinaisons possibles de médicaments et de traitements. Il est possible de le faire en ajoutant une autre variable, le médicament (Drug), dans la case « Split by ». Facile, bien que parfois, si vous divisez trop, il n'y a pas assez de données dans chaque combinaison de décomposition pour faire des calculs significatifs. Dans ce cas, Jamovi vous le dit en disant quelque chose comme NaN ou Inf.[29]

---

[29] Parfois, Jamovi présentera aussi des chiffres d'une manière inhabituelle. Si un nombre est très petit, ou très grand, alors Jamovi passe à une forme exponentielle pour les nombres. Par exemple, 6.51e-4 revient à dire que le point décimal est déplacé de 4 positions vers la gauche, donc le nombre réel est 0.000651. S'il y a un signe plus (c'est-à-dire 6.51e+4), la virgule décimale est déplacée vers la droite, c'est-à-dire 65 100,00. Habituellement, seuls des nombres très petits ou très grands sont exprimés de cette façon, par exemple 6.51e-16, ce qui serait très difficile à écrire de la manière habituelle.

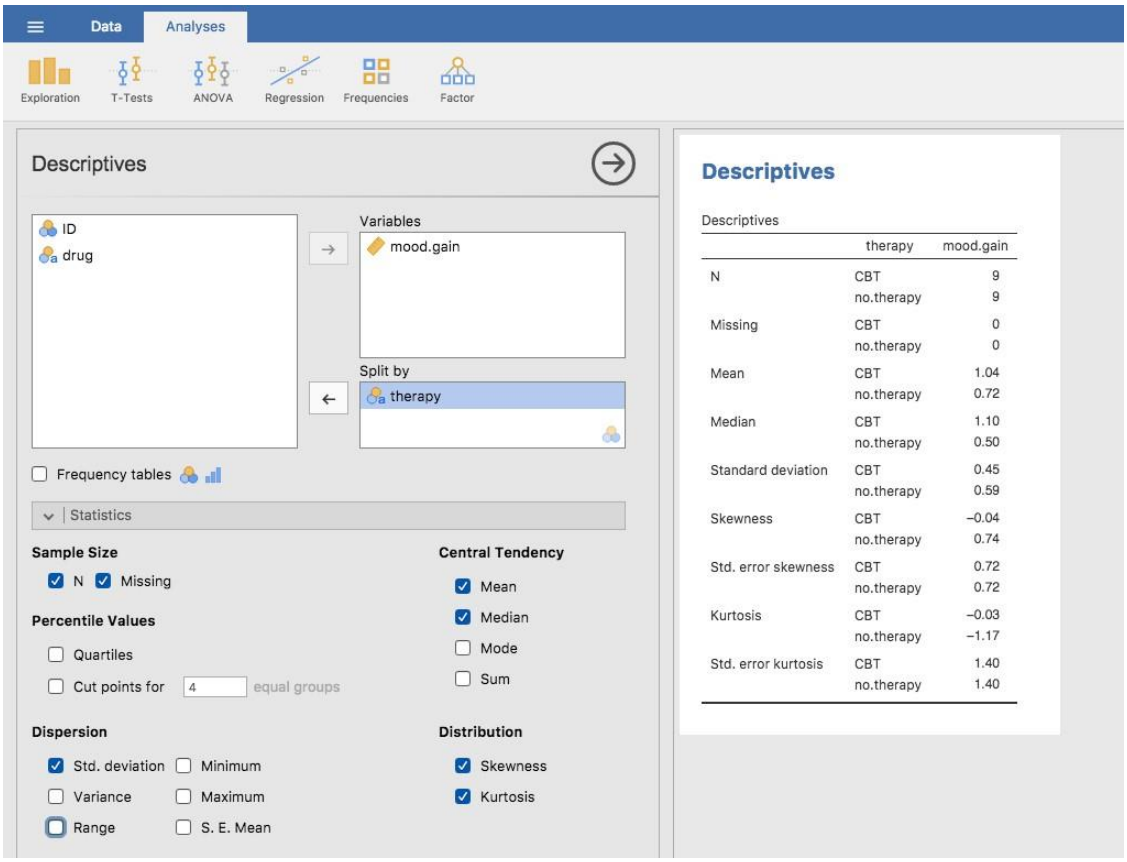


Figure 4-14 : Une capture d'écran de Jamovi montrant les descriptions par type de thérapie

Scores standards

Supposons que mon ami est en train de mettre au point un nouveau questionnaire destiné à mesurer le « grincheux ». L'enquête comporte 50 questions auxquelles vous pouvez répondre de façon grincheuse ou non. Sur un grand échantillon (hypothétiquement, imaginons un million de personnes environ !), les données sont distribuées assez normalement, le score moyen de grincheux étant de 17 sur 50 questions auxquelles on répond de façon grincheuse, et l'écart-type étant de 5. En revanche, lorsque je réponds au questionnaire, je le fais d'une façon grincheuse à 35 questions sur 50. Alors, à quel point suis-je grincheux ? Une façon de le voir serait de dire que j'ai une grinchiosité de 35/50, donc on pourrait dire que je suis grincheux à 70%. Mais c'est un peu bizarre, quand on y pense. Si mon amie avait formulé ses questions un peu différemment, les gens auraient pu y répondre d'une manière différente, de sorte que la distribution globale des réponses pourrait facilement monter ou descendre en fonction de la façon précise dont les questions ont été posées. Donc, je ne suis grincheux à 70 p. 100 qu'en ce qui concerne *cet ensemble de questions du sondage*. Même s'il s'agit d'un très bon questionnaire, ce n'est pas une déclaration très informative.

Une façon plus simple de contourner ce problème est de décrire ma grinchiosité en me comparant à d'autres personnes.

Étonnamment, sur l'échantillon de 1 000 000 de personnes de mon ami, seulement 159 étaient aussi grincheux que… moi (ce n'est pas du tout irréaliste, franchement) suggérant que je suis dans le top 0,016% des grincheux. Cela a beaucoup plus de sens que d'essayer d'interpréter les données brutes. Cette idée selon laquelle nous devrions décrire ma grinchiosité en termes de distribution globale de la grinchiosité des humains, est l'idée qualitative à laquelle la normalisation tente d'aboutir. Une façon d'y parvenir est de faire exactement ce que je viens de faire et de tout décrire en termes de percentiles. Cependant, le problème en faisant ça, c'est qu'on se sent seul au sommet. Supposons que mon ami n'ait collecté qu'un échantillon de 1000 personnes (j'aimerais ajouter encore que c'est un échantillon assez important pour tester un nouveau questionnaire) et cette fois-ci, nous ayons obtenu, disons, une moyenne de 16 sur 50 avec un écart-type de 5. Le problème, c'est qu'il est presque certain qu'aucune personne de cet échantillon ne serait aussi grincheuse… que moi. Cependant, tout n'est pas perdu. Une approche différente consiste à convertir mon score grincheux en un **score standard**, également appelé *score z*. Le score standard est défini comme le nombre d'écarts-types au-dessus de la moyenne de mon score de grinchiosité. Pour l'exprimer en « pseudo-maths », on calcule le score standard de la manière suivante :

$$\text{Score standard} = \frac{\text{Score observe} - \text{moyenne}}{e\text{cart} - \text{type}}$$

En mathématiques réelles, l'équation pour le *z-score* est la suivante

$$z_i = \frac{X_i - \bar{X}}{\hat{\sigma}}$$

Donc, pour en revenir aux données sur la grinchiosité, nous pouvons maintenant transformer la grinchiosité brute de Dani en un score de grinchiosité standardisé.

$$z = \frac{35 - 17}{5} 3{,}6$$

Pour interpréter cette valeur, rappelez-vous l'heuristique grossière que j'ai fournie à la section 4.2.5 dans laquelle j'ai noté que 99,7 % des valeurs devraient se situer dans les 3 écarts-types de la moyenne. Le fait que ma grinchiosité correspond à un score *z* de 3,6 indique que je suis très grincheux en effet. En fait, cela suggère que je suis plus grincheux que 99,98% des gens. C'est à peu près ça.

En plus de vous permettre d'interpréter un score brut par rapport à une population plus large (et donc de donner un sens à des variables qui reposent sur des échelles arbitraires), les scores standard ont une deuxième fonction utile. Les scores standards peuvent être comparés les uns aux autres dans des situations où les scores bruts ne le peuvent pas. Supposons, par exemple, que mon ami avait aussi un autre questionnaire qui mesurait l'extraversion à l'aide d'un questionnaire à 24 items. Comme vous pouvez l'imaginer, cela n'a pas beaucoup de sens d'essayer de comparer mon score brut de 2 sur le questionnaire d'extraversion à mon score brut de 35 sur le questionnaire de grincheux. Les scores bruts

pour les deux variables sont « à peu près » fondamentalement différents, donc ce serait comme comparer des pommes à des oranges.

Qu'en est-il des scores standards ? Eh bien, c'est un peu différent. Si nous calculons les scores standards, nous obtenons z = (35-17)/5 = 3,6 pour la grinchiosité et z = (2-13)/4 = 2,75 pour l'extraversion.

Ces deux nombres *peuvent* être comparés l'un à l'autre.[30] Je suis beaucoup moins extraverti que la plupart des gens (*z* = -2,75) et beaucoup plus grincheux que la plupart des gens (*z* = 3.6). Mais l'ampleur de mon étrangeté est beaucoup plus extrême pour la grinchiosité, puisque 3,6 est un chiffre supérieur à 2,75. Parce que chaque score standardisé est une indication de la position d'une observation par rapport à sa propre population, il est possible de comparer des scores standardisés pour des variables complètement différentes.

Résumé

Calculer quelques statistiques descriptives de base est l'une des toutes premières choses que vous faites lorsque vous analysez des données réelles, et les statistiques descriptives sont beaucoup plus simples à comprendre que les statistiques inférentielles, donc comme dans tous les autres manuels de statistiques, j'ai commencé avec des statistiques descriptives. Dans ce chapitre, nous avons abordé les sujets suivants :

- *Mesures de tendance centrale*. D'une manière générale, les mesures de tendance centrale vous indiquent où se trouvent les données. Trois mesures sont habituellement mentionnées dans la documentation : la moyenne, la médiane et le mode. (Section 4.1)
- *Mesures de la variabilité*. Par contre, les mesures de variabilité vous renseignent sur la façon dont les données sont « étalées «. Les principales mesures sont : l'intervalle, l'écart-type et l'intervalle interquartile.(Section 4.2)
- *Mesures de l'asymétrie et de l'aplatissement*. Nous avons également étudié l'asymétrie dans la distribution d'une variable (skew) et l'applatissement (kurtosis). (Section 4.3)
- *Obtenir des résumés de groupes de variables dans Jamovi*. Puisque ce livre se concentre sur l'analyse des données dans Jamovi, nous avons passé un peu de temps à parler de la façon dont les statistiques descriptives sont calculées pour différents sous-groupes. (Section 4.4)
- *Score standard*. Le *z-score* est une bête un peu inhabituelle. Ce n'est pas tout à fait une statistique descriptive, ni une inférence. Nous en avons parlé à la section 4.5. Assurez-vous de bien comprendre cette section. On en reparlera plus tard.

Dans le prochain chapitre, nous passerons à une discussion sur la façon de dessiner des images ! Tout le monde aime les belles photos, non ? Mais avant de le faire, je voudrais

---

[30] Bien qu'une certaine prudence soit généralement de mise. Il n'est pas toujours vrai qu'un écart-type sur la variable A correspond au même « type » de chose qu'un écart-type sur la variable B. Faites preuve de bon sens lorsque vous essayez de déterminer si les scores *z* de deux variables peuvent être comparés de façon significative.

terminer sur un point important. Un premier cours traditionnel en statistique ne consacre qu'une petite partie de la classe à la statistique descriptive, peut-être une ou deux conférences tout au plus. La grande majorité du temps du conférencier est consacrée aux statistiques inférentielles parce que c'est là que se trouvent toutes les choses difficiles. C'est logique, mais cela cache l'importance pratique quotidienne de choisir de bonnes descriptions. Avec cela à l'esprit….

Epilogue : Les bonnes statistiques descriptives sont descriptives !

La mort d'un homme est une tragédie. La mort de millions de personnes est une statistique. - Josef Staline, Potsdam 1945

950 000 – 1 200 000 - Estimation du nombre de victimes de la répression soviétique, 1937-1938 (Ellman 2002)

La citation tristement célèbre de Staline sur le caractère statistique de la mort de millions de personnes mérite réflexion. L'intention claire de sa déclaration est que la mort d'un individu nous touche personnellement et sa force ne peut être niée, mais que les morts d'une multitude sont incompréhensibles et, par conséquent, sont de simples statistiques plus facilement ignorées. Je dirais que Staline avait à moitié raison. Une statistique est une abstraction, une description d'événements au-delà de notre expérience personnelle, et si difficile à visualiser. Peu d'entre nous, sinon aucun d'entre nous, ne peut imaginer à quoi ressemble « vraiment » la mort de millions de personnes, mais nous pouvons imaginer une seule mort et cela donne à la seule mort son sentiment de tragédie immédiate, un sentiment qui manque dans la description statistique froide d'Ellman.

Pourtant, ce n'est pas si simple. Sans chiffres, sans dénombrement, sans description de ce qui s'est passé, nous *n'avons aucune chance* de comprendre ce qui s'est réellement passé, aucune occasion même d'essayer d'invoquer le sentiment manquant. Et en vérité, au moment où j'écris ces lignes, assis confortablement un samedi matin, à la moitié d'une vie et loin des goulags, lorsque je mets l'estimation d'Ellman à côté de la citation de Staline, une crainte obsédante me prend l'estomac et un frisson me parcourt.

La répression stalinienne est quelque chose qui dépasse vraiment mon expérience, mais avec une combinaison de données statistiques et d'histoires personnelles enregistrées qui nous sont parvenues, ce n'est pas entièrement au-delà de ma compréhension. Parce que les chiffres d'Ellman nous disent ceci : sur une période de deux ans, la répression stalinienne a anéanti l'équivalent de chaque homme, femme et enfant vivant actuellement dans la ville où je vis. Chacun de ces décès avait sa propre histoire, sa propre tragédie, et nous n'en connaissons que quelques-uns à l'heure actuelle. Malgré tout, avec quelques statistiques soigneusement choisies, l'ampleur de l'atrocité commence à se faire sentir.

Il n'est donc pas anodin de dire que la première tâche du statisticien et du scientifique est de résumer les données, de trouver un ensemble de chiffres qui puissent transmettre à un public une idée de ce qui s'est passé. C'est le travail des statistiques descriptives, mais ce n'est pas un travail qui est fait uniquement à l'aide des chiffres. Vous êtes un analyste de données et non un logiciel statistique. Une partie de votre travail consiste à prendre ces *statistiques* et à les transformer en une *description*. Lorsque vous analysez des données, il ne suffit pas d'énumérer une collection de chiffres. Rappelez-vous toujours que ce que vous essayez vraiment de faire, c'est de communiquer avec un public humain. Les chiffres sont

importants, mais ils doivent être rassemblés en une histoire significative que votre auditoire peut interpréter. Cela signifie que vous devez penser à la structure. Vous devez penser au contexte. Et vous devez penser aux événements individuels que vos statistiques résument.

Réaliser des graphiques

Par-dessus tout, affichez les données. -Edward Tufte[31]

La visualisation des données est l'une des tâches les plus importantes de l'analyste de données. C'est important pour deux raisons distinctes mais étroitement liées. Tout d'abord, il s'agit de dessiner des « représentations graphique », l'affichage de vos données d'une manière propre et visuellement attrayante facilite la compréhension de ce que vous essayez de leur dire par votre lecteur. Le fait que dessiner des graphiques *vous* aide à comprendre les données est tout aussi important, voire plus important encore. Pour ce faire, il est important de dessiner des « graphiques exploratoires » qui vous aideront à en apprendre davantage sur les données au fur et à mesure que vous les analysez. Ces points peuvent sembler assez évidents, mais je ne peux pas compter le nombre de fois où j'ai vu des gens les oublier.

Pour donner une idée de l'importance de ce chapitre, je veux commencer par une illustration classique de la puissance d'un bon graphique. Pour ce faire, la Figure 5-1 présente une reproduction de l'une des visualisations de données les plus célèbres de tous les temps. C'est la carte des décès dus au choléra de John Snow en 1854. La carte est élégante dans sa simplicité. En arrière-plan, nous avons un plan des rues qui aide à orienter le spectateur. En haut, on voit un grand nombre de petits points, chacun représentant l'emplacement d'un cas de choléra. Les plus grands symboles indiquent l'emplacement des pompes à eau, étiquetées par leur nom. Même l'inspection la plus superficielle du graphique montre très clairement que la source de l'éclosion est presque certainement la pompe de Broad Street. En montrant ce graphique, le Dr Snow a fait en sorte que la poignée soit retirée de la pompe et a mis fin à l'épidémie qui avait tué plus de 500 personnes. Telle est la puissance d'une bonne visualisation des données.

Les objectifs de ce chapitre sont doubles. Tout d'abord, discuter de plusieurs graphiques assez standard que nous utilisons beaucoup lors de l'analyse et de la présentation des données, et ensuite vous montrer comment créer ces graphiques en Jamovi. Les graphiques eux-mêmes ont tendance à être assez simples, de sorte qu'à certains égards, ce chapitre est assez simple. Là où les gens ont habituellement des difficultés, c'est pour apprendre à produire des graphiques et surtout d'apprendre à produire de bons graphiques. Heureusement, apprendre à dessiner des graphiques avec Jamovi est assez simple tant que vous n'êtes pas trop pointilleux sur l'aspect de votre graphique. Ce que je veux dire en disant cela, c'est que Jamovi offre beaucoup de *très* bons graphiques par défaut, ou tracés,

---

[31] L'origine de cette citation est le beau livre de Tufte, *The Visual Display of Quantitative Information*.

qui produisent la plupart du temps un graphique propre et de haute qualité. Cependant, dans les cas où vous voudriez faire quelque chose de non standard, ou si vous avez besoin d'apporter des changements très spécifiques à la figure, sachez que les fonctionnalités graphiques de Jamovi ne sont pas encore capables de supporter un travail avancé ou une édition détaillées.

Figure 5-1 : Une reproduction stylisée de la carte originale de John Snow sur le choléra. Chaque petit point représente l'emplacement d'un cas de choléra et chaque grand cercle indique l'emplacement d'un puits. Comme le montre clairement le graphique, l'épidémie de choléra est centrée très étroitement sur la pompe Broad St.

Histogrammes

Commençons par l'humble **histogramme**. Les histogrammes sont l'un des moyens les plus simples et les plus utiles de visualiser les données. Ils sont plus pertinents lorsque vous disposez d'une variable sur une échelle d'intervalle ou de ratio (par exemple, les données de marges afl. du chapitre 4) et que vous voulez obtenir une impression générale de la variable. La plupart d'entre vous savent probablement comment fonctionnent les histogrammes, puisqu'ils sont largement utilisés, mais par souci d'exhaustivité, je vais les décrire. Tout ce que vous faites est de diviser les valeurs possibles en **compartiments**, puis de compter le nombre d'observations qui tombent dans chaque compartiment. Ce comptage est appelé fréquence ou densité du compartiment et est affiché sous la forme d'une barre verticale. Dans les données sur les marges gagnantes de l'AFL, il y a 33 jeux où la marge gagnante était inférieure à 10 points et c'est ce fait qui est représenté par la hauteur de la barre la plus à gauche que nous avons montrée précédemment au chapitre 4, Figure 4-2. Avec ces graphiques précédents, nous avons utilisé un package de traçage avancé en R qui, pour l'instant, est au-delà des capacités de Jamovi. Mais Jamovi s'en rapproche, et dessiner cet histogramme en Jamovi est assez simple. Ouvrez les options « tracés » sous « Exploration » - « Descriptives » et cliquez sur la case à cocher « Histogram », comme dans

la Figure 5-2. Jamovi étiquette par défaut l'axe des y « density » et l'axe des x avec le nom de variable. Les **compartiments** sont sélectionnés automatiquement, et il n'y a pas d'information sur l'échelle, ou de comptage, sur l'axe des y, contrairement à la Figure 4-2 précédente. Mais cela n'a pas trop d'importance car ce qui nous intéresse vraiment, c'est notre impression sur la forme de la distribution : est-elle normalement distribuée ou y a-t-il une asymétrie ou un aplatissement ? Nos premières impressions sur ces caractéristiques proviennent du tracé d'un **histogramme**.

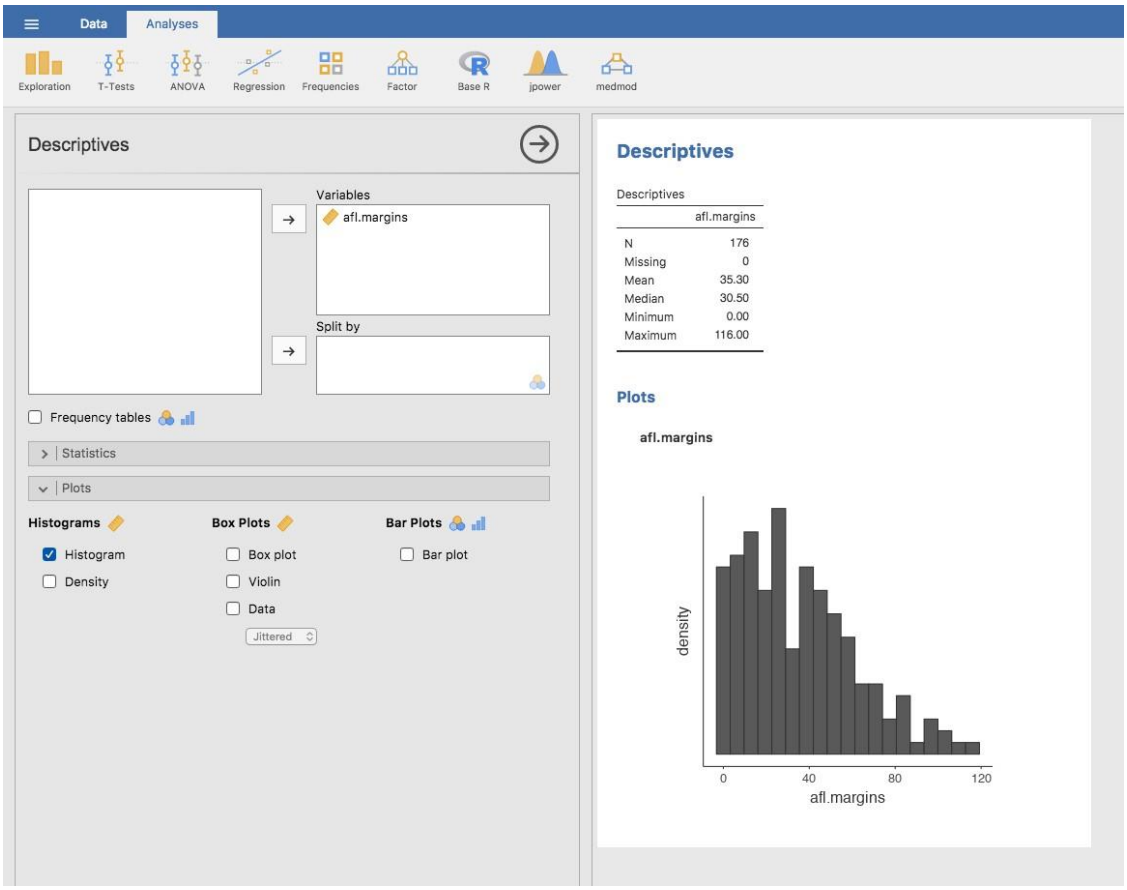


Figure 5-2 : Ecran Jamovi montrant la case à cocher histogramme

Une caractéristique supplémentaire que Jamovi offre est la possibilité de tracer une courbe de « densité ». Vous pouvez le faire en cliquant sur la case à cocher « Density » sous les options « Graphiques » (et en décochant « Histogramme »), ce qui nous donne le graphique présenté dans la Figure 5-3. Un graphe de densité visualise la distribution des données sur un intervalle continu ou une période de temps. Ce graphique est une variante d'un histogramme qui utilise **l'estimation par noyau** pour tracer les valeurs, ce qui permet des distributions plus fines en lissant le bruit. Les pics d'un graphe de densité permettent d'afficher l'endroit où les valeurs sont concentrées sur l'intervalle. L'avantage des courbes de densité par rapport aux histogrammes est qu'elles sont plus aptes à déterminer la forme de distribution parce qu'elles ne sont pas affectées par le nombre de compartiments utilisés (chaque barre utilisée dans un histogramme typique). Un histogramme ne comprenant que 4 compartiments ne produirait pas une forme de distribution suffisamment distincte

comme le ferait un histogramme de 20 compartiments. En revanche, dans le cas les graphiques de densité, ce n'est pas un problème.

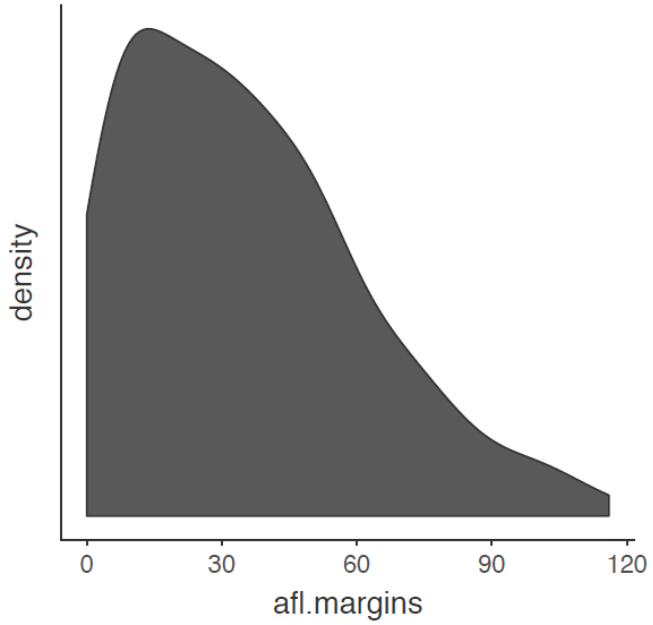


Figure 5-3 : Un graphe de densité de la variable afl.margins tracé avec Jamovi

Bien que cette image ait besoin de beaucoup de nettoyage pour faire une bonne présentation graphique (c.-à-d. une présentation que vous incluriez dans un rapport), elle fait néanmoins un assez bon travail pour décrire les données. En fait, la grande force d'un histogramme ou d'un graphe de densité est qu'il montre (correctement utilisé) toute la dispersion des données, de sorte que vous pouvez avoir une idée assez précise de ce à quoi elle ressemble. L'inconvénient des histogrammes est qu'ils ne sont pas très compacts. Contrairement à d'autres graphiques, il faut souligner le fait qu'il est difficile d'entasser 20 à 30 histogrammes dans une seule image sans submerger le lecteur. Et bien sûr, si vos données sont à l'échelle nominale, les histogrammes sont inutiles.

Boxplots

Une autre alternative aux histogrammes est un **boxplot**, que l'on appelle parfois un tracé « boîte et moustaches ». Comme les histogrammes, ils sont plus adaptés aux données sur une échelle d'intervalle ou de rapport. L'idée derrière un boxplot est de fournir une représentation visuelle simple de la médiane, de l'écart interquartile et de l'étendue des données. Et parce qu'ils le font d'une manière assez compacte, les boxplots sont devenus un graphique statistique très populaire, surtout pendant la phase exploratoire de l'analyse des données lorsque vous essayez de comprendre les données vous-même. Voyons comment ils fonctionnent, encore une fois en utilisant les données afl.margins comme exemple.

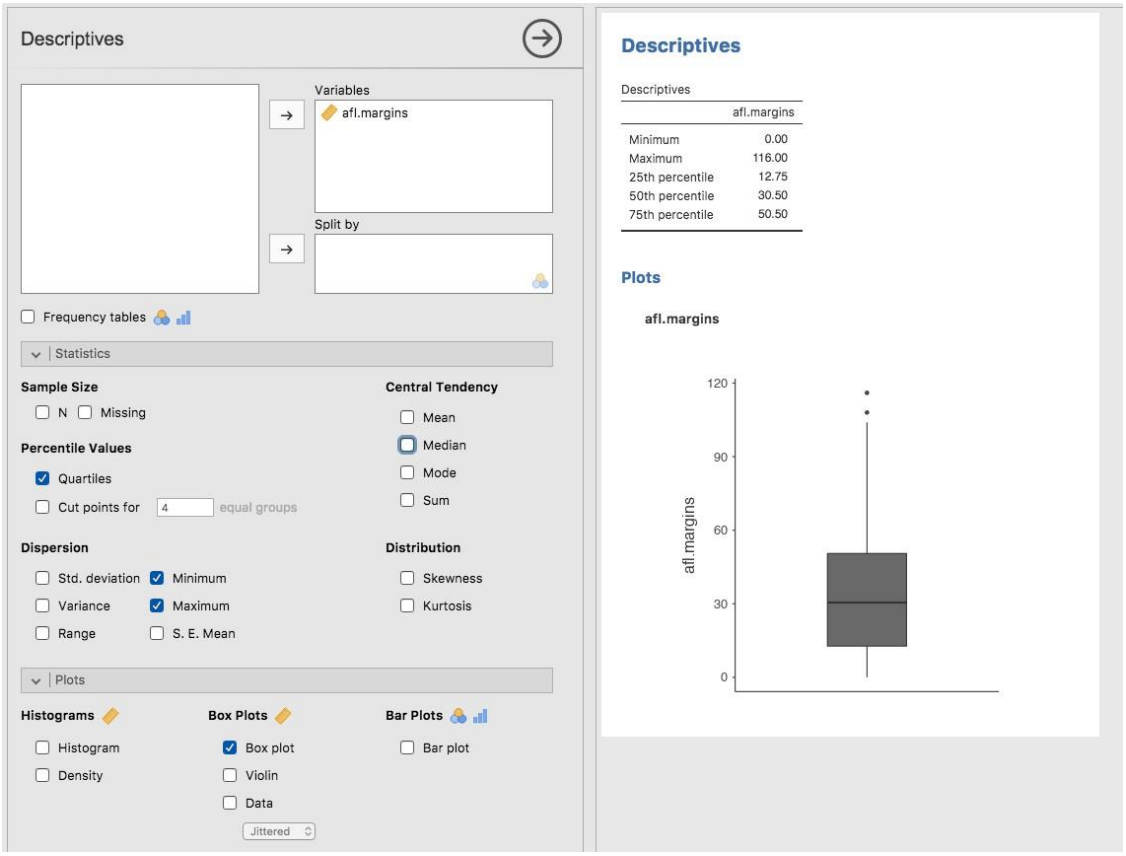


Figure 5-4 : Un box plot de la variable afl.margins tracée dans Jamovi

La façon la plus simple de décrire à quoi ressemble un boxplot est d'en dessiner un. Cliquez sur la case à cocher « Box plot » et vous obtiendrez le graphique en bas à droite de la Figure 5.4. Jamovi a dessiné le boxplot le plus simple possible. Lorsque vous regardez ce graphique, voici comment vous devez l'interpréter : la ligne épaisse au milieu de la boîte est la médiane ; la boîte elle-même s'étend du 25e centile au 75e centile ; et les « moustaches » vont jusqu'au point de données le plus extrême qui ne dépasse pas une certaine limite. Par défaut, cette valeur est 1,5 fois l'écart interquartile (IQR), calculé comme suit 25e percentile - (1,5*IQR) pour la limite inférieure, et 75e percentile + (1,5*IQR) pour la limite supérieure. Toute observation dont la valeur se situe en dehors de cette plage est tracée sous la forme d'un cercle ou d'un point au lieu d'être couverte par les moustaches, et est communément désignée comme une **valeur aberrante**. Pour nos données de marges AFL, il y a deux observations qui se situent en dehors de cette plage, et ces observations sont tracées sous forme de points (la limite supérieure est 107, et en regardant au-dessus de la colonne de données dans le tableur, il y a deux observations avec des valeurs supérieures, 108 et 116, ce sont ici les points).

Graphique en violon

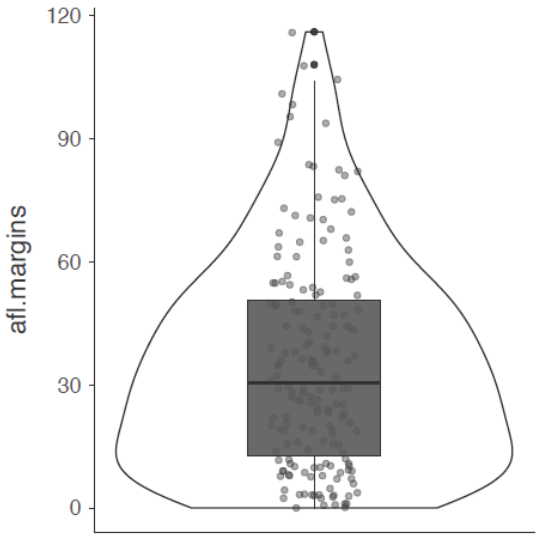


Figure 5-5 : Un graphique en violon de la variable afl.margins tracé in Jamovi, alsow montrant un tracé en boîte et des points de données.

Le graphique en violon est une variante du box plot traditionnel. Les graphiques en violon sont semblables aux graphiques en boîtes, sauf qu'ils montrent également la densité de probabilité du noyau des données à des valeurs différentes. Habituellement, les graphiques en violon comprennent un marqueur pour la médiane des données et une boite indiquant l'écart interquartile, comme dans les boxplots standards. Dans Jamovi, vous pouvez réaliser ce type de graphiques en cochant les cases « Violin » et « Boxplot ». Voir la Figure 5-5, où la case à cocher « Data » a également été cochée pour faire apparaitre les points de données réels sur le graphique. Cela tend cependant à rendre le graphique un peu trop surchargé, à mon avis. La clarté est la simplicité, donc dans la pratique, il peut être préférable d'utiliser un simple diagramme en boîtes.

Dessiner plusieurs boxplots

Une dernière chose. Que faire si vous voulez dessiner plusieurs boxplots à la fois ? Supposons, par exemple, que je veuille des boxplots séparés montrant les marges AFL non seulement pour 2010 mais pour chaque année entre 1987 et 2010. Pour ce faire, la première chose à faire est de trouver les données. Ceux-ci sont stockés dans le fichier aflsmall2.csv. Alors chargeons-le dans Jamovi et voyons ce qu'il y a dedans. Vous verrez qu'il s'agit d'un assez grand ensemble de données. Il contient 4296 observations et les variables qui nous intéressent. Ce que nous voulons faire, c'est dessiner des boxplots avec Jamovi pour la variable margin, mais les tracés séparément pour chaque année. Pour ce faire, il suffit de faire glisser la variable année dans la case « Split by », comme dans la Figure 5-6.

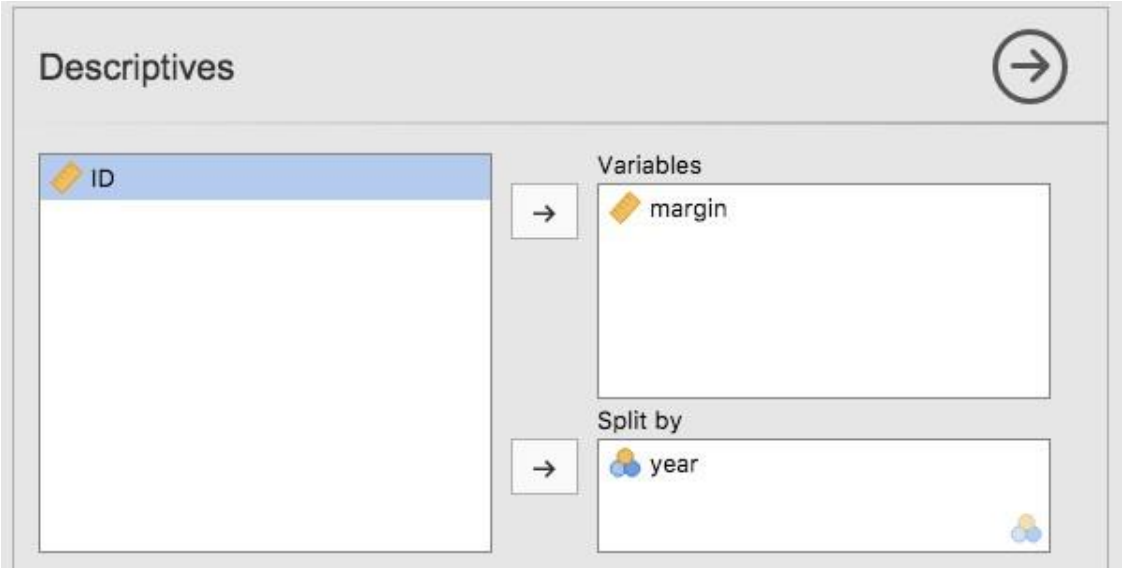


Figure 5-6 : Capture d'écran de Jamovi montrant la fenêtre « Split by »

Le résultat est illustré à la Figure 5-7. Cette version du diagramme en boîtes, divisé par année, donne une idée de la raison pour laquelle il est parfois utile de choisir des diagrammes en boîtes plutôt que des histogrammes. Il est possible d'avoir une bonne idée de l'aspect des données d'une année à l'autre sans être submergé par trop de détails. Imaginez maintenant ce qui se serait passé si j'avais essayé d'entasser 24 histogrammes dans cet espace : aucune chance que le lecteur apprenne quelque chose d'utile.

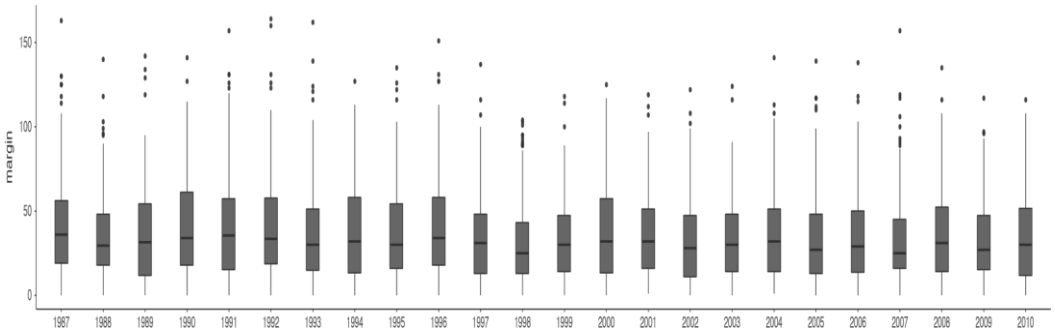


Figure 5-7 : Boxplots multiples tracés dans Jamovi, pour les variables de marge par année dans l'ensemble de données aflsmall2

Utilisation de diagrammes en boîtes pour détecter les valeurs aberrantes

Parce que le boxplot sépare automatiquement les observations qui se situent en dehors d'une certaine plage, les dépeignant avec un point dans le Jamovi, les gens les utilisent souvent comme une méthode informelle pour la détection des **valeurs aberrantes** : observations qui sont « étrangement » éloignées du reste des données.

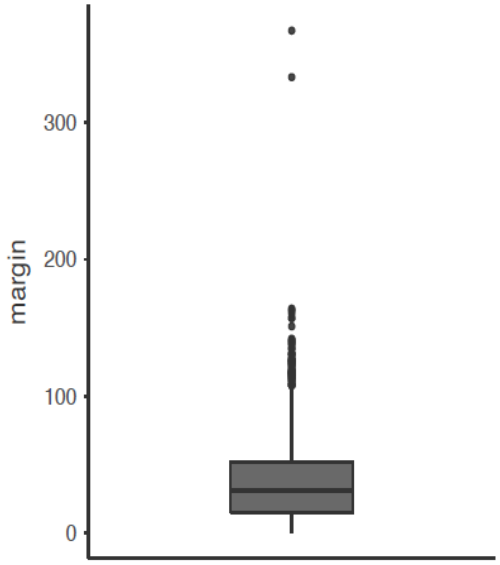


Figure 5-8 : Un boxplot montrant deux valeurs aberrantes très suspectes !

En voici un exemple. Supposons que j'ai dessiné le boxplot pour les données des marges AFL et qu'il apparaisse comme dans la Figure 5-8. Il est assez clair qu'il se passe quelque chose de bizarre avec deux des observations. Apparemment, il y a eu deux matchs où la marge était de plus de 300 points ! Ça ne me semble pas correct. Maintenant que j'ai un doute, il est temps d'examiner un peu plus attentivement les données. Dans Jamovi vous pouvez rapidement découvrir lesquelles de ces observations sont suspectes et ensuite vous pouvez retourner aux données brutes pour voir s'il y a eu une erreur dans la saisie des données. Pour ce faire, vous devez configurer un filtre de sorte que seules les observations dont les valeurs dépassent un certain seuil soient incluses. Dans notre exemple, le seuil est supérieur à 300, c'est donc le filtre que nous allons créer. Tout d'abord, cliquez sur le bouton « Filtres » en haut de la fenêtre Jamovi, puis tapez « margin>300 » dans le champ filtre, comme dans la Figure 5-9.

Ce filtre crée une nouvelle colonne dans la vue feuille de calcul où seules les observations qui passent le filtre sont incluses. Une façon simple d'identifier rapidement ces observations est de dire à Jamovi de produire une « Table de fréquences » (dans la fenêtre « Exploration » - « Descriptives ») pour la variable ID (qui doit être une variable nominale sinon la table de fréquences ne sera pas produite). Dans la Figure 5-10, vous pouvez voir que les valeurs ID pour les observations où la marge était supérieure à 300 sont 14 et 134. Il s'agit de cas suspects, ou d'observations, où vous devriez retourner à la source de données originale pour savoir ce qui se passe.

Habituellement, vous constatez que quelqu'un a tapé le mauvais chiffre. Bien que cela puisse sembler un exemple stupide, je dois souligner que ce genre de chose arrive souvent. Les ensembles de données du monde réel sont souvent truffés d'erreurs stupides, surtout lorsqu'une personne a dû saisir quelque chose sur un ordinateur à un moment donné. En

fait, il y a un nom pour cette phase de l'analyse des données et dans la pratique, cela peut prendre une grande partie de notre temps : le **nettoyage des données**.

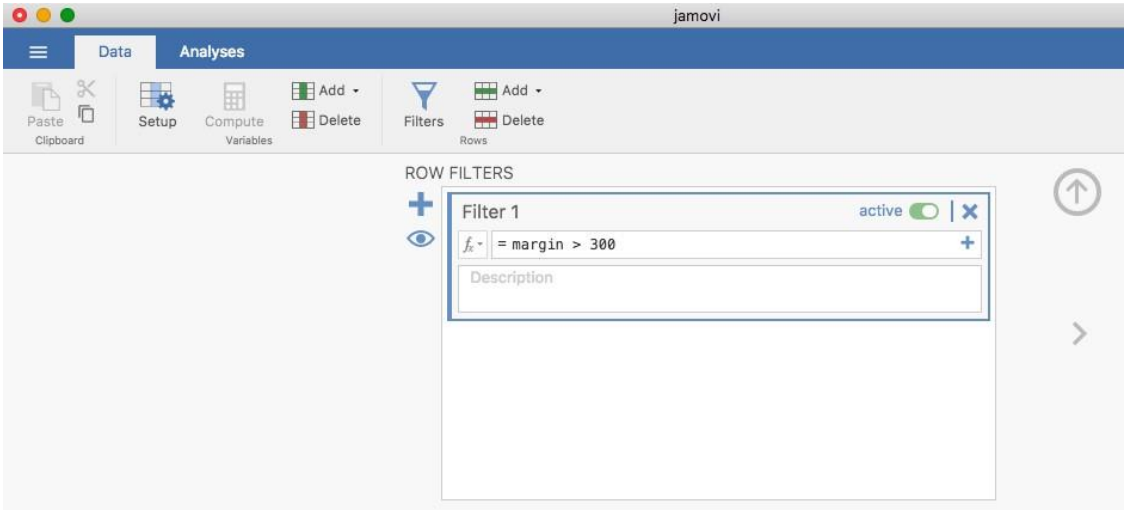


Figure 5-9 : La grille du filtre Jamovi

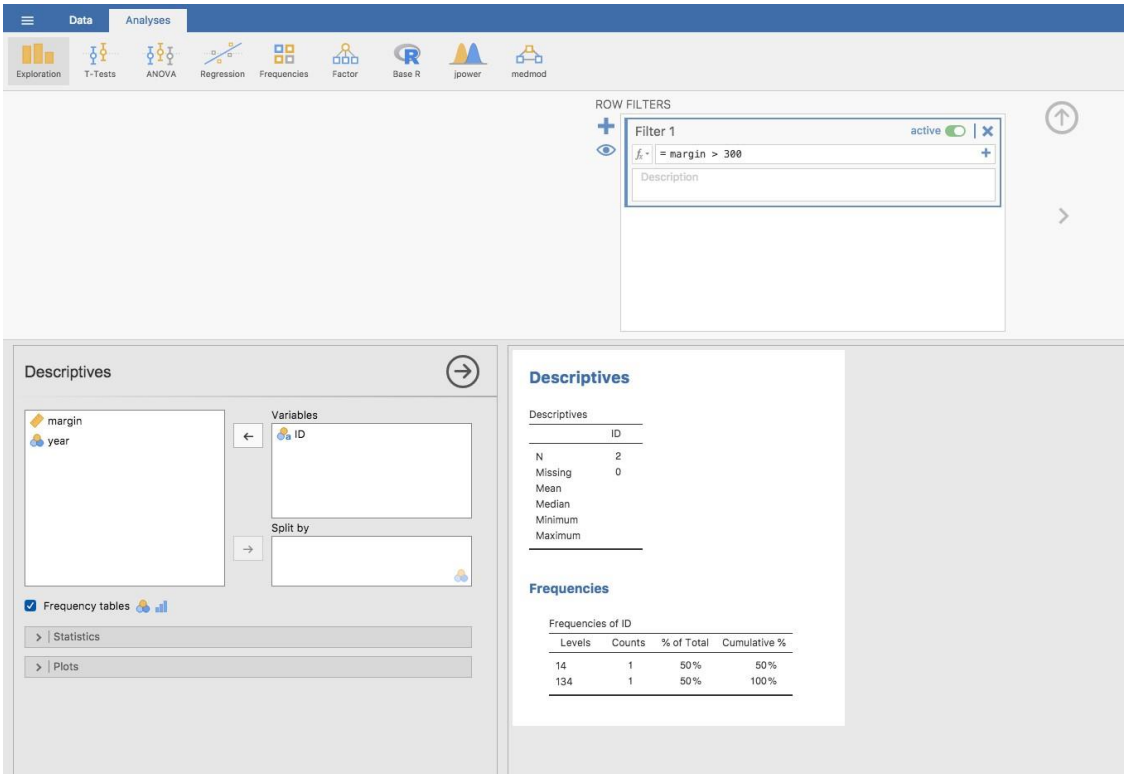


Figure 5-10 : Tableau de fréquence des numéros d'identification indiquant les numéros d'identification des deux valeurs aberrantes suspectes : 14 et 134

Il s'agit de rechercher les fautes de frappe (« typos »), les données manquantes et toutes sortes d'autres erreurs insupportables dans les fichiers de données brutes.

Pour des valeurs moins extrêmes, même si elles sont marquées dans un boxplot comme valeurs aberrantes, la décision d'inclure ou non des valeurs aberrantes dans une analyse dépend fortement de la *raison pour laquelle* vous pensez que les données ressemblent à ce qu'elles sont et de l'usage que vous voulez en *faire*. Vous devez vraiment faire preuve de jugement. Si la valeur aberrante vous semble légitime, gardez-la. Quoi qu'il en soit, je reviendrai sur le sujet à la section 12.10.

diagramme en barres

Une autre forme de graphique que vous voudrez souvent tracer est le **diagramme en barres**. Utilisons l'ensemble de données afl.finalists avec la variable afl.finalists que j'ai présentée à la section 4.1.6. Ce que je veux faire, c'est dessiner un graphique en barres qui affiche le nombre de finales auxquelles chaque équipe a participé au cours de la période couverte par l'ensemble de données afl.finalists. Il y a beaucoup d'équipes, mais je suis particulièrement intéressé par quatre d'entre elles : Brisbane, Carlton, Fremantle et Richmond. La première étape consiste donc à mettre en place un filtre pour que seules ces quatre équipes soient incluses dans le diagramme à barres. C'est très simple dans Jamovi et vous pouvez le faire en utilisant la fonction « Filters » que nous avons utilisée précédemment. Ouvrez la fenêtre « Filters » et tapez ce qui suit :

afl.finalistes ==‘Brisbane’ ou afl.finalistes ==‘Carlton’ ou afl.finalistes ==‘Fremantle’ ou afl.finalistes ==‘Richmond’.[Jamovi utilise ici le symbole “==” pour signifier « correspondre »]

Lorsque vous aurez fait cela, vous verrez, dans la vue « Data », que Jamovi a filtré toutes les valeurs à l'exception de celles que nous avons spécifiées. Ensuite, ouvrez la fenêtre « Exploration » - « Descriptives » et cliquez sur la case « Bar plot » (n'oubliez pas de déplacer la variable afl.finalists dans la case « Variables » pour que Jamovi sache quelle variable utiliser). Vous devriez alors obtenir un diagramme à barres, quelque chose comme celui illustré à la Figure 5-11.

Enregistrer des fichiers image en utilisant Jamovi

En attendant, vous vous dites peut-être : Quel est l'intérêt de pouvoir dessiner de jolis graphiques avec Jamovi si je ne peux pas les sauvegarder et les envoyer à des amis pour me vanter de l'incroyable qualité de mes données ? Comment sauvegarder l'image ? Simple. Il suffit de cliquer avec le bouton droit de la souris sur l'image du graphique et de l'enregistrer dans un fichier, aux formats « .eps », « svg » ou « pdf ». Ces formats produisent tous de belles images que vous pouvez envoyer à vos amis, ou inclure dans vos devoirs ou papiers.

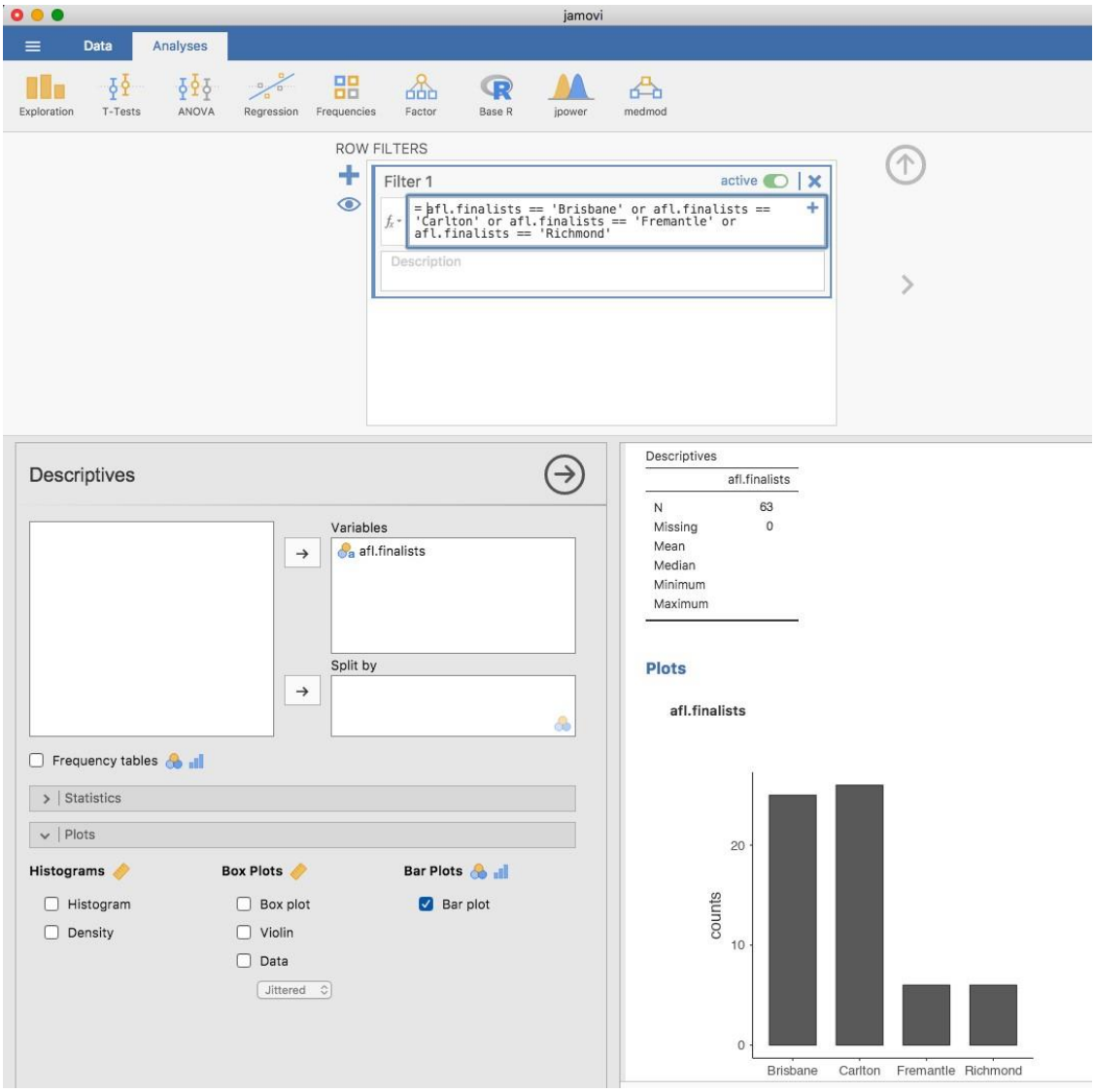


Figure 5-11 : Filtrage pour n'inclure que quatre équipes de l'AFL, et dessin d'un tracé de bar à Jamovi

Résumé

Je suis peut-être une personne simple d'esprit, mais j'adore les images. Chaque fois que j'écris un nouvel article scientifique, l'une des premières choses que je fais est de m'asseoir et de réfléchir à ce que seront les illustrations. Dans ma tête, un article n'est en fait qu'une suite d'images reliées entre elles par une histoire. Tout le reste n'est que de la poudre aux yeux. Ce que j'essaie vraiment de dire ici, c'est que le système visuel humain est un outil d'analyse de données très puissant. Donnez-lui le bon type d'information et il fournira très rapidement à un lecteur humain une énorme quantité de connaissances. Ce n'est pas pour rien que nous avons le dicton « une image vaut mille mots ». Dans cet esprit, je pense qu'il s'agit d'un des chapitres les plus importants du livre. Les sujets abordés étaient :

- *Graphiques communs*. Une grande partie du chapitre a porté sur les graphiques standards que les statisticiens aiment produire : histogrammes (section 5.1), diagrammes de boxplots (section 5.2) et diagrammes en barres (section 5.3).
- *Enregistrer des fichiers image*. Il est important de noter que nous avons également abordé la façon d'exporter vos photos (Section 5.4)

Une dernière chose à souligner. Bien que Jamovi produise des graphiques par défaut très soignés, l'édition des tracés n'est actuellement pas possible. Pour des graphismes et des capacités de traçage plus avancés, les packages disponibles dans R sont beaucoup plus puissants. L'un des systèmes graphiques les plus populaires est fourni par le paquet ggplot2 (voir http:www//ggplot2.org/), qui est basé sur « The grammar of graphics » (Wilkinson 2005). Ce n'est pas pour les novices. Il faut avoir une bonne maîtrise de R avant de pouvoir commencer à l'utiliser, et même là, il faut un certain temps pour vraiment s'y habituer. Mais quand vous êtes prêt, cela vaut la peine de prendre le temps de vous enseigner, car c'est un système beaucoup plus puissant et plus propre.

Problèmes pratiques

Le jardin de la vie ne semble jamais se limiter aux intrigues que les philosophes ont tracées pour son confort. Peut-être que quelques tracteurs de plus feraient l'affaire. - Roger Zelazny[32]

C'est un chapitre un peu étrange, même selon mes critères. Mon objectif dans ce chapitre est de parler un peu plus honnêtement des réalités du travail avec les données que vous ne le verrez ailleurs dans le livre. Le problème avec les ensembles de données du monde réel, c'est qu'ils sont *désordonnés*. Très souvent, le fichier de données avec lequel vous commencez n'a pas les variables stockées dans le bon format pour l'analyse que vous voulez faire. Parfois, il peut y avoir beaucoup de valeurs manquantes dans votre ensemble de données. Parfois, vous ne voulez analyser qu'un sous-ensemble des données. Et cetera. En d'autres termes, il y a beaucoup de **manipulation de données** que vous devez faire juste pour obtenir les variables dans votre ensemble de données dans le format dont vous avez besoin. Le but de ce chapitre est de fournir une introduction de base à ces sujets pragmatiques. Bien que le chapitre soit motivé par le genre de problèmes pratiques qui surgissent lors de la manipulation de données réelles, je m'en tiendrai à la pratique que j'ai adoptée dans la majeure partie du livre et je m'appuierai sur de très petits ensembles de données qui illustrent le problème sous-jacent. Comme ce chapitre est essentiellement un recueil de techniques et qu'il ne raconte pas une seule histoire cohérente, il peut être utile de commencer par une liste de sujets :

- Section 6.1. Mise en tableaux des données.
- Section 6.2. Utiliser des expressions logiques.
- Section 6.3. Transformer ou recoder une variable.
- Section 6.4. Quelques fonctions mathématiques utiles.

---

[32] La citation provient de *Home is the* Hangman, publié en 1975.

- Section 6.5. Extraction d'un sous-ensemble d'un ensemble de données.

Comme vous pouvez le constater, la liste des sujets abordés dans le chapitre est assez vaste, et il y a *beaucoup* de contenu. Même s'il s'agit d'un des chapitres les plus longs et les plus difficiles du livre, je ne fais qu'effleurer plusieurs sujets assez différents et importants. Mon conseil, comme d'habitude, est de lire le chapitre une fois et d'essayer de le suivre autant que possible. Ne vous inquiétez pas trop si vous ne pouvez pas tout saisir d'un coup, surtout dans les sections suivantes. Le reste du livre ne s'appuie que légèrement sur ce chapitre pour que vous puissiez vous en sortir avec une simple compréhension des bases. Cependant, ce que vous découvrirez probablement plus tard, c'est que vous devrez revenir à ce chapitre pour comprendre certains des concepts auxquels je fais référence ici.

Mise en tableaux et recoupement des données

Une tâche très courante lors de l'analyse des données est la construction de tableaux de fréquence, ou de tableaux croisés d'une variable par rapport à une autre. Ces tâches peuvent être réalisées avec Jamovi et je vais vous montrer comment dans cette section.

Création de tables pour des variables individuelles

Commençons par un exemple simple. En tant que père d'un petit enfant, je passe naturellement beaucoup de temps à regarder des émissions de télévision comme *In the Night Garden*. Dans le fichier nightgarden.csv, j'ai transcrit une courte section du dialogue. Le fichier contient deux variables d'intérêt, le locuteur (speaker) et l'énoncé (Utterance). Ouvrez cet ensemble de données dans Jamovi et jetez un coup d'œil aux données dans la vue « feuille de calcul ». Vous verrez que les données ressemblent à ceci :

Variable « Speaker » :

upsy-daisy upsy-daisy upsy-daisy upsy-daisy upsy-daisy tombliboo tombliboo makka-pakka makka-pakka makka-pakka makka-pakka makka-pakka makka-pakka

variable « Utterance » :

pip pip pip onk onk onk ee oo pip pip onk onk onk

En regardant cela apparait clairement ce qui est arrivé à ma santé mentale ! Avec des données comme celle-ci, une tâche à laquelle je pourrais me trouver confronter est de construire un compte de fréquence du nombre de mots que chaque personnage parle pendant l'émission. Le menu Jamovi « Descriptives » comporte une case à cocher appelée « Frequencies Tables» qui fait cela, voir Figure 6-1.

Frequencies of speaker

Levels	Counts	% of total	Cumulative %
makka-pakka	4	40%	40%
tomliboo	2	20%	60%
upsy-daisy	4	40%	100%

Figure 6-1 : Tableau des fréquences pour la variable speaker

La sortie ici nous indique sur la première ligne que ce que nous regardons est une distribution de la variable du speaker. Dans la colonne « Levels », elle liste tous les locuteurs qui existent dans les données, et dans la colonne « Counts », il vous indique combien de fois ce locuteur apparaît dans les données. En d'autres termes, ceci est une table de fréquences.

Dans Jamovi, la case à cocher « Frequencies Tables » ne produira qu'un tableau pour une seule variable.

Pour un tableau à deux variables, par exemple en combinant le speaker et Utterance de sorte que nous puissions voir combien de fois chaque locuteur a prononcé un énoncé particulier, nous avons besoin d'un tableau croisé ou tableau de contingence. Dans Jamovi vous pouvez le faire en sélectionnant les « Frequencies » - « Contingence Tables » - « Independants Samples », et en déplaçant la variable speaker dans la case « Rows », et la variable Utterance dans la case « Columns ». Vous devriez alors avoir un tableau de contingence comme celui illustré à la Figure 6-2.

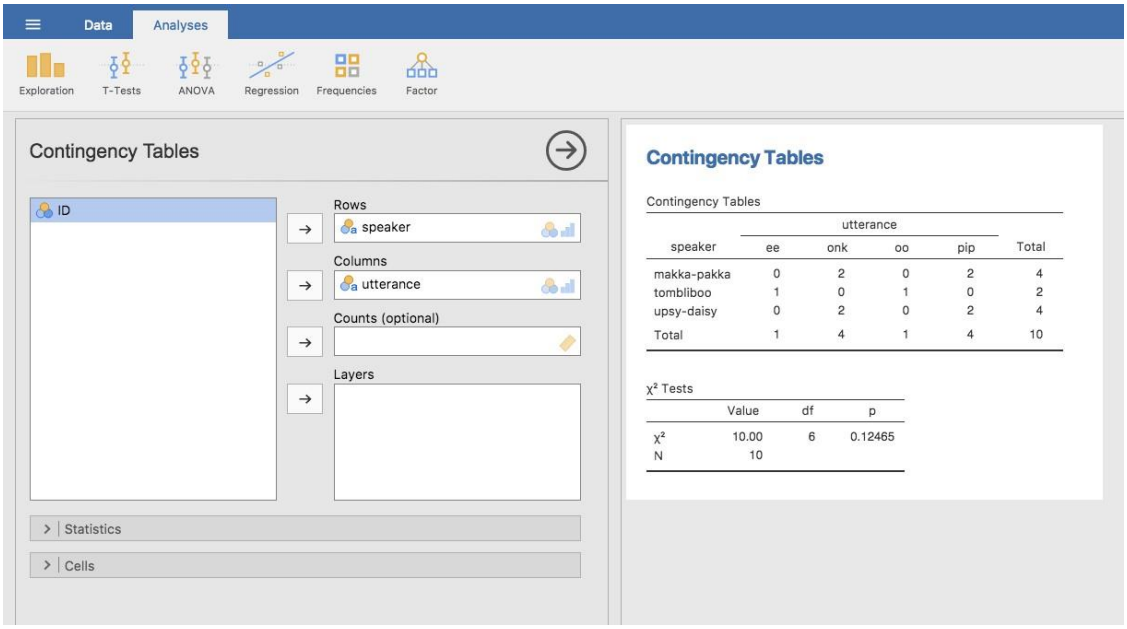


Figure 6-2 : Tableau de contingence pour le locuteur et les variables des énoncés

Ne vous inquiétez pas pour le tableau « χ^2 Tests » qui est produit. Nous reviendrons sur ce point plus loin au chapitre 10. Lors de l'interprétation du tableau de contingence, n'oubliez pas qu'il s'agit de comptages, de sorte que le fait que la première ligne et la deuxième colonne de chiffres correspondent à une valeur de 2 indique que Makka-Pakka (ligne 1) a dit « onk » (colonne 2) deux fois dans cet ensemble de données.

Ajout de pourcentages à un tableau de contingence

Le tableau de contingence illustré à la Figure 6-2 présente un tableau des fréquences brutes. C'est-à-dire, un compte du nombre total de cas pour différentes combinaisons de niveaux des variables spécifiées. Cependant, vous voulez souvent que vos données soient organisées

en termes de pourcentages aussi bien que de comptages. Vous pouvez trouver les cases à cocher pour différents pourcentages sous l'option « Cellules » dans la fenêtre « Tableaux de contingence ». Tout d'abord, cliquez sur la case à cocher « Ligne » et le tableau de contingence dans la fenêtre de sortie deviendra celui de la Figure 6-3.

Contingency Tables

		utterance				
speaker		ee	onk	oo	pip	Total
makka-pakka	Observed	0	2	0	2	4
	% within row	0%	50%	0%	50%	
tombliboo	Observed	1	0	1	0	2
	% within row	50%	0%	50%	0%	
upsy-daisy	Observed	0	2	0	2	4
	% within row	0%	50%	0%	50%	
Total	Observed	1	4	1	4	10
	% within row	10%	40%	10%	40%	

Figure 6-3 : Tableau de contingence pour le Speaker et la variable Utterrance, avec les pourcentages des lignes

Ce que nous examinons ici est le pourcentage d'énoncés faits par chaque personnage. En d'autres termes, 50% des énoncés de Makka-Pakka sont des « pip », et les 50% restants sont des « onk ». Comparons ceci avec le tableau que nous obtenons lorsque nous calculons les pourcentages des colonnes (décochez « Ligne »et cochez « Colonne » dans la fenêtre des options des cellules), voir Figure 6-4.

Contingency Tables

		utterance				
speaker		ee	onk	oo	pip	Total
makka-pakka	Observed	0	2	0	2	4
	% within column	0%	50%	0%	50%	
tombliboo	Observed	1	0	1	0	2
	% within column	100%	0%	100%	0%	
upsy-daisy	Observed	0	2	0	2	4
	% within column	0%	50%	0%	50%	
Total	Observed	1	4	1	4	10
	% within column	100%	100%	100%	100%	

Figure 6-4 : Tableau de contingence pour le speaker et le variable Utterrance, avec les pourcentages en colonnes

Dans cette version, ce que nous voyons est le pourcentage de caractères associés à chaque énoncé. Par exemple, chaque fois que l'énoncé « ee » est prononcé (dans cet ensemble de données), 100% du temps c'est un Tombliboo qui le dit.

Expressions logiques dans Jamovi

Un concept clé sur lequel s'appuient beaucoup de transformations de données dans Jamovi est l'idée d'une **valeur logique**. Une valeur logique est une affirmation sur le fait que quelque chose est vrai ou faux. Ceci est implémenté dans Jamovi d'une manière assez simple. Il y a deux valeurs logiques, à savoir VRAI et FAUX. Malgré leur simplicité, les valeurs logiques sont très utiles. Voyons comment ils fonctionnent.

Évaluer les vérités mathématiques

Dans le livre classique de George Orwell de *1984, l'*un des slogans utilisés par le Parti totalitaire était « deux plus deux égalent cinq ». L'idée étant que la domination politique de la liberté humaine devient complète lorsqu'il est possible de renverser même la vérité la plus fondamentale. C'est une pensée terrifiante, surtout quand le protagoniste Winston Smith s'effondre finalement sous la torture et accepte la proposition. « L'homme est infiniment malléable », dit le livre. Je suis presque sûr que ce n'est pas vrai pour les humains[33] et ce n'est certainement pas vrai pour Jamovi. Jamovi n'est pas infiniment malléable, il a des opinions plutôt fermes à propos de ce qui est et n'est pas vrai, du moins en mathématiques de base. Si je lui demande de calculer 2 + 2[34], il donne toujours la même réponse, et ce n'est pas 5 !

Bien sûr, pour l'instant, Jamovi ne fait que les calculs. Je ne lui ai pas demandé d'affirmer explicitement que 2 + 2 = 4 est une vraie affirmation. Si je veux que Jamovi fasse un jugement explicite, je peux utiliser une commande comme celle-ci : 2 + 2 == 4

Ce que j'ai fait ici est d'utiliser l'**opérateur d'égalité**, ==, pour forcer Jamovi à faire un jugement «vrai ou faux».[Notez qu'il s'agit d'un opérateur très différent de l'opérateur égal =. Une coquille commune que les gens font lorsqu'ils essaient d'écrire des commandes logiques en Jamovi (ou dans d'autres langues, puisque la distinction "= versus ==" est importante dans de nombreux programmes informatiques et statistiques) est de saisir accidentellement = quand vous voulez vraiment dire ==. Soyez particulièrement prudent avec cela, je programme dans plusieurs langues depuis mon adolescence et je foire *encore* beaucoup cela. Hmm. Je crois comprendre pourquoi je n'étais pas cool quand j'étais ado. Et pourquoi je ne suis toujours pas cool.] Bien, voyons ce que Jamovi pense du slogan du Parti, alors tapez ceci dans la boîte de calcul de la nouvelle variable « Formula » :

---

[33] J'offre mes tentatives d'adolescence d'être « cool » comme preuve que certaines choses ne peuvent tout simplement pas être faites.

[34] Vous pouvez le faire dans l'écran Calculer nouvelle variable, bien que calculer 2 + 2 pour chaque cellule d'une nouvelle variable ne soit pas très utile !

2 + 2 == 5

Qu'obtenez-vous ? Il devrait s'agir d'un ensemble complet de valeurs « FALSE » dans la colonne de la feuille de calcul pour votre variable nouvellement calculée. Youpi ! Liberté et poneys pour tous ! Ou quelque chose comme ça.

Quoi qu'il en soit, ça valait la peine de jeter un coup d'œil sur ce qui se passe si j'essaie de *forcer* Jamovi à croire que deux plus deux font cinq en écrivant une formule comme 2 + 2 = 5. Je sais que si je fais cela dans un autre programme, par exemple R, alors il affiche un message d'erreur. Mais attendez, si vous faites cela en Jamovi, vous obtenez tout un ensemble de valeurs « FALSE » valeurs. Alors, que se passe-t-il ? Eh bien, il semble que Jamovi est assez intelligent et réalise que vous testez si c'est VRAI ou FAUX que 2 + 2 = 5, peu importe si vous utilisez le bon **opérateur d'égalité**, ==, ou le signe égal =.

Opérations logiques

Bien maintenant, nous avons vu des opérations logiques à l'œuvre. Mais jusqu'à présent, nous n'avons vu que l'exemple le plus simple possible. Vous ne serez probablement pas surpris de découvrir que nous pouvons combiner des opérations logiques avec d'autres opérations et fonctions d'une manière plus complexe, comme celle-ci :

3*3 + 4*4 == 5*5

ou ceci

SQRT(25) == 5

Non seulement cela, mais comme l'illustre le Tableau 6-1, il existe plusieurs autres opérateurs logiques que vous pouvez utiliser correspondant à certains concepts mathématiques de base. J'espère que tous ceux-ci se comprennent d'eux-mêmes. Par exemple, le l'opérateur **moins que** < vérifie si le nombre à gauche est inférieur au nombre à droite. Si c'est moins, alors Jamovi retourne une réponse VRAI, mais si les deux nombres sont égaux, ou si celui de droite est plus grand, alors Jamovi retourne une réponse FAUX.

En revanche, l'opérateur **inférieur ou égal à** <= fera exactement ce qu'il dit. Elle retourne une valeur de VRAI si le numéro du côté gauche est inférieur ou égal au numéro du côté droit. A ce stade, j'espère que ce que font les opérateurs **plus grand que** > et **plus grand ou égal que** >= est assez évident.

Dans la liste des opérateurs logiques est l'opérateur **pas égal à** != comme avec tous les autres, fait ce qu'il dit qu'il fait. Il retourne une valeur de VRAI lorsque les choses ne sont pas identiques de part et d'autre. Par conséquent, puisque 2 + 2 n'est pas égal à 5, nous obtiendrions VRAI comme valeur pour notre variable nouvellement calculée. Essayez et vous verrez :

2 + 2! = 5

Nous n'avons pas encore tout à fait fini. Il y a trois autres opérations logiques énumérées dans le Tableau 6-2 qu'il vaut la peine de connaître. Ce sont les opérateurs **non** ! , opérateur **et** and, et l'opérateur **ou** or. Comme les autres opérateurs logiques, leur comportement est plus ou moins exactement ce à quoi on pourrait s'attendre étant donné leur nom.

Tableau 6-1 : Quelques opérateurs logiques. Techniquement, je devrais les appeler « opérateurs relationnels binaires », mais franchement, je n'en ai pas envie. C'est mon livre donc que personne ne m'y oblige.

Operation	operateur	exemple d'entrée	Résultats
moins de	<	2 < 3	VRAI
inférieur ou égal à	<=	2 <= 2	VRAI
supérieur à	>	2 > 3	FAUX
supérieur ou égal à	>=	2 >= 2	VRAI
égal à	==	2 == 3	FAUX
n'est pas égal à	!=	2 != 3	VRAI

Tableau 6-2 : Quelques opérateurs plus logiques

Opération	Opérateur	exemple d'entrée	Résultats
ne pas	NON	NOT(1==1)	FAUX
ou	ou	(1==1) or (2==3)	VRAI
et	et	(1==1) and (2==3)	FAUX

Par exemple, si je vous demande d'évaluer l'affirmation selon laquelle « 2 + 2 = 4 *ou* 2+ 2 = 5 » vous verrez que c'est vrai. Puisqu'il s'agit d'une alternative où on a « soit l'un, soit l'autre », tout ce dont nous avons besoin, c'est que l'un des deux termes soit vrai. C'est ce que fait l'opérateur or :[35]

(2+2 == 4) or (2+2 == 5)

Par contre, si je vous demande d'évaluer l'affirmation selon laquelle « 2 - 2 = 4 *et* 2 - 2 = 2 = 5 », vous diriez que c'est faux. Puisqu'il s'agit d'une conjonction, nous avons besoin que les deux parties soient vraies. Et c'est ce que fait l'opérateur and :

(2+2 == 4) and (2+2 == 5)

---

[35] Voilà une bizarrerie de Jamovi. Lorsque vous avez des expressions logiques simples comme celles que nous avons déjà rencontrées, par exemple 2 + 2 == 5, Jamovi indique clairement « FAUX » (ou « VRAI ») dans la colonne correspondante de la feuille de calcul. Lorsque nous avons des expressions logiques plus complexes, telles que (2+2 == 4) ou (2+2 == 5), Jamovi affiche simplement 0 ou 1, selon que l'expression logique est évaluée comme fausse, ou vraie.

Enfin, il y a l'opérateur *not*, qui est simple mais délicat à décrire en français. Si je vous demande de considérer mon affirmation selon laquelle « il n'est pas vrai que 2+2 = 5 », vous diriez que mon affirmation est vraie, parce qu'en fait mon affirmation est que «2+2 = 5 est fausse ». J'ai donc raison. Pour écrire ça dans Jamovi, on utilise ça :

NOT(2+2 == 5)

En d'autres termes, puisque 2+2 == 5 est une fausse déclaration, il doit être vrai que NOT(2+2 == 5) est vrai. Pour l'essentiel, ce que nous avons vraiment fait, c'est prétendre que « non faux » est la même chose que « vrai ». Évidemment, ce n'est pas tout à fait juste dans la vraie vie. Mais Jamovi vit dans un monde beaucoup plus en noir ou blanc. Pour Jamovi, tout est vrai ou faux. Aucune nuance de gris n'est autorisée.

Bien sûr, dans notre exemple 2=2 = 5, nous n'avions pas vraiment besoin d'utiliser deux opérateurs séparés « non » NOT et « Egal à » == comme. Nous aurions pu simplement utiliser l'opérateur « n'est pas égal à » != comme ça :

2+2 != 5

Application d'une opération logique au texte

Je tiens également à souligner brièvement que vous pouvez appliquer ces opérateurs logiques aussi bien au texte qu'aux données logiques. Nous devons simplement être un peu plus prudents pour comprendre comment Jamovi interprète les différentes opérations. Dans cette section, je vais parler de la façon dont l'opérateur « égal à » == s'applique au texte, puisque c'est le plus important. Évidemment, puisque l'opérateur « pas égal à » != donne les réponses exactement opposées à ==, alors je parle implicitement de lui aussi, mais je ne donnerai pas d'instructions spécifiques montrant l'utilisation de !=.

Bien, voyons comment ça marche. Dans un sens, c'est très simple. Par exemple, je peux demander à Jamovi si le mot « chat » est le même que le mot « chien », comme ceci :

« chat » == « chien »

C'est assez évident, et c'est bon de savoir que même Jamovi peut le découvrir. De même, Jamovi reconnaît qu'un « chat » est un « chat » :

« chat » == « chat »

Encore une fois, c'est exactement ce à quoi nous nous attendions. Cependant, ce que vous devez garder à l'esprit est que Jamovi n'est pas du tout tolérant quand il s'agit de grammaire et d'espacement. Si deux chaines diffèrent de quelque façon que ce soit, Jamovi dira qu'elles ne sont pas égales l'une à l'autre, comme dans les cas suivants :

« chat» ==« chat »

« chat » ==« CHAT »

« chat » == « c h a t »

Vous pouvez également utiliser d'autres opérateurs logiques. Par exemple, Jamovi vous permet également d'utiliser les opérateurs < et > pour déterminer laquelle des deux

« chaînes » de texte vient en premier, alphabétiquement parlant. En fait, c'est un peu plus compliqué que ça, mais commençons par un exemple simple :

« chat » <« chien »

Dans Jamovi, cet exemple est considéré comme VRAI. C'est parce que « chat » vient avant « chien » dans l'ordre alphabétique que Jamovi juge que la déclaration est vraie. Cependant, si nous demandons à Jamovi de nous dire si « chat » vient avant « fourmilier » alors il évaluera l'expression comme fausse. Pour l'instant, tout va bien. Mais les données textuelles sont un peu plus compliquées que ne le suggère le dictionnaire. Qu'en est-il du « CHAT » et du « chat » ? Lequel d'entre eux vient en premier ? Essayez-le et vous le saurez :

« CAT » < « cat »

Ceci est en fait évalué comme vrai. En d'autres termes, Jamovi suppose que les lettres majuscules passent avant les minuscules. D'accord, c'est exact. Personne n'en sera surpris. Ce qui peut vous surprendre, c'est que Jamovi suppose que *toutes les* lettres majuscules passent avant *toutes les* minuscules. C'est-à-dire que si « fourmilier » < « zèbre » est une affirmation vraie, et l'équivalent en majuscules « FOURMILIER » < « ZEBRE » est également vrai, il n'est pas *vrai de* dire que « fourmilier » < « ZEBRE », comme l'extrait suivant l'illustre. Essayez ceci :

« fourmilier » < « ZEBRE »

Cette proposition est considérée comme « fausse », ce qui peut sembler un peu contre-intuitif. En gardant cela à l'esprit, il peut être utile de jeter un coup d'œil rapide au Tableau 6-3 qui énumère divers caractères de texte dans l'ordre dans lequel Jamovi les traite.

Tableau 6-3 : L'ordre des différents caractères de texte utilisés par les opérateurs < et >. Le caractère « espace », qui vient en premier sur la liste, n'est pas affiché.

! « # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z [\ ] ^ _ '

a b c d e f g h i j k l m n o p q r s t u v w x y z } | {

Transformer et recoder une variable

Il n'est pas rare dans l'analyse des données du monde réel de constater que l'une de vos variables n'est pas tout à fait équivalente à la variable que vous voulez vraiment. Par exemple, il est souvent pratique de prendre une variable à valeur continue (p. ex. l'âge) et de la diviser en un petit nombre de catégories (p. ex. jeune, adulte, plus âgé). À d'autres moments, vous devrez peut-être convertir une variable numérique en une variable numérique différente (p. ex. vous voudrez peut-être analyser à la valeur absolue de la variable originale). Dans cette section, je décrirai quelques principales façons de faire ces choses avec Jamovi.

Créer une variable transformée

Le premier truc à discuter est l'idée de **transformer** une variable. Prise littéralement, *tout ce que* vous faites à une variable est une transformation, mais en pratique, cela signifie habituellement que vous appliquez une fonction mathématique relativement simple à la variable originale afin de créer une nouvelle variable qui (a) fournit une meilleure façon de décrire ce qui vous intéresse réellement, ou (b) est plus en accord avec les hypothèses des tests statistiques que vous voulez faire. Comme, à ce stade, je n'ai pas parlé des tests statistiques ou de leurs hypothèses, je vais vous montrer un exemple basé sur le premier cas.

Supposons que j'ai fait une courte étude dans laquelle je pose une seule question à 10 personnes :

Sur une échelle de 1 (fortement en désaccord) à 7 (fortement d'accord), dans quelle mesure êtes-vous d'accord avec la proposition selon laquelle « les dinosaures sont impressionnants » ?

Maintenant, chargeons et regardons les données. Le fichier de données likert.omv contient une variable unique qui contient les réponses brutes à l'échelle de Likert pour ces 10 personnes. Cependant, si vous y réfléchissez bien, ce n'est pas la meilleure façon de représenter ces réponses. En raison de la façon assez symétrique dont nous avons établi l'échelle de réponse, il y a un sens dans lequel le point médian de l'échelle aurait dû être codé 0 (sans opinion), et les deux paramètres devraient être +3 (fortement d'accord) et -3 (fortement en désaccord). En recodant les données de cette façon, on reflète un peu mieux la façon dont nous pensons vraiment aux réponses. Le recodage ici est assez simple, il suffit de soustraire 4 des scores bruts. Dans Jamovi vous pouvez le faire en calculant une nouvelle variable : cliquez sur le bouton « Data » - « Compute » et vous verrez qu'une nouvelle variable a été ajoutée à la feuille de calcul. Appelons cette nouvelle variable likert.centred (saisissez son nom) et ajoutons ce qui suit dans la boîte de formule, comme dans la Figure 6-5: « likert.raw – 4 ».

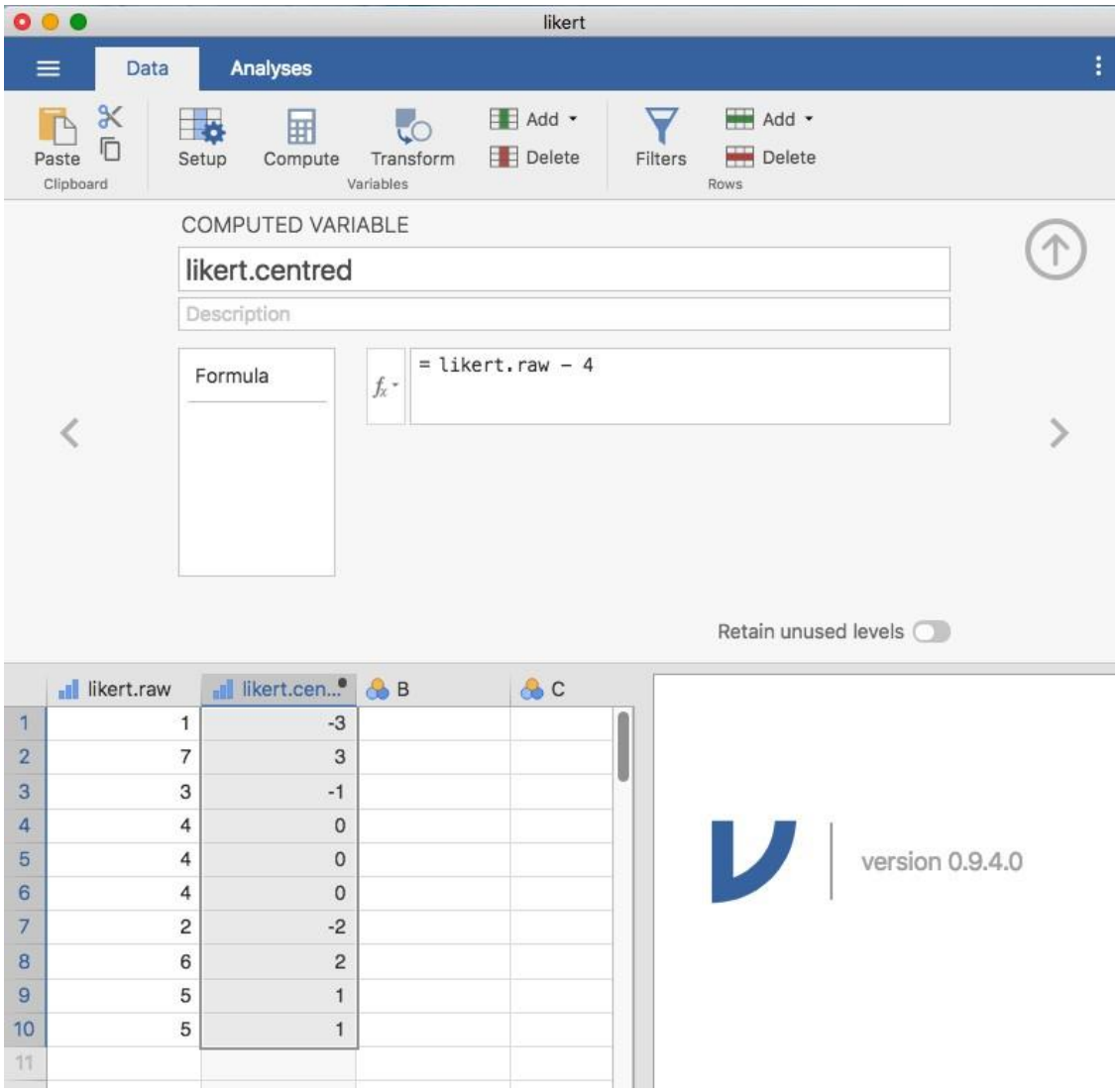


Figure 6-5 : Créer une nouvelle variable calculée dans Jamovi

L'une des raisons pour lesquelles il pourrait être utile d'avoir les données dans ce format est qu'il existe de nombreuses situations où vous pourriez préférer analyser la *force de l'*opinion séparément de la *direction de l'*opinion. Nous pouvons faire deux transformations différentes sur cette variable likert.centred afin de distinguer ces deux concepts différents. Tout d'abord, pour calculer une variable opinion.strength, nous voulons prendre la valeur absolue des données centrées (en utilisant la fonction « ABS »).[36] Dans Jamovi, créez une autre variable en utilisant le bouton « Compute ». Nommez la variable opinion.strength et cette fois, cliquez sur le bouton « *fx* » à côté de la case « Formula ». Ceci montre les différentes « Fonctions » et « Variables » que vous pouvez ajouter à la boîte « Formula », double-cliquez donc sur « ABS » puis double-cliquez sur likert.centred et vous verrez que la

[36] La valeur absolue d'un nombre est sa distance à zéro, alors que son signe est négatif ou positif.

boîte « Formula » se remplit avec ABS(likert.centred) et qu'une nouvelle variable a été créée dans la feuille de calcul, comme dans Figure 6-6:

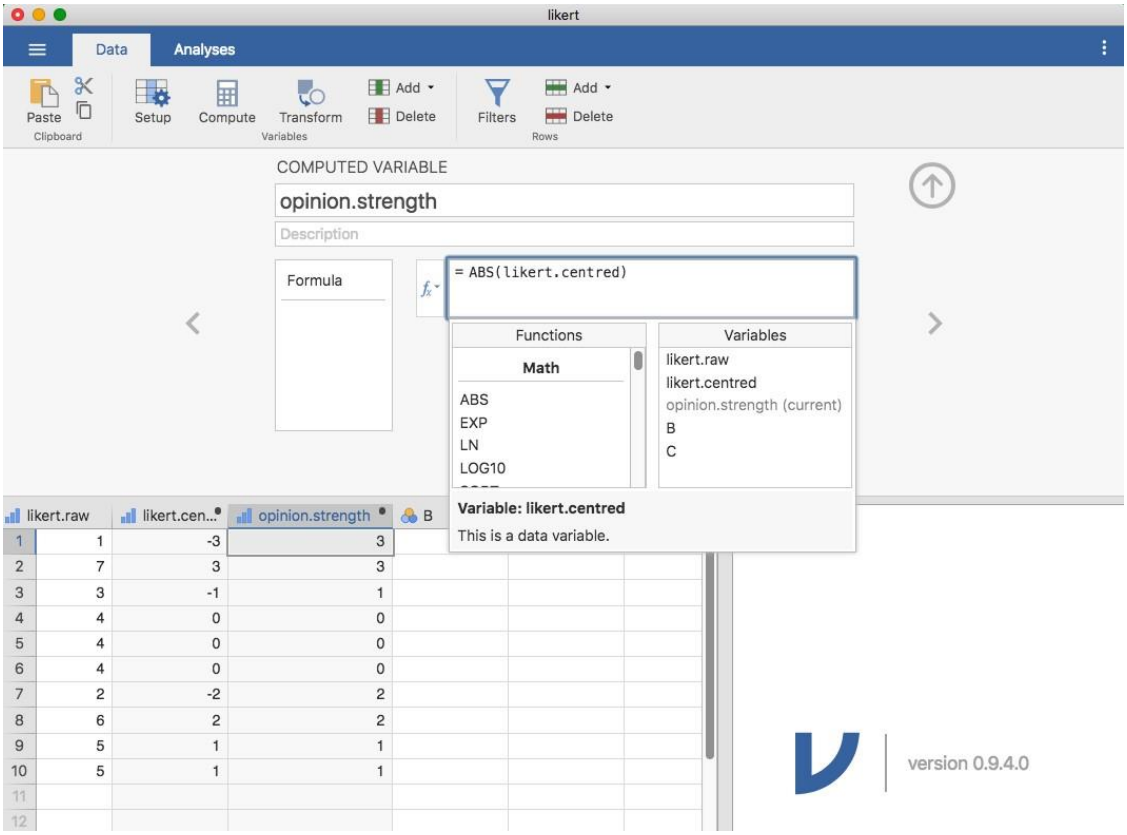


Figure 6-6 : Sélection de fonctions et de variables à l'aide du bouton fx

Deuxièmement, pour calculer une variable qui ne contient que la direction de l'opinion et ignore la force, nous voulons calculer le « signe » de la variable. Dans Jamovi nous pouvons utiliser la fonction IF pour cela. Créez une autre variable à l'aide du bouton « Compute », nommez-la en opinion.sign, puis tapez ce qui suit dans la boîte de fonction :

IF(likert.centred == 0, 0, likert.centred / opinion.strength)

Une fois fait, vous verrez que tous les nombres négatifs de la variable likert.centered sont convertis en -1, tous les nombres positifs sont convertis en 1 et zéro reste à 0, comme ici :

-1 1 -1 0 0 0 -1 1 1 1

Décomposons ce que fait cette commande « IF ». Dans Jamovi il y a trois parties à une déclaration « IF », écrite ainsi « IF(expression, value, else) ». La première partie, « expression », peut être un énoncé logique ou mathématique. Dans notre exemple, nous avons spécifié « likert.centred == 0 », ce qui est VRAI pour les valeurs où likert.centered est zéro. La partie suivante, « value », est la nouvelle valeur à retourner lorsque l'expression dans la première partie est VRAIE. Dans notre exemple, nous avons dit que pour toutes les valeurs où likert.centred est zéro, les gardez à zéro. Dans la partie suivante, « else », nous pouvons entrer une autre instruction logique ou mathématique à utiliser si la première

partie a pour résultat FAUX, c'est-à-dire les cas où likert.centred n'est pas nul. Dans notre exemple, nous avons divisé likert.centred par opinion.strength pour obtenir « -1 » ou « +1 » selon le signe de la valeur originale dans likert.centred.[37]

Et c'est fini. Nous avons maintenant trois nouvelles variables géniales, qui sont toutes des transformations utiles des données originales de likert.raw.

Réduire une variable en un plus petit nombre de niveaux ou en catégories discrètes

Une tâche pratique qui revient assez souvent est le problème du regroupement d'une variable en un plus petit nombre de niveaux ou de catégories distincts. Par exemple, supposons que je m'intéresse à la répartition par âge des participants à une réunion sociale :

60, 58, 24, 26, 34, 42, 31, 30, 33, 2, 9

Dans certaines situations, il peut être très utile de les regrouper en un petit nombre de catégories. Par exemple, nous pourrions regrouper les données en trois grandes catégories : jeunes (0-20 ans), adultes (21-40 ans) et plus âgés (41-60 ans). Il s'agit d'une classification assez grossière, et les étiquettes que j'ai jointes n'ont de sens que dans le contexte de cet ensemble de données (p. ex. en général, une personne de 42 ans ne se considérerait pas comme étant « plus âgés »). Nous pouvons découper cette variable en tranches assez facilement en utilisant la fonction Jamovi « IF » que nous avons déjà utilisée. Cette fois, nous devons spécifier des instructions « IF » imbriquées, ce qui signifie simplement que SI la première expression logique est VRAIE, insérer une première valeur, mais SI une deuxième expression logique est VRAIE, insérer une deuxième valeur, mais SI une troisième expression logique est VRAIE, insérer ensuite une troisième valeur. Cela peut s'écrire comme suit :

IF(Age >= 0 et Age <= 20, 1, IF(Age >= 21 ans et Age <= 40 ans, 2, IF(Age >= 41 ans et Age <= 60 ans, 3))))

Notez qu'il y a trois parenthèses gauches utilisées dans l'imbrication, donc l'instruction entière doit se terminer par trois parenthèses droites sinon vous aurez un message d'erreur. La capture d'écran Jamovi de cette manipulation de données, ainsi qu'un tableau de fréquences l'accompagnant, est présentée à la Figure 6-7.

---

[37] La raison pour laquelle nous devons utiliser la commande'IF' et garder zéro comme zéro est que vous ne pouvez pas simplement utiliser likert.centred / opinion.strength pour calculer le signe de likert.centrred, car la division mathématique de zéro par zéro ne fonctionne pas. Essayez-le et vous verrez

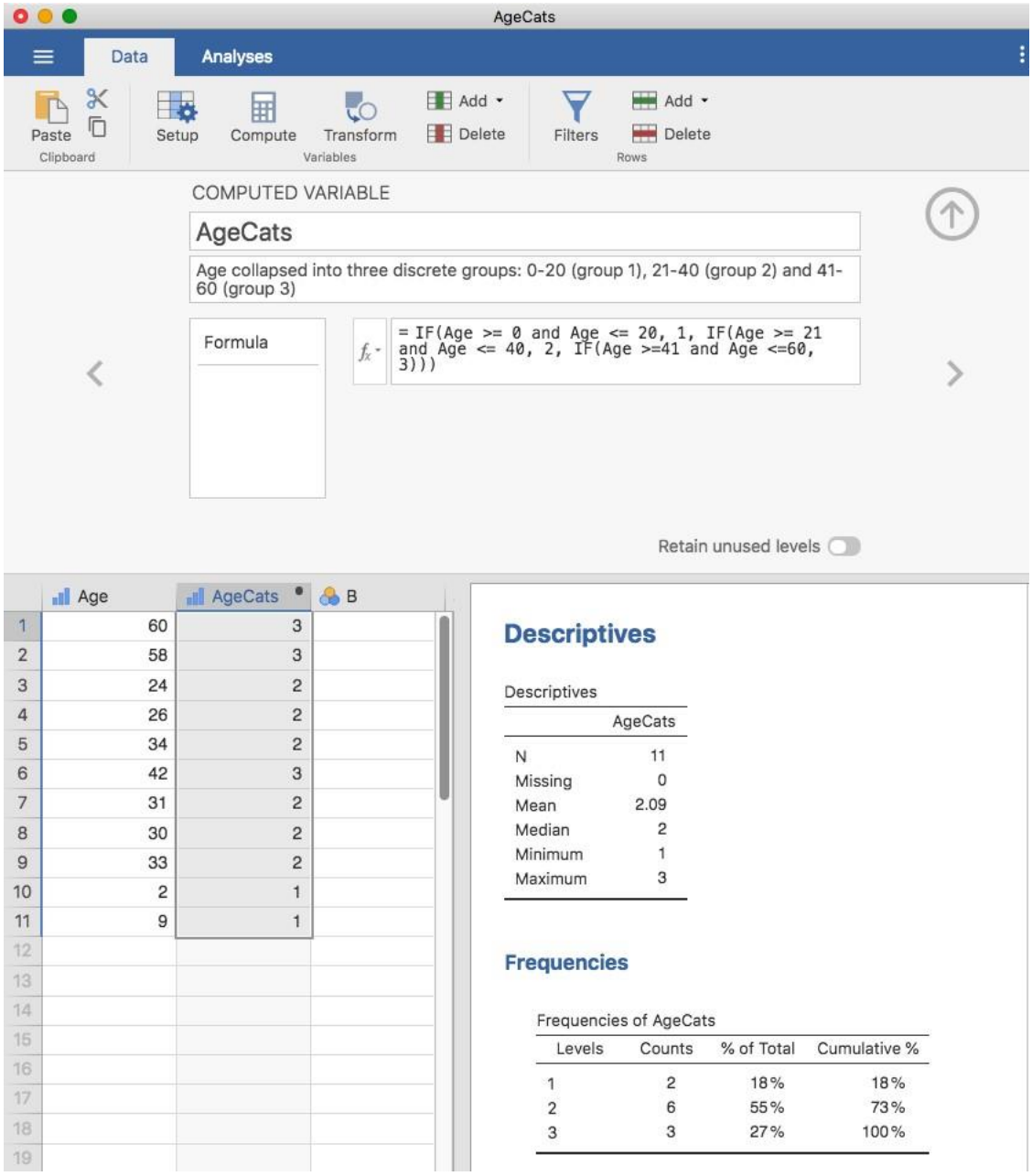


Figure 6-7 : Réduire une variable en un plus petit nombre de niveaux discrets à l'aide de la fonction « IF » de Jamovi

Il est important de prendre le temps de déterminer si les catégories qui en résultent ont un sens pour votre projet de recherche. Si elles n'ont aucun sens pour vous en tant que catégories, alors toute analyse de données qui utilise ces catégories est susceptible d'être tout aussi dénuée de sens. Plus généralement, dans la pratique, j'ai remarqué que les gens ont un désir très fort de découper leurs données (continues et désordonnées) en quelques catégories (discrètes et simples), puis d'effectuer des analyses en utilisant les données

catégorisées plutôt que les données originales.[38] Je n'irais pas jusqu'à dire qu'il s'agit d'une mauvaise idée en soi, mais elle comporte parfois des inconvénients assez graves, je vous conseille donc de faire preuve de prudence si vous envisagez de le faire.

Créer une transformation qui peut être appliquée à plusieurs variables

Parfois, vous voulez appliquer la même transformation à plus d'une variable, par exemple lorsque vous avez plusieurs items de questionnaire qui doivent tous être recalculés ou recodés de la même manière. L'une des caractéristiques intéressantes de Jamovi est que vous pouvez créer une transformation, en utilisant le bouton « Data » - « Transform », qui peut ensuite être enregistrée et appliquée à plusieurs variables. Revenons au premier exemple ci-dessus, en utilisant le fichier de données likert.omv qui contient une seule variable avec des réponses brutes à l'échelle de Likert pour 10 personnes. Pour créer une transformation que vous pouvez enregistrer puis appliquer à plusieurs variables (en supposant que vous ayez plus de variables de ce type dans votre fichier de données), sélectionnez d'abord dans l'éditeur de feuille de calcul (c'est-à-dire cliquez sur) la variable que vous voulez utiliser pour créer initialement la transformation. Dans notre exemple, il s'agit de likert.raw. Cliquez ensuite sur le bouton « Transform » dans le ruban Jamovi « Data », et vous devriez avoir quelque chose comme sur la Figure 6-8.

Donnez un nom à votre nouvelle variable, appelons-la opinion.strength, puis cliquez sur la case de sélection « Using transform » et sélectionnez « Create New Transform… ». C'est ici que vous allez créer, et nommer, la transformation qui peut être réappliquée à autant de variables que vous le souhaitez. La transformation est automatiquement nommée pour nous comme « Transform 1 » (Bien pensé, non ? Vous pouvez changer ceci si vous voulez). Tapez ensuite l'expression « ABS($source - 4) » dans la zone de texte de la fonction, comme dans la Figure 6-9, appuyez sur Entrée ou Retour sur votre clavier et, rapidement, vous avez créé une nouvelle transformation et l'avez appliquée à la variable likert.raw ! Bien ! Notez qu'au lieu d'utiliser l'étiquette de la variable dans l'expression, nous avons plutôt utilisé « $source ». C'est pour que nous puissions ensuite utiliser la même transformation avec autant de variables différentes que nous le souhaitons - Jamovi vous demande d'utiliser « $source » pour faire référence à la variable source que vous transformez. Votre transformation a également été sauvegardée et peut être réutilisée à tout moment (à

---

[38] Si vous avez lu plus loin dans le livre, et que vous relisez cette section, alors un bon exemple de cela serait que quelqu'un choisisse de faire une analyse de variance en utilisant AgeCats comme variable de regroupement, au lieu de faire une régression en utilisant Age comme prédicteur. Il y a parfois de bonnes raisons de le faire. Par exemple, si la relation entre l'âge et votre variable de résultat est très non linéaire et que vous n'êtes pas à l'aise d'essayer de faire une régression non linéaire ! Cependant, à moins que vous n'ayez vraiment une bonne raison de le faire, il vaut mieux ne pas le faire. Elle tend à introduire toutes sortes d'autres problèmes (par exemple, les données violeront probablement l'hypothèse de normalité) et vous pouvez perdre beaucoup de puissance statistique.

condition que vous sauvegardiez l'ensemble de données dans un fichier « .omv », sinon vous la perdrez !

Vous pouvez également créer une transformation à l'aide du deuxième exemple que nous avons examiné, la répartition par âge des participants à une réunion sociale. Essayez si vous en avez envie ! N'oubliez pas que nous avons regroupé cette variable en trois groupes : les jeunes, les adultes et les plus âgés. Cette fois-ci, nous allons faire la même chose, mais en utilisant le bouton « Transform » - « Add condition » de Jamovi. Avec cet ensemble de données (y retourner ou le créer à nouveau si vous ne l'avez pas sauvegardé), configurez une nouvelle transformation de variable. Appelez la variable transformée AgeCats et la transformation que vous allez créer Agegroupings. Cliquez ensuite sur le grand signe « + » à côté de la case de fonction. C'est le bouton « Add condition » et j'ai collé une grosse flèche rouge sur la Figure 6-10 pour que vous puissiez voir exactement où cela se trouve. Recréez la transformation illustrée à la Figure 6-10 et lorsque vous aurez terminé, vous verrez apparaître les nouvelles valeurs dans la fenêtre du tableur.

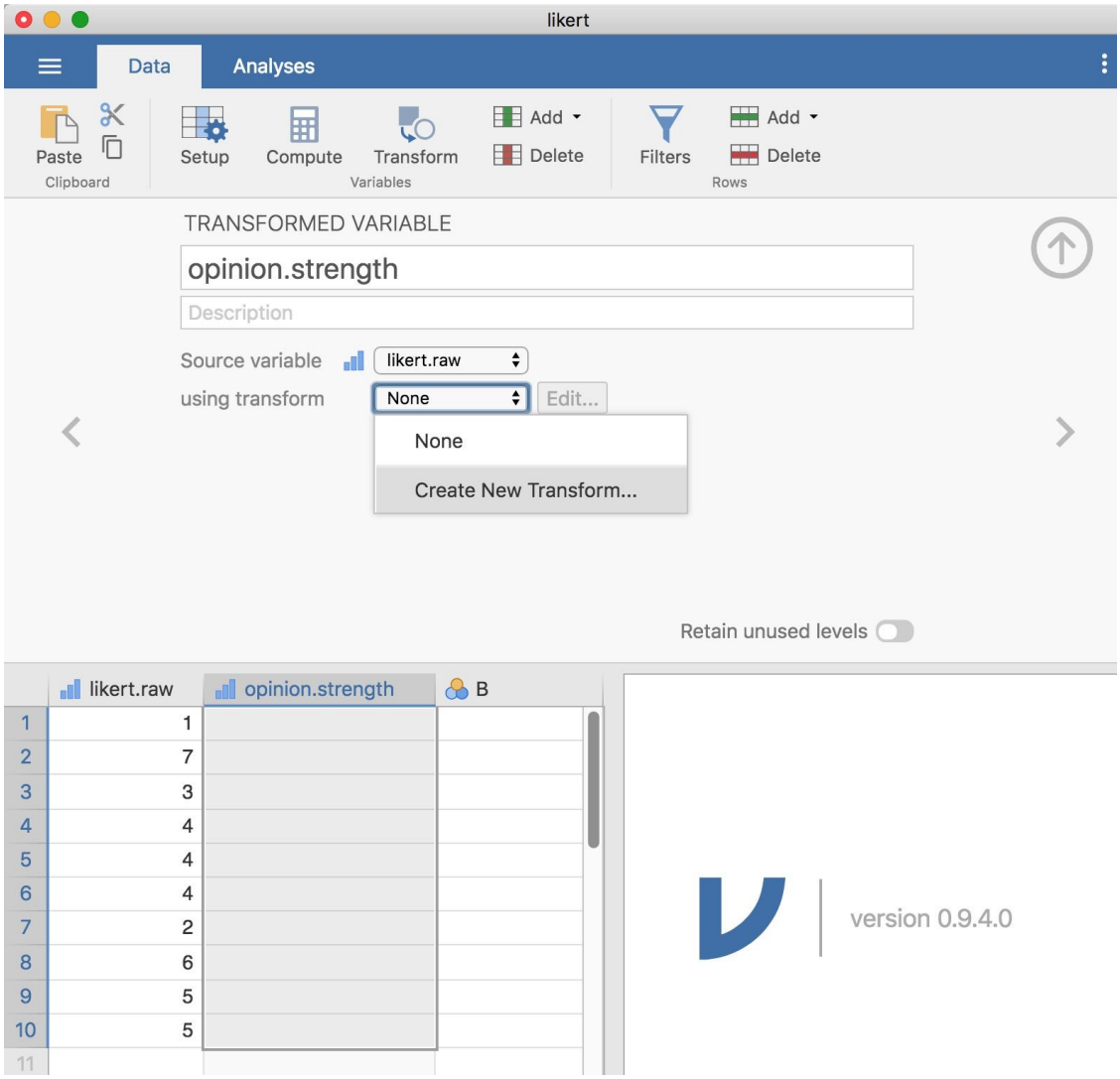


Figure 6-8 : Création d'une nouvelle transformation de variable à l'aide de la commande Jamovi « Transform ».

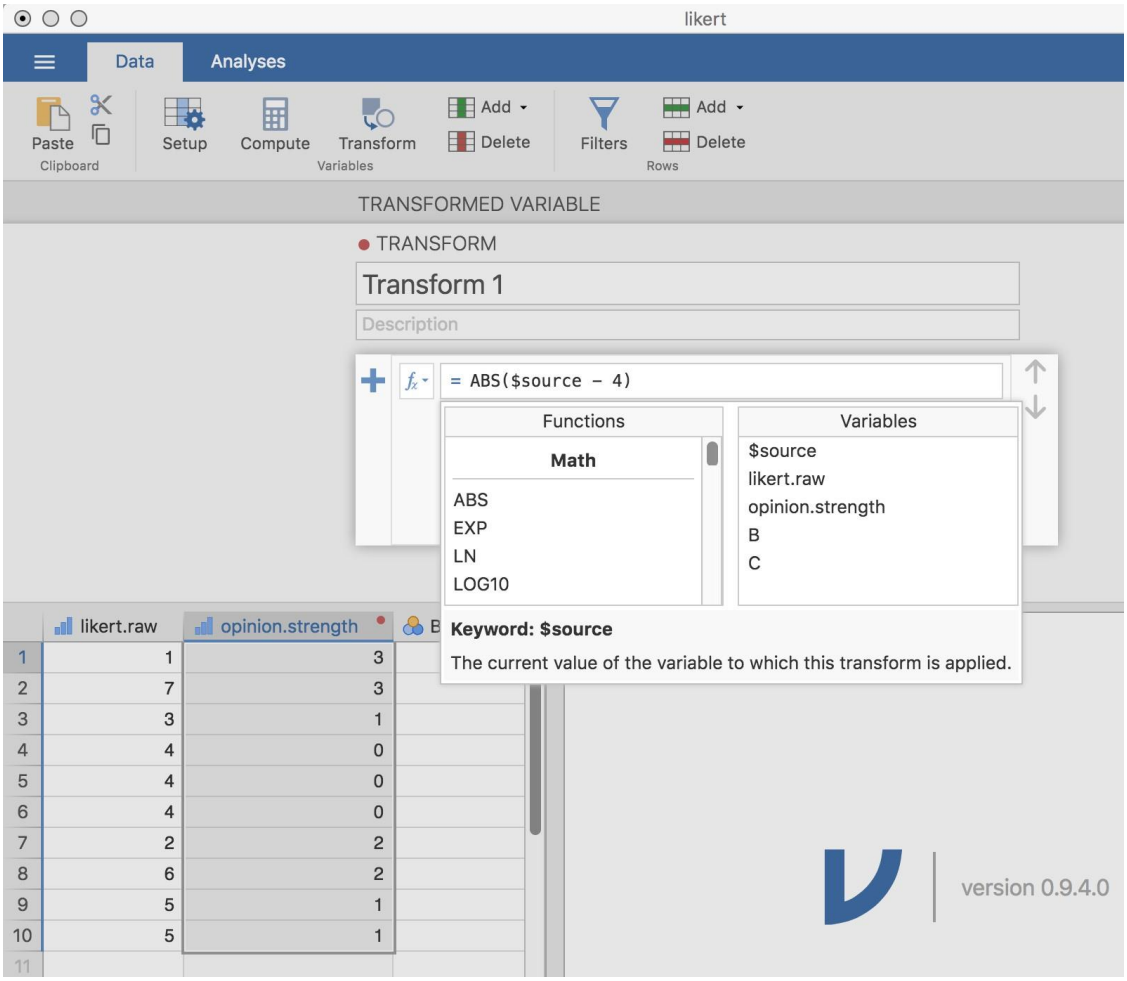


Figure 6-9 : Spécification d'une transformation dans Jamovi, à sauvegarder sous le nom imaginaire « Transform 1 ».

De plus, la transformation des groupes d'âge a été sauvegardée et peut être réappliquée à tout moment. Entendu, je sais qu'il est peu probable que vous ayez plus d'une variable « Age », mais vous savez maintenant comment configurer les transformations dans Jamovi, donc vous pouvez suivre cet exemple avec d'autres types de variables. Un scénario typique est celui où vous avez un questionnaire avec des échelles, disons, 20 items (variables) et chaque item a été initialement noté de 1 à 6 mais, pour une raison ou une autre, vous décidez de recoder tous les items de 1 à 3. Vous pouvez facilement le faire avec Jamovi en créant puis en appliquant de nouveau votre transformation pour chaque variable que vous voulez recoder.

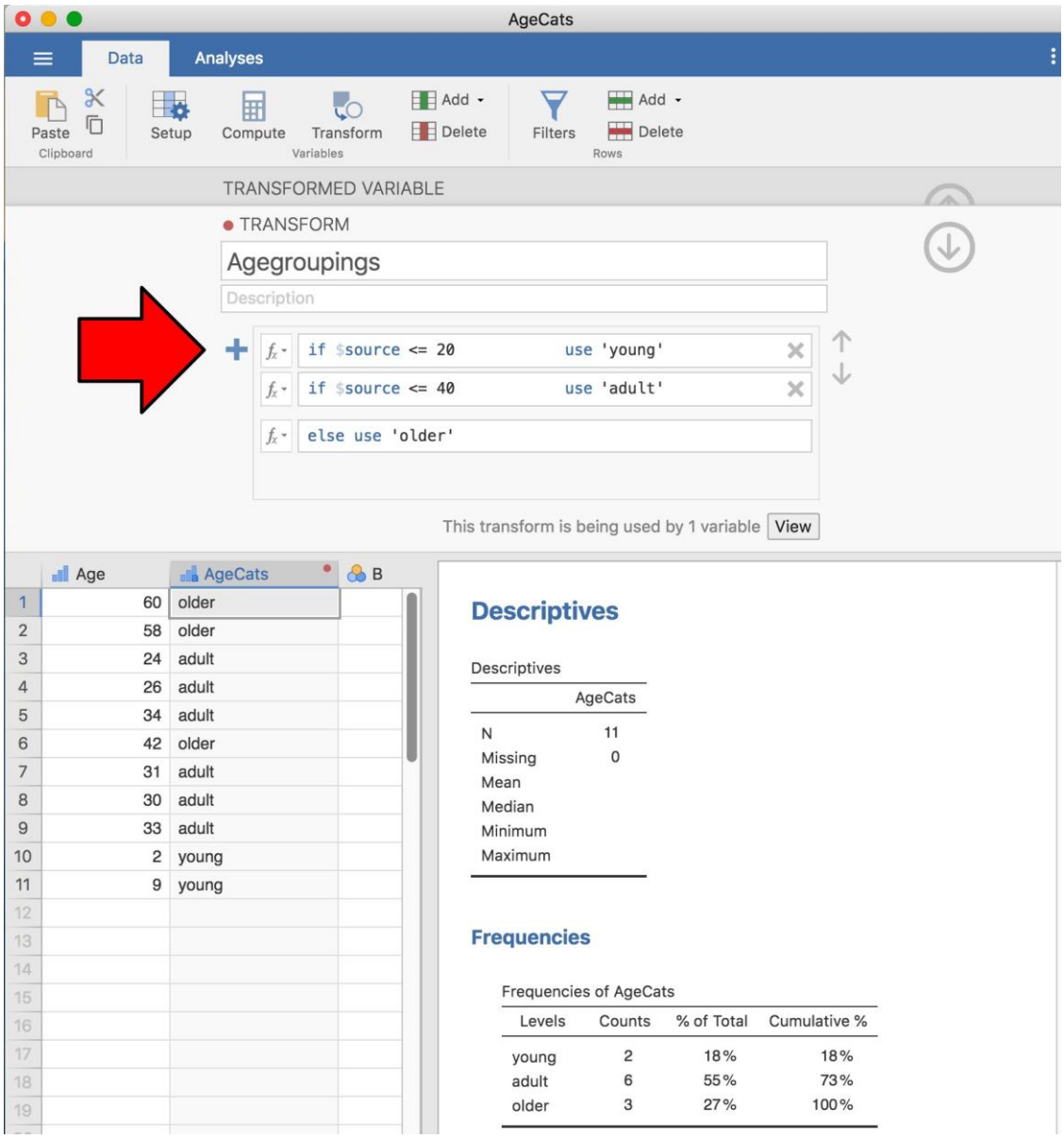


Figure 6-10 : Transformation de Jamovi en trois catégories d'âge, à l'aide du bouton 'Ajouter condition'.

Tableau 6-4 : Certaines des fonctions mathématiques disponibles dans Jamovi

	fonction	exemple d'entrée	(réponse)
racine carrée	SQRT(x)	SQRT(25)	5
valeur absolue	ABS(x)	ABS(-23)	23
logarithme (base 10)	LOG10(x)	LOG10(1000)	3
logarithme (base *e*)	LN(x)	LN(1000)	6.908

exponentiation	EXP(x)	EXP(6.908)	1000.245
box-cox	BOXCOX(x, lamda)	BOXCOX(6.908, 3)	109.551

Quelques fonctions et opérations mathématiques supplémentaires

Dans la section 6.3, j'ai discuté des idées qui sous-tendent les transformations des variables et j'ai montré qu'un grand nombre des transformations que vous pourriez vouloir appliquer à vos données sont basées sur des fonctions et opérations mathématiques assez simples. Dans cette section, je veux revenir sur cette discussion et mentionner plusieurs autres fonctions mathématiques et opérations arithmétiques qui sont en fait très utiles pour beaucoup d'analyses de données réelles. Le Tableau 6-4 donne un bref aperçu des diverses fonctions mathématiques dont je veux parler ici ou plus loin.[39] Évidemment, cela ne constituent pas un catalogue complet des possibilités disponibles, mais cela couvre un éventail de fonctions qui sont utilisées régulièrement dans l'analyse des données et qui sont disponibles dans Jamovi.

Logarithmes et exponentielles

Comme je l'ai mentionné plus tôt, Jamovi possède une gamme pratique de fonctions mathématiques intégrées et il ne serait pas vraiment utile d'essayer de les décrire ou même de les énumérer toutes. Pour l'essentiel, je me suis concentré uniquement sur les fonctions qui sont strictement nécessaires pour ce livre. Cependant, je veux faire une exception pour les logarithmes et les exponentielles. Bien qu'ils ne soient nécessaires nulle part ailleurs dans ce livre, ils sont *partout* dans les statistiques. En plus, il y a *beaucoup de* situations dans lesquelles il est pratique d'analyser le logarithme d'une variable (c'est-à-dire de prendre une « log-transformation » de la variable). Je soupçonne que beaucoup (peut-être la plupart) des lecteurs de ce livre ont déjà rencontré des logarithmes et des exponentielles auparavant, mais d'après mon expérience passée, je sais qu'il y a une proportion importante d'étudiants qui suivent un cours de statistiques en sciences sociales et qui n'ont pas touché aux logarithmes depuis le secondaire, et j'aimerais faire un petit rappel.

Pour comprendre les logarithmes et les exponentielles, le plus simple est de les calculer et de voir comment ils se rapportent à d'autres calculs simples. Il y a trois fonctions Jamovi en particulier dont je veux parler, à savoir LN(), LOG10() et EXP(). Pour commencer, considérons LOG10(), qui est connu sous le nom de « logarithme en base 10 ». L'astuce pour comprendre un **logarithme** est de comprendre qu'il s'agit essentiellement du « contraire » l'élévation à la puissance. Plus précisément, le logarithme en base 10 est étroitement lié aux puissances de 10. Commençons donc par noter que 10 au cube, c'est 1000. Mathématiquement, on écrirait ceci :

$$10^3 = 1000$$

[39] Nous laisserons la fonction box-cox à plus tard, voir section 12.10.4.

L'astuce pour comprendre un logarithme est de reconnaître que l'affirmation que « 10 à la puissance de 3 est égale à 1000 » est équivalente à l'affirmation que « le logarithme (en base 10) de 1000 est égal à 3 ». Mathématiquement, nous écrivons ceci comme suit,

$$(1000) - 3$$

Bien, puisque la fonction LOG10() est liée aux puissances de 10, vous pouvez vous attendre à ce qu'il y ait d'autres logarithmes (dans des bases autres que 10) qui sont également liés aux autres puissances. Et bien sûr, c'est vrai : il n'y a rien de mathématiquement spécial dans le chiffre 10. Il se trouve que vous et moi le trouvons utile parce que les nombres décimaux sont construits autour du chiffre 10, mais le terrible monde des mathématiques se moque de nos nombres décimaux. Malheureusement, l'univers ne se soucie pas vraiment de la façon dont nous écrivons les chiffres. Quoi qu'il en soit, la conséquence de cette indifférence cosmique est qu'il n'y a rien de particulier à calculer les logarithmes en base 10. Vous pourriez, par exemple, calculer vos logarithmes en base 2. Alternativement, un troisième type de logarithme, et on en voit beaucoup plus dans les statistiques que la base 10 ou la base 2, s'appelle le **logarithme naturel**, et correspond au logarithme de la base *e*. Comme vous pourriez un jour le rencontrer, je ferais mieux de vous expliquer ce qu'est *e*. Le nombre *e*, connu sous le nom de **nombre d'Euler**, est l'un de ces nombres « irrationnels » ennuyeux dont l'expansion décimale est infiniment longue, et est considéré comme l'un des nombres les plus importants en mathématiques. Les premiers chiffres de *e* sont :

e = 2,718282

Il y a pas mal de situation dans les statistiques qui nous obligent à calculer les puissances de *e*, bien qu'aucun d'entre eux n'apparaissent dans ce livre. Elever *e* à la puissance *x* s'appelle l'**exponentielle** de *x*, et il est donc très commun de voir *ex* écrit comme exp(x). Il n'est donc pas surprenant que Jamovi ait une fonction qui calcule les exponentielles, appelée EXP(). Étant donné que le nombre *e* apparaît si souvent dans les statistiques, le logarithme naturel (c.-à-d. le logarithme en base *e*) a aussi tendance à apparaître. Les mathématiciens l'écrivent souvent comme loge(x) ou ln(x). En fait, Jamovi fonctionne de la même manière : la fonction LN() correspond au logarithme naturel.

Et avec ça, je pense que nous avons eu assez d'exponentielles et de logarithmes pour ce livre !

Extraction d'un sous-ensemble de données

Un type très important de traitement des données est la possibilité d'extraire un sous-ensemble particulier de données. Par exemple, vous pourriez n'être intéressé que par l'analyse des données d'une condition expérimentale, ou vous pourriez vouloir examiner de près les données de personnes âgées de plus de 50 ans. Pour ce faire, la première étape consiste à faire filtrer avec Jamovi le sous-ensemble des données correspondant aux observations qui vous intéressent.

Cette section revient sur l'ensemble de données nightgarden.csv. Si vous lisez tout ce chapitre en une seule fois, alors vous devriez déjà avoir cet ensemble de données chargé dans une fenêtre Jamovi. Pour cette section, concentrons-nous sur les deux variables

speaker et Utterance (voir Section 6.1 si vous avez oublié à quoi ressemblent ces variables). Supposons que ce que je veuille faire, c'est retirer seulement les énoncés qui ont été faits par Makka-Pakka. Pour cela, nous devons spécifier un filtre dans Jamovi. Ouvrez d'abord une fenêtre de filtre en cliquant sur « Filters » dans la barre d'outils principale Jamovi « Data ». Puis, dans la zone de texte « Filter 1 », à côté du signe « = », tapez ce qui suit :

speaker ==‘makka-pakka’.

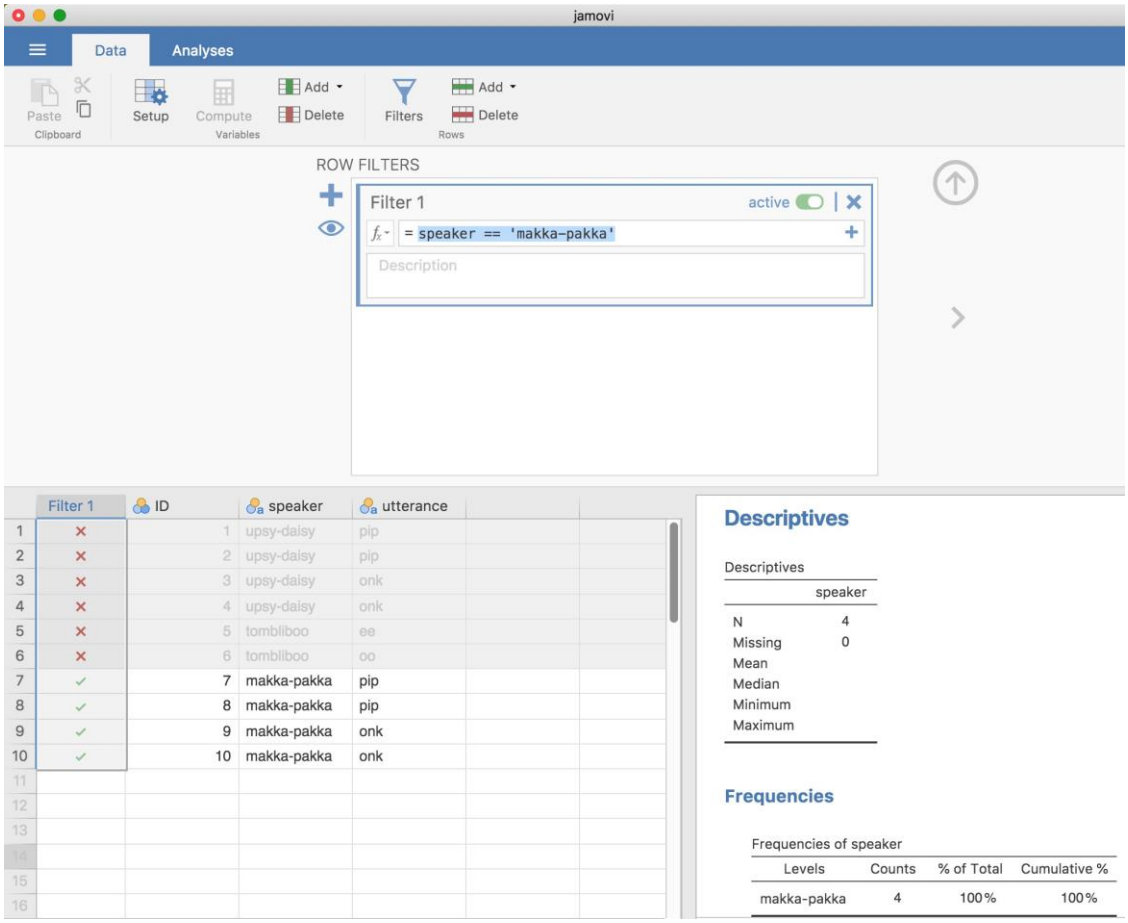


Figure 6-11 : Création d'un sous-ensemble de données de jardin de nuit à l'aide de l'option Jamovi ‘Filters’.

Une fois cette opération terminée, vous verrez qu'une nouvelle colonne a été ajoutée à la fenêtre du tableur (voir Figure 6-11), intitulée « Filter 1 », avec les cas où le speaker *n'est pas* « makka-pakka « grisée (c'est-à-dire filtrée) et, inversement, où le speaker *est* « makka-pakka «, avec une coche verte indiquant que le filtre est activé. Vous pouvez tester ceci en exécutant « Exploration''Descriptifs » - « Tableaux de fréquences » pour la variable speaker et voir ce que cela indique. Allez-y, essayez-le !

En suivant cet exemple simple, vous pouvez aussi construire des filtres plus complexes en utilisant des expressions logiques de Jamovi. Par exemple, supposons que je veuille garder

seulement les cas où l'énoncé est « pip » ou « oo ». Dans ce cas, dans la zone de texte « Filter 1 », à côté du signe « = », vous devez taper ce qui suit :

prononciation ==‘pip’ ou prononciation ==‘oo’.

Résumé

Il est évident que ce chapitre n'a pas vraiment de cohérence. C'est juste un ensemble de sujets et d'astuces qu'il peut être utile de connaître, alors le meilleur résumé que je puisse donner ici est de répéter cette liste :

- Section 6.1. Mise en tableau des données.
- Section 6.2. Utiliser des expressions logiques.
- Section 6.3. Transformer ou recoder une variable.
- Section 6.4. Quelques fonctions mathématiques utiles.
- Section 6.5. Extraction d'un sous-ensemble d'un ensemble de données.

Introduction à la probabilité

[Dieu] ne nous a accordé que le crépuscule de la Probabilité. - John Locke

Jusqu'à maintenant, nous avons discuté de certaines des idées clés de la conception expérimentale, et nous avons parlé un peu de la façon dont vous pouvez résumer un ensemble de données. Pour beaucoup de gens, c'est tout ce qu'il y a à faire avec les statistiques : recueillir tous les chiffres, calculer les moyennes, dessiner des graphiques et les mettre toutes quelque part dans un rapport. C'est un peu comme la philatélie, mais avec des chiffres. Cependant, les statistiques c'est bien plus que cela. Cependant, les statistiques englobent beaucoup plus que cela. En fait, la statistique descriptive est l'une des plus petites composantes de la statistique et l'une des moins puissantes. La plus importante et la plus utile caractéristique des statistiques est qu'elles fournissent des informations qui vous permettent de faire des inférences sur les données.

Une fois que l'on commence à penser aux statistiques en ces termes, que les statistiques sont là pour nous aider à tirer des conclusions à partir des données, on commence à en voir des exemples partout. Par exemple, voici un petit extrait d'un article paru dans le Sydney Morning Herald (30 octobre 2010) :

« J'ai un travail difficile », a déclaré le Premier ministre en réponse à un sondage qui a révélé que son gouvernement est maintenant l'administration travailliste la plus impopulaire de l'histoire des sondages, avec un vote lors de la primaire de seulement 23 pour cent ».

Ce genre de remarque est tout à fait banale dans les journaux ou dans la vie de tous les jours, mais réfléchissons un peu à ce qu'elle implique. Une société de sondage a effectué une enquête, habituellement assez important parce qu'elle peut se le permettre. Je suis trop paresseux pour restituer l'enquête originale, alors imaginons qu'ils ont appelé 1000 électeurs de Nouvelle-Galles du Sud (NSW) au hasard, et 230 (23%) d'entre eux ont déclaré avoir l'intention de voter pour le Parti travailliste australien (ALP). Pour les élections fédérales de 2010, la Commission électorale australienne a déclaré 4 610 795 électeurs

inscrits en Nouvelle-Galles du Sud, de sorte que les opinions des 4 609 795 électeurs restants (environ 99,98 % des électeurs) nous sont inconnues. Même en supposant que personne n'a menti à la société de sondage, la seule chose que nous pouvons dire avec 100% de confiance est que le vrai vote primaire ALP se situe entre 230/4610795 (environ 0,005%) et 4610025/4610795 (environ 99,83%). Alors, sur quelle base est-il légitime pour la société de sondage, le journal et le lectorat de conclure que le vote à la primaire de l'ALP n'est que d'environ 23% ?

La réponse à la question est assez évidente. Si j'appelle 1000 personnes au hasard et que 230 d'entre elles disent qu'elles ont l'intention de voter pour l'ALP, il semble très peu probable qu'il s'agisse des 230 *seules* personnes sur l'ensemble des électeurs qui ont l'intention de voter. En d'autres termes, nous supposons que les données recueillies par la société de sondage sont assez représentatives de la population en général. Mais quelle représentativité ? Serait-on surpris d'apprendre que le véritable vote à la primaire de l'ALP est en fait de 24% ? 29% ? 37% ? C'est à ce moment-là que l'intuition quotidienne commence à s'effondrer un peu. Personne ne serait surpris de 24 p. 100 et tout le monde serait surpris de 37 p. 100, mais il est un peu difficile de dire si 29 p. 100 est plausible. Nous avons besoin d'outils plus puissants que le simple examen des chiffres et des hypothèses.

Les statistiques inférentielles fournissent les outils dont nous avons besoin pour répondre à ce genre de questions, et puisque ces questions sont au cœur de l'entreprise scientifique, elles occupent la part du lion dans chaque cours d'introduction à la statistique et aux méthodes de recherche. Cependant, la théorie de l'inférence statistique est construite sur la **théorie des probabilités**. Et c'est vers la théorie des probabilités que nous devons maintenant nous tourner. Cette discussion de la théorie des probabilités est essentiellement un détail de fond. Il n'y a pas beaucoup de statistiques en soi dans ce chapitre, et vous n'avez pas besoin de comprendre ce matériel aussi en profondeur que les autres chapitres de cette partie du livre. Néanmoins, comme la théorie des probabilités sous-tend une grande partie des statistiques, cela vaut la peine d'en aborder certains aspects fondamentaux.

En quoi la probabilité et les statistiques sont-elles différentes ?

Avant de commencer à parler de la théorie des probabilités, il est utile de réfléchir un moment à la relation entre probabilité et statistiques. Les deux disciplines sont étroitement liées, mais elles ne sont pas identiques. La théorie des probabilités est « la doctrine des chances ». C'est une branche des mathématiques qui vous indique la fréquence à laquelle différents types d'événements se produisent. Par exemple, toutes ces questions sont des choses auxquelles vous pouvez répondre en utilisant la théorie des probabilités :

- Quelles sont les chances qu'une belle pièce de monnaie tombe sur face 10 fois de suite ?
- Si je lance deux fois un dé à six faces, quelle est la probabilité que je lance deux six ?
- Quelle est la probabilité que cinq cartes tirées d'un jeu de cartes parfaitement mélangées soient toutes des cœurs ?
- Quelles sont les chances que je gagne à la loterie ?

Notez que toutes ces questions ont quelque chose en commun. Dans chaque cas, la « vérité du monde » est connue et ma question porte sur le « genre d'événements » qui va se produire. Dans la première question, je *sais que* la pièce de monnaie est correcte et qu'il y a donc 50 p. 100 de chances que n'importe quelle pièce de monnaie tombe sur pile ou face. Dans la deuxième question, je *sais que* la chance d'avoir un 6 sur un seul dé est de 1 sur 6. Dans la troisième question, je *sais que* le jeu est bien mélangé. Et dans la quatrième question, je *sais que* la loterie suit des règles spécifiques. Vous comprenez l'idée. Le point critique est que les questions probabilistes partent d'un **modèle** connu du monde, et nous utilisons ce modèle pour faire certains calculs. Le modèle sous-jacent peut être assez simple. Par exemple, dans l'exemple de pile ou face, nous pouvons écrire le modèle comme ceci :

$$p(\text{face}) = 0{,}5$$

Ce que l'on peut lire comme « la probabilité de face est de 0,5 ». Comme nous le verrons plus loin, de la même façon que les pourcentages sont des nombres qui vont de 0 % à 100 %, les probabilités ne sont que des nombres qui vont de 0 à 1. Lorsque j'utilise ce modèle de probabilité pour répondre à la première question, je ne sais pas exactement ce qui va se passer. J'aurai peut-être 10 faces, comme le dit la question. Mais peut-être que j'aurai trois faces. C'est le point clé. En théorie des probabilités, le *modèle* est connu, mais les *données* ne le sont pas.

C'est donc ça, la probabilité. Qu'en est-il des statistiques ? Les questions statistiques fonctionnent dans l'autre sens. Dans les statistiques, nous *ne* connaissons *pas la* vérité sur le monde. Tout ce que nous avons, ce sont les données et c'est à partir de ces données que nous voulons *apprendre la* vérité sur le monde. Les questions statistiques ont tendance à ressembler davantage à celles-ci :

- Si mon ami tire à pile ou face 10 fois et obtient 10 face, est-ce qu'il me joue un tour ? *Si cinq cartes sur le dessus du paquet sont toutes des cœurs, quelle est la probabilité que le paquet ait été mélangé ?* Si le conjoint du commissaire de la loterie gagne, quelle est la probabilité que la loterie ait été truquée ?

Cette fois, nous n'avons que des données. Ce que je *sais,* c'est que j'ai vu mon ami tirer à pile ou face 10 fois et qu'à chaque fois, il y a eu des problèmes. Et ce que je veux en **déduire,** c'est si oui ou non, je dois conclure que ce que je viens de voir est en fait une pièce de monnaie qui a été lancée à pile ou face 10 fois de suite, ou si je dois soupçonner que mon ami me joue un tour. Les données que j'ai ressemblent à ceci :

F F F F F F F F F F

Ce que j'essaie de faire, c'est de trouver en quel « modèle du monde » je dois avoir confiance. Si la pièce est juste alors le modèle que je devrais adopter est celui qui dit que la probabilité de faces est de 0,5, c'est-à-dire

$$p(\text{face}) = 0{,}5)$$

. Si la pièce n'est pas correcte alors je devrais conclure que la probabilité de face *n'*est *pas de* 0,5, ce que nous écririons comme

$$p(face) \neq 0{,}5$$

. En d'autres termes, le problème de l'inférence statistique consiste à déterminer lequel de ces modèles de probabilité est le bon. De toute évidence, la question statistique n'est pas la même que la question de probabilité, mais elles sont étroitement liées les unes aux autres. C'est pourquoi, une bonne introduction à la théorie statistique commencera par une discussion sur ce qu'est la probabilité et comment elle fonctionne.

Que signifie probabilité ?

Commençons par la première de ces questions. Qu'est-ce que la « probabilité » ? Cela peut vous surprendre, mais bien que les statisticiens et les mathématiciens s'entendent (pour la plupart) sur les *règles de* probabilité, il y a beaucoup moins de consensus sur ce que le mot *signifie* vraiment. Cela semble bizarre parce que nous sommes tous très à l'aise d'utiliser des mots comme « chance », « probable », « possible » et « probable », et il ne semble pas que ce soit une question à laquelle il devrait être très difficile de répondre. Mais si vous avez déjà vécu cette expérience dans la vie réelle, vous pourriez avoir l'impression de quitter la discussion en ayant le sentiment que vous ne l'avez pas bien comprise, et que (comme beaucoup de concepts courants) vous ne savez pas vraiment de quoi il s'agit.

Alors je vais essayer. Supposons que je veuille parier sur un match de football entre deux équipes de robots, *Arduino Arsenal* et *C Milan*. Après y avoir réfléchi, je décide qu'il y a 80% de chances qu'*Arduino Arsenal* gagne. Qu'est-ce que je veux dire par là ? Voici trois possibilité :

- Ce sont des équipes robotisées, donc je peux les faire jouer encore et encore, et si je le faisais, *Arduino Arsenal* gagnerait en moyenne 8 matchs sur 10.
- Pour n'importe quel jeu donné, je serais d'accord que parier sur ce jeu n'est « satisfaisant » que si un pari de $1 sur *C Milan* donne un gain de $5 (i.e. je reçois mon $1 plus un gain de $4 pour être juste), tout comme un pari de $4 sur *Arduino Arsenal* (i.e., mes $4 de mise plus une récompense de $1).
- Ma « croyance » subjective ou « confiance » en une victoire de l'*Arduino Arsenal* est quatre fois plus forte que ma croyance en une victoire du *C Milan.*

Chacune d'entre elles semble raisonnable. Cependant, ils ne sont pas identiques et les statisticiens ne les approuveraient pas tous. La raison à cela est qu'il existe différentes idéologies statistiques (oui, vraiment !) et selon celle dans laquelle vous souscrivez, vous pourriez dire que certaines de ces affirmations sont dénuées de sens ou non pertinentes. Dans cette section, je donne une brève introduction aux deux principales approches qui existent dans la littérature. Ce ne sont pas les seules approches, mais ce sont les deux plus importantes.

La vision fréquentiste

La première des deux grandes approches de la probabilité, et la plus dominante en statistique, est appelée le **point de vue fréquentiste** et elle définit la probabilité comme une **fréquence récurrente**. Supposons que nous essayions de jouer à pile ou face encore et

encore. Par définition, il s'agit d'une pièce de monnaie qui a $p(\text{face}) = 0{,}5$. Que pourrions-nous observer ? Une possibilité est que les 20 premiers lancers ressemblent à ceci :

```
P, F, F, F, F, P, P, F, F, F, F, P, F, F, P, P, P, P, P, F
```

Dans ce cas, 11 de ces 20 tours de pièces de monnaie (55 %) ont donné face. Supposons maintenant que j'ai compté le nombre de face (que j'appellerai *NF*) que j'ai vues, dans les *N* premiers lancers, et que je calcule à chaque fois la proportion de faces N_F/N. Voici ce que j'obtiendrais (j'ai vraiment fait pile ou face pour produire cela !) :

number of flips	1	2	3	4	5	6	7	8	9	10
number of heads	0	1	2	3	4	4	4	5	6	7
proportion	.00	.50	.67	.75	.80	.67	.57	.63	.67	.70

number of flips	11	12	13	14	15	16	17	18	19	20
number of heads	8	8	9	10	10	10	10	10	10	11
proportion	.73	.67	.69	.71	.67	.63	.59	.56	.53	.55

Notez qu'au début de la séquence, la *proportion* de faces fluctue énormément, à commencer par 0,00 et jusqu'à 0,80. Après, on a l'impression qu'elle diminue un peu, avec de plus en plus de valeurs qui se rapprochent de la « bonne » réponse de 0,50. C'est, en résumé, la définition fréquentiste de la probabilité. Tirez à pile ou face sur une pièce de monnaie et à mesure que *N* devient grand (s'approche de l'infini, noté $N \rightarrow \infty$, la proportion de faces converge vers 50%. Les mathématiciens se soucient de certaines subtilités techniques, mais qualitativement parlant, c'est ainsi que les fréquentistes définissent la probabilité. Malheureusement, je n'ai pas un nombre infini de pièces de monnaie ou la patience infinie requise pour lancer une pièce de monnaie un nombre infini de fois. Cependant, j'ai un ordinateur et les ordinateurs excellent dans les tâches répétitives sans intelligence. J'ai donc demandé à mon ordinateur de simuler 1000 fois le lancer d'une pièce de monnaie, puis j'ai fait un graphique de ce qui arrive à la proportion de N_f/N lorsque *N* augmente. En fait, je l'ai fait quatre fois juste pour m'assurer que ce n'était pas un hasard. Les résultats sont présentés à la Figure 7-1. Comme vous pouvez le constater, la *proportion de faces observées* finit par cesser de fluctuer et se stabiliser. Lorsqu'elle le fait, le nombre auquel elle se stabilise finalement est la véritable probabilité des faces.

La définition fréquentiste de la probabilité présente certaines caractéristiques souhaitables. Premièrement, elle est objective. La probabilité d'un événement est *nécessairement* ancrée dans le monde. La seule façon dont les énoncés de probabilité peuvent avoir un sens, c'est lorsqu'ils font référence à des (une séquence d') événements qui se produisent dans l'univers physique.[40] Deuxièmement, elle est sans ambiguïté. Deux personnes qui regardent

[40] Cela ne veut pas dire que les fréquentistes ne peuvent pas faire des affirmations hypothétiques, bien sûr. C'est simplement que si vous voulez faire une affirmation sur la probabilité, il doit être possible de la redécrire en termes d'une séquence d'événements

la même séquence d'événements se dérouler, en essayant de calculer la probabilité d'un événement, doivent inévitablement trouver la même réponse.

Cependant, elle a aussi des caractéristiques indésirables. Premièrement, les séquences infinies n'existent pas dans le monde physique. Supposons que vous avez pris une pièce dans votre poche et que vous avez commencé à la retourner. Chaque fois qu'elle atterrit, elle a un impact sur le sol. Chaque impact use un peu la pièce. La pièce finira par être détruite. On peut donc se demander s'il est vraiment sensé de prétendre qu'une séquence « infinie » de lancers de pièces de monnaie est même un concept ayant du sens ou objectif. Nous ne pouvons pas dire qu'une « séquence infinie » d'événements est une chose réelle dans l'univers physique, car l'univers physique ne permet rien d'infini. Plus sérieusement, la définition du fréquentiste a une vision étroite. Il y a beaucoup de choses auxquelles les êtres humains sont heureux d'attribuer des probabilités dans le langage courant, mais qui ne peuvent (même en théorie) être mises en correspondance avec une séquence hypothétique d'événements. Par exemple, si un météorologue vient à la télévision et dit « la probabilité de pluie à Adélaïde le 2 novembre 2048 est de 60% », nous sommes heureux de l'accepter. Mais on ne voit pas comment définir cela en termes fréquentistes. Il n'y a qu'une seule ville d'Adélaïde, et un seul le 2 novembre 2048. Il n'y a pas d'enchaînement infini d'événements ici, juste une chose unique. La probabilité fréquentiste nous *interdit* véritablement de faire des déclarations sur la probabilité d'un seul événement. Du point de vue fréquentiste, il pleuvra demain ou il ne pleuvra pas. Il n'y a pas de « probabilité » qui s'attache à un seul événement non répétable. Maintenant, il faut noter qu'il y a des astuces que les fréquentistes peuvent utiliser pour contourner ce problème. Il est possible que le météorologiste veuille dire quelque chose comme « Il y a une catégorie de jours pour lesquels je prévois une probabilité de 60 % de pluie, et si nous ne regardons que les jours pour lesquels je fais cette prédiction, alors, pour 60 % de ces jours, il pleuvra réellement ». C'est très bizarre et contre-intuitif de voir les choses de cette façon, mais on voit parfois des fréquentistes faire cela. Et il en *sera question* plus loin dans ce livre (voir section 8.5).

---

potentiellement observables, avec les fréquences relatives des différents résultats qui apparaissent dans cette séquence.

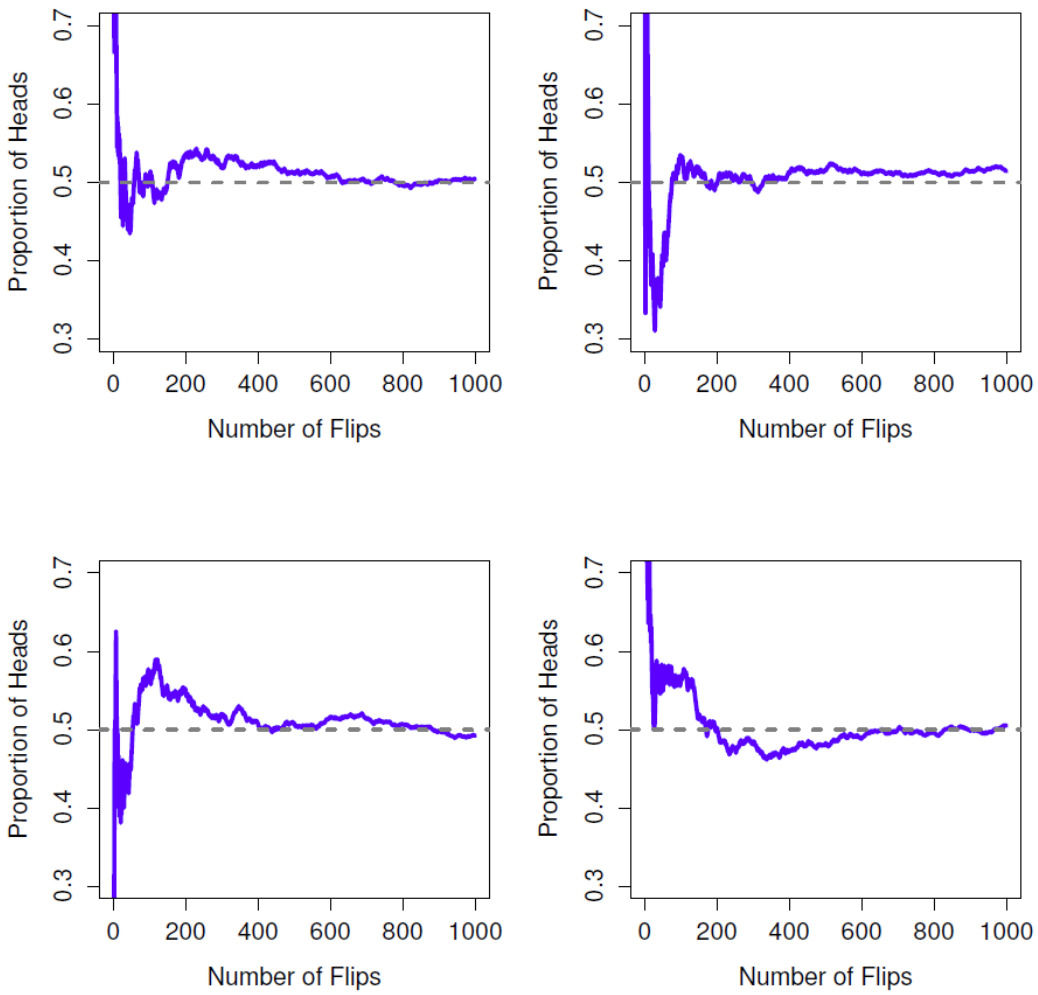


Figure 7-1 : Une illustration du fonctionnement de la probabilité fréquentiste. Si vous tirez à pile ou face sur une pièce de monnaie, la proportion de faces (proportion of heads) que vous avez obtenues finit par diminuer et converge vers la probabilité réelle de 0,5. Chaque graphique présente quatre expériences simulées différentes. Dans chaque cas, nous faisons semblant d'avoir tiré à pile ou face 1000 fois et nous gardons une trace de la proportion de lancers (flips) qui tombait sur face au fur et à mesure que nous avancions. Bien qu'aucune de ces séquences n'ait atteint une valeur exacte de 0,5, si nous avions prolongé l'expérience pour un nombre infini de tours de pièces de monnaie, elles auraient pu le faire.

Le point de vue bayésienne

La **vision bayésienne** de la probabilité est souvent appelée la vision subjectiviste, et bien qu'elle ait été une vision minoritaire chez les statisticiens, elle n'a cessé de gagner du terrain au cours des dernières décennies. Il existe de nombreuses variante du bayésianisme, ce qui rend difficile de dire exactement ce qu'est « la » vision bayésienne. La façon la plus courante de penser la probabilité subjective est de définir la probabilité d'un événement comme le **degré de croyance** qu'un agent intelligent et rationnel attribue à cette vérité de

cet événement. De ce point de vue, les probabilités n'existent pas dans le monde, mais plutôt dans les pensées et les hypothèses des gens et autres êtres intelligents.

Cependant, pour que cette approche fonctionne, nous avons besoin d'un moyen d'opérationnaliser le « degré de croyance ». Une façon de le faire est de le formaliser en termes de « paris rationnels », mais il existe de nombreuses autres façons. Supposons que je croie qu'il y a 60% de chances qu'il pleuve demain. Si quelqu'un me propose un pari selon lequel s'il pleut demain, je gagne 5 $, mais s'il ne pleut pas, je perds 5 $, de toute évidence, selon moi, c'est un assez bon pari. Par contre, si je pense que la probabilité de pluie n'est que de 40%, c'est un mauvais pari à prendre. Ainsi, nous pouvons opérationnaliser la notion de « probabilité subjective » en termes de ce que je suis prêt à accepter de parier.

Quels sont les avantages et les inconvénients de l'approche bayésienne ? L'avantage principal est qu'il vous permet d'assigner des probabilités à n'importe quel événement. Vous n'avez pas besoin d'être limité aux événements qui sont répétables. Le principal inconvénient (pour beaucoup de gens) est que nous ne pouvons pas être purement objectifs. Pour spécifier une probabilité, nous devons spécifier une entité qui a le degré de conviction pertinent. Cette entité peut être un humain, un extra-terrestre, un robot ou même un statisticien. Mais il doit y avoir un agent intelligent qui croit aux choses. Pour beaucoup de gens, c'est inconfortable, cela semble rendre la probabilité arbitraire. Alors que l'approche bayésienne exige que l'agent en question soit rationnel (c'est-à-dire qu'il obéisse aux règles des probabilités), elle permet à chacun d'avoir ses propres croyances. Je peux croire que la pièce est correcte et vous n'avez pas à le faire, même si nous sommes tous les deux rationnels. Le point de vue fréquentiste ne permet pas à deux observateurs d'attribuer des probabilités différentes à un même événement. Quand cela se produit, au moins l'un d'entre eux doit se tromper. Le point de vue bayésien n'empêche pas que cela se produise. Deux observateurs ayant des connaissances de base différentes peuvent légitimement avoir des croyances différentes sur le même événement. Bref, là où la vision fréquentiste est parfois jugée trop étroite (interdit beaucoup de choses auxquelles on veut attribuer des probabilités), la vision bayésienne est parfois jugée trop large (permet trop de différences entre observateurs).

Quelle est la différence ? Et qui a raison ?

Maintenant que vous avez examiné chacune de ces deux perspectives indépendamment, il est utile de vous garantir que vous pouvez comparer les deux. Revenez à l'hypothétique jeu de robots footballeurs au début de la section. Que pensez-vous qu'un fréquentiste et un Bayésien diraient de ces trois énoncés ? Selon un fréquentiste, quelle est la définition correcte de la probabilité ? Lequel choisirait un Bayésien ? Certaines de ces affirmations seraient-elles dénuées de sens pour un fréquentiste ou un Bayésien ? Si vous avez compris les deux points de vue, vous devriez avoir une idée de la façon de répondre à ces questions.

Bien, en supposant que vous compreniez la différence alors vous vous demandez peut-être lequel d'entre eux est *vrai* ? Honnêtement, je ne sais pas s'il y a une bonne réponse. Pour autant que je sache, il n'y a rien de mathématiquement incorrect dans la façon dont les fréquentistes pensent aux séquences d'événements, et il n'y a rien de mathématiquement incorrect dans la façon dont les Bayésiens définissent les croyances d'un agent rationnel. En

fait, lorsque vous creusez dans les détails, les Bayésiens et les fréquentistes sont d'accord sur beaucoup de choses. De nombreuses méthodes fréquentistes conduisent à des décisions pour lesquelles les Bayésiens s'accordent à dire qu'un agent rationnel les prendrait. De nombreuses méthodes bayésiennes ont de très bonnes propriétés fréquentistes.

En général, je suis pragmatique, alors j'utiliserai n'importe quelle méthode statistique en laquelle j'ai confiance. Il s'avère que cela me fait préférer les méthodes bayésiennes pour des raisons que j'expliquerai vers la fin du livre. Mais je ne suis pas fondamentalement opposé aux méthodes fréquentistes. Tout le monde n'est pas aussi détendu. Prenons l'exemple de Sir Ronald Fisher, l'une des figures dominantes de la statistique du XXe siècle et un farouche opposant à tout ce qui est bayésien, dont l'article sur les fondements mathématiques des statistiques qualifiait la probabilité bayésienne de « jungle impénétrable [qui] arrête tout progrès vers la précision des concepts statistiques « (Fisher 1922b, p. 311). Ou le psychologue Paul Meehl, qui suggère que s'appuyer sur des méthodes fréquentistes pourrait faire de vous « un râteau intellectuel puissant mais stérile qui laisse dans son joyeux chemin un long train de jeunes filles violées mais pas de descendance scientifique viable » (Meehl (1967), p. 114). L'histoire des statistiques, comme vous pouvez le constater, n'est pas dénuée de divertissement.

En tout état de cause, si je préfère personnellement la vision bayésienne, la majorité des analyses statistiques sont basées sur l'approche fréquentiste. Mon raisonnement est pragmatique. Le but de ce livre est de couvrir à peu près le même territoire qu'une classe typique de statistiques de premier cycle en psychologie, et si vous voulez comprendre les outils statistiques utilisés par la plupart des psychologues, vous aurez besoin d'une bonne compréhension des méthodes fréquentistes. Je vous promets que ce n'est pas un effort inutile. Même si vous finissez par vouloir passer à la perspective bayésienne, vous devriez vraiment lire au moins un livre sur la vision fréquentiste « orthodoxe ». De plus, je n'ignorerai pas complètement la perspective bayésienne. De temps à autre, j'ajouterai quelques commentaires d'un point de vue bayésien, et je reviendrai sur le sujet plus en détail au chapitre 15.

Théorie de base des probabilités

En dépit des arguments idéologiques entre les Bayésiens et les fréquentistes, il s'avère que la plupart des gens s'entendent sur les règles auxquelles les probabilités doivent obéir. Il y a beaucoup de façons différentes d'en arriver à ces règles. L'approche la plus couramment utilisée est basée sur les travaux d'Andrey Kolmogorov, l'un des grands mathématiciens soviétiques du XXe siècle. Je n'entrerai pas dans les détails, mais je vais essayer de vous donner un petit aperçu de la façon dont cela fonctionne. Et pour ce faire, je vais devoir parler de mon pantalon.

Introduction des distributions de probabilités

Une des vérités troublantes de ma vie est que je ne possède que 5 pantalons. Trois jeans, la moitié inférieure d'un costume et un pantalon en survêtement. Encore plus triste, je leur ai donné des noms : Je les appelle *X1*, *X2*, *X3*, *X4* et *X5*., C'est vraiment pour ça qu'on m'appelle M. Imaginatif. Maintenant, n'importe quel jour, je choisis exactement un pantalon à porter.

Je ne suis pas assez stupide pour essayer de porter deux pantalons, et grâce à des années d'entraînement, je ne sors plus sans pantalon. Si je devais décrire cette situation en utilisant le langage de la théorie des probabilités, je dirais que chaque pantalon (c.-à-d. chaque *X*) est un **événement élémentaire**. La caractéristique clé des événements élémentaires est que chaque fois que nous faisons une observation (par exemple, chaque fois que je mets un pantalon), le résultat sera un et un seul de ces événements. Comme je l'ai dit, de nos jours, je porte toujours un seul pantalon pour que mes pantalons satisfassent à cette contrainte. De même, l'ensemble de tous les événements possibles s'appelle un **l'espace d'échantillonnage**. Certes, certains l'appelleraient une « garde-robe », mais c'est parce qu'ils refusent de penser à mon pantalon en termes probabilistes. C'est triste.

Ok, maintenant que nous avons un espace d'échantillonnage (une garde-robe), qui est construit à partir d'un grand nombre d'événements élémentaires possibles (pantalon), ce que nous voulons faire est d'assigner une **probabilité** à l'un de ces événements élémentaires. Pour un événement *X*, la probabilité de cet événement *P(X)* est un nombre compris entre 0 et 1 ; plus la valeur de *P(X)* est élevée, plus l'événement est probable. Ainsi, par exemple, *P(X)=0* signifie que l'événement *X* est impossible (c'est-à-dire que je ne porte jamais ce pantalon). Par contre, si *P(X)=1* signifie que l'événement *X* est certain (c'est-à-dire que je porte toujours ce pantalon). Pour les valeurs de probabilité au milieu, cela signifie que je porte parfois ce pantalon. Par exemple, si *P(X)=0,5* signifie que je porte ce pantalon la moitié du temps.

A ce stade, on a presque fini. La dernière chose que nous devons reconnaître, c'est que « quelque chose arrive toujours ». Chaque fois que je mets un pantalon, je finis vraiment par porter un pantalon (c'est fou, non ?). Ce que cet énoncé quelque peu banal signifie, en termes probabilistes, c'est que les probabilités des événements élémentaires doivent être égales à 1, ce que l'on appelle la **loi de la probabilité totale**, sans que personne ne s'en soucie vraiment. Plus important encore, si ces exigences sont satisfaites, nous avons une **distribution de probabilités**. Par exemple, ceci est une distribution de probabilités :

Quel pantalon ?	Étiquette	Probabilité
jeans bleu	*X1*	$P(X_1) = 0{,}5$
Jeans gris	*X2*	$P(X_1) = 0{,}3$
Jeans noir	*X3*	$P(X_1) = 0{,}1$
Costume noir	*X4*	$P(X_1) = 0$
Survêtement bleu	*X5*	$P(X_1) = 0{,}1$

Chacun des événements a une probabilité comprise entre 0 et 1, et si l'on additionne la probabilité de tous les événements, ils totalisent 1. Génial. Génial. On peut même dessiner un beau graphique à barres (voir section 5.3) pour visualiser cette distribution, comme le montre la Figure 7-2. Et, à ce stade, nous avons tous accompli quelque chose. Vous avez appris ce qu'est une distribution de probabilité, et j'ai finalement réussi à trouver un moyen

de créer un graphique qui se concentre entièrement sur mes pantalons. Tout le monde y gagne !

La seule autre chose que je dois souligner, c'est que la théorie des probabilités vous permet de parler d'**événements non élémentaires** aussi bien que d'événements élémentaires. La façon la plus simple d'illustrer le concept est d'utiliser un exemple. Dans l'exemple du pantalon, il est parfaitement légitime de parler de la probabilité que je porte un jean. Dans ce scénario, l'événement « Dan porte un jean » est censé s'être produit aussi longtemps que l'événement élémentaire qui s'est réellement produit est l'un des événements appropriés. Dans ce cas, « jeans bleu », « jeans noir » ou « jeans gris ». En termes mathématiques, nous avons défini l'événement E « jeans » pour correspondre à l'ensemble des événements élémentaires (X_1, X_2, X_3). Si l'un de ces événements élémentaires se produit alors E est également dit s'être produit. Après avoir décidé d'écrire la définition du E *de* cette façon, il est assez simple d'indiquer quelle est la probabilité P(E): on additionne tout. Dans ce cas particulier

$$P(E) = P(X_1) + P(X_2) + P(X_3)$$

et, puisque les probabilités de jeans bleu, gris et noir sont respectivement 0,5 ; 0,3 et 0,1, la probabilité que je porte un jeans est égale à 0,9.

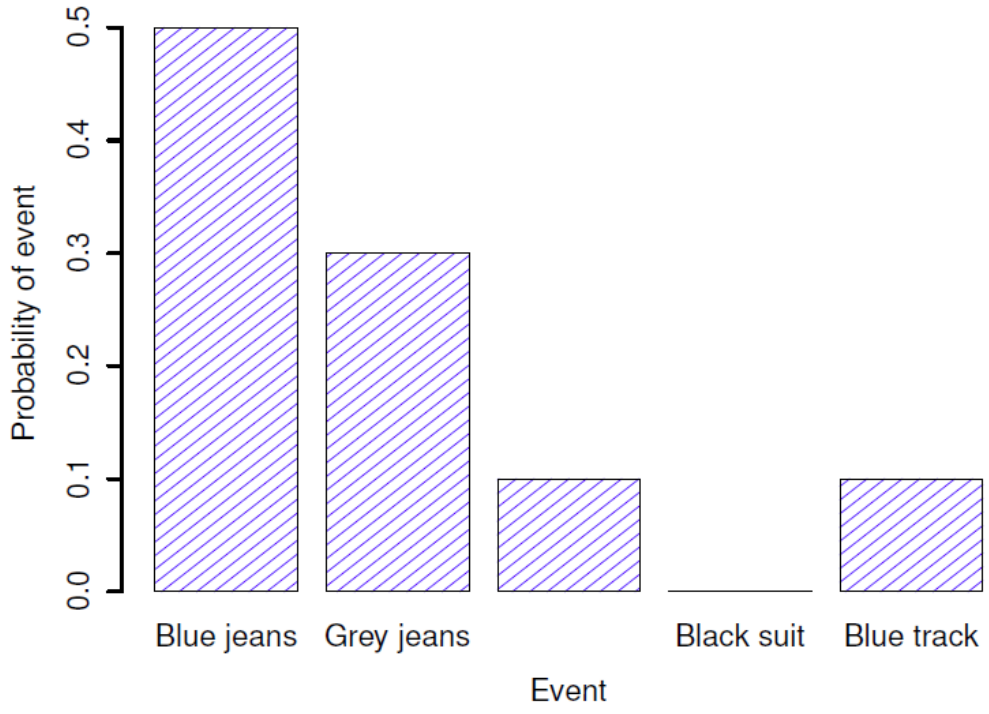


Figure 7-2 : Représentation visuelle de la distribution de probabilité « pantalon ». Il y a cinq « événements élémentaires », correspondant aux cinq pantalons que je possède. Chaque événement a une certaine probabilité de se produire : cette probabilité est un nombre compris entre 0 et 1, la somme de ces probabilités étant égale à 1.

À ce stade, vous pensez peut-être que tout cela est terriblement évident et simple et vous auriez raison. Tout ce que nous avons vraiment fait, c'est d'envelopper quelques notions mathématiques de base autour de quelques intuitions de bon sens. Cependant, à partir de ces débuts simples, il est possible de construire des outils mathématiques extrêmement puissants. Je ne vais certainement pas entrer dans les détails de ce livre, mais ce que je vais faire, c'est énumérer, dans le Tableau 7-1, certaines des autres règles que les probabilités satisfont. Ces règles peuvent être dérivées des hypothèses simples que j'ai décrites ci-dessus, mais comme nous n'utilisons pas ces règles pour quoi que ce soit dans ce livre, je ne le ferai pas ici.

Tableau 7-1 : Quelques règles de base que les probabilités doivent satisfaire. Vous n'avez pas vraiment besoin de connaître ces règles pour comprendre les analyses dont nous parlerons plus loin dans le livre, mais elles sont importantes si vous voulez comprendre un peu plus profondément la théorie des probabilités.

En français	Notation	Formule
Non A	$P(\neg A)$	$= 1 - P(A)$
A ou B	$P(A \cup B)$	$= P(A) + P(B) - P(A \cap B)$
A et B	$P(A \cap B)$	$= P(A\|B)P(B)$

La distribution binomiale

Comme vous pouvez l'imaginer, les distributions de probabilités varient énormément et il existe une large gamme de distributions. Cependant, elles ne sont pas toutes d'égale importance. En fait, la grande majorité du contenu de ce livre repose sur l'une des cinq distributions suivantes : la distribution binomiale, la distribution normale, la distribution *t*, la distribution χ^2 (chi carré) et la distribution *F.* C'est pourquoi, au cours des prochaines sections, je vais vous présenter brièvement ces cinq distributions, en accordant une attention particulière à la binomiale et à la normale. Je vais commencer par la distribution binomiale puisque c'est la plus simple des cinq.

Présentation de la distribution binomiale

La théorie des probabilités est née de la tentative de décrire le fonctionnement des jeux de hasard ; il semble donc approprié que notre discussion sur la **distribution binomiale** comprenne une discussion sur les lancer de dés et de pièces à pile ou face. Imaginons une simple « expérience ». Dans ma petite main chanceuse, je tiens 20 dés identiques à six faces. Sur une face de chaque dé, il y a l'image d'un crâne, les cinq autres faces sont toutes vides. Si je lance les 20 dés, quelle est la probabilité que j'obtienne exactement 4 crânes ? En supposant que les dés sont corrects, nous savons que la probabilité qu'une personne meure en tirant le crâne est de 1 sur 6. En d'autres termes, la probabilité d'avoir un crâne pour un seul dé est d'environ *0,167*. C'est assez d'information pour répondre à notre question, regardons comment cela se fait.

Comme d'habitude, nous allons vous présenter quelques noms et quelques notations. Nous laisserons N indiquer le nombre de jets de dés dans notre expérience, qui est souvent appelé **paramètre de taille** de notre distribution binomiale. Nous utiliserons aussi θ pour faire référence à la probabilité qu'un seul dé présente un crâne, une quantité que l'on appelle habituellement la **probabilité de succès** de la binomiale.[41] Enfin, nous utiliserons X pour faire référence aux résultats de notre expérience, à savoir le nombre de crânes que j'obtiens lorsque je lance les dés.

Tableau 7-2 : Formules pour les distributions binomiale et normale. Nous n'utilisons pas vraiment ces formules pour quoi que ce soit dans ce livre, mais elles sont assez importantes pour un travail plus avancé, alors j'ai pensé qu'il serait peut-être préférable de les mettre ici dans un tableau, où elles ne peuvent pas gêner le texte. Dans l'équation de la distribution binomiale, X ! est la factorielle (c.-à-d. multiplier tous les nombres entiers de 1 à X), et pour la distribution normale, « exp » désigne la fonction exponentielle, dont nous avons parlé au chapitre 6. Si ces équations n'ont pas beaucoup de sens pour vous, ne vous en faites pas trop.

Binomiale	$P(X\|\theta, N) = \frac{N!}{X!\,(N-X)!}\theta^X(1-\theta)^{N-X}$
Normale	$p(X\|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$

Puisque la valeur réelle de X est due au hasard, nous l'appelons **variable aléatoire**. Quoi qu'il en soit, maintenant que nous avons toute cette terminologie et cette notation, nous pouvons l'utiliser pour énoncer le problème un peu plus précisément. La quantité que l'on veut calculer est la probabilité que $X = 4$ étant donné que l'on sait que $\theta = 0{,}167$ et N=20. La « forme » générale de la chose que je veux calculer pourrait s'écrire comme suit :

$$P(X|\theta, N)$$

et nous nous intéressons au cas particulier où $X = 4$, $\theta = 0{,}167$ et $N = 20$. Il ne me reste plus qu'un seul élément de notation à mentionner avant de passer à la discussion sur la solution au problème. Si je veux dire que X est généré aléatoirement à partir d'une distribution binomiale avec les paramètres θ et N, la notation que je devrais utiliser est la suivante :

$$X \sim Binomiale(\theta, N)$$

Ouais, ouais !. Je sais ce que vous pensez : notation, notation, notation, notation. Vraiment, qui s'y intéresse ? Très peu de lecteurs de ce livre sont ici pour la notation, donc je devrais probablement passer à autre chose et parler de la façon d'utiliser la distribution binomiale.

---

[41] Notez que le terme "succès" est plutôt arbitraire et n'implique pas que le résultat est quelque chose à désirer. Si θ faisait référence à la probabilité qu'un passager se blesse dans un accident d'autobus, j'appellerais cela la probabilité de succès, mais cela ne veut pas dire que je veux que les gens soient blessés dans un accident d'autobus !

J'ai inclus la formule pour la distribution binomiale dans le Tableau 7-2, puisque certains lecteurs voudront peut-être jouer avec, mais comme la plupart des gens ne s'en soucient probablement pas beaucoup et parce que nous n'avons pas besoin de la formule dans ce livre, je ne vais pas en parler en détail. Au lieu de cela, je veux juste vous montrer à quoi ressemble la distribution binomiale.

Pour cela, la Figure 7-3 présente les probabilités binomiales pour toutes les valeurs possibles de X pour notre expérience de lancement de dés, de $X = 0$ (sans crânes) jusqu'à $X = 20$ (tous les crânes). Notez qu'il s'agit essentiellement d'un diagramme à barres qui n'est pas différent de celui que j'ai dessiné à la Figure 7-2. Sur l'axe horizontal, nous avons tous les événements possibles, et sur l'axe vertical, nous pouvons lire la probabilité de chacun de ces événements. Ainsi, la probabilité d'obtenir 4 crânes sur 20 lancers est d'environ 0,20 (la réponse réelle est 0,2022036, comme nous allons le voir dans un instant). En d'autres termes, on s'attendrait à ce que cela se produise environ 20 % du temps où vous avez répété cette expérience.

Pour vous donner une idée de la façon dont la distribution binomiale change lorsque nous modifions les valeurs de θ et N, supposons qu'au lieu de lancer des dés, je suis en train de lancer des pièces. Cette fois-ci, mon expérience consiste à lancer une pièce de monnaie à plusieurs reprises et le résultat qui m'intéresse est le nombre de faces que j'observe. Dans ce scénario, la probabilité de succès est maintenant $\theta = 1/2$.

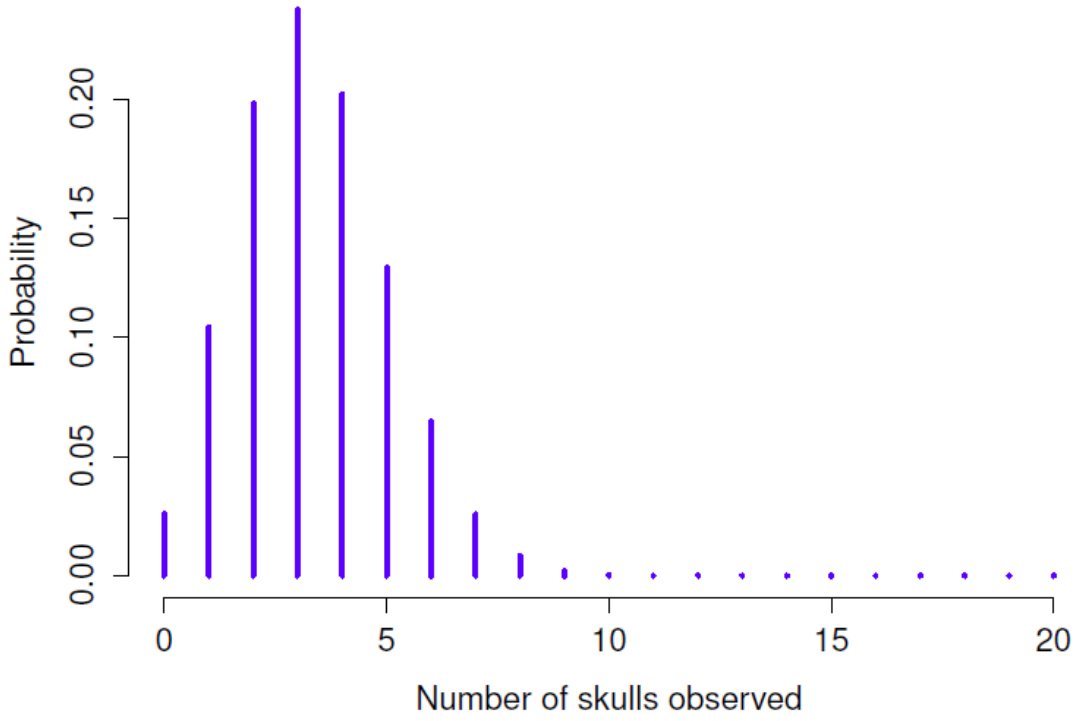


Figure 7-3 : La distribution binomiale avec un paramètre de taille de N = 20 et une probabilité de succès sous-jacente de $\theta = 1/6$ Chaque barre verticale représente la probabilité d'un résultat spécifique (c.-à-d. une valeur possible de X). Comme il s'agit d'une

distribution de probabilités, chacune des probabilités doit être un nombre compris entre 0 et 1, et la somme des hauteurs des barres doit être égale à 1.

Supposons que je devais lancer la pièce $N = 20$ fois. Dans cet exemple, j'ai changé la probabilité de succès mais j'ai gardé la même taille d'expérience. Qu'est-ce que cela change à notre distribution binomiale ? Eh bien, comme le montre la Figure 7-4a, l'effet principal est de déplacer toute la distribution, comme on peut s'y attendre. Bien, et si on jouait à pile ou face 100 fois ? Eh bien, dans ce cas, nous obtenons la Figure 7-4b. La distribution reste à peu près au milieu, mais il y a un peu plus de variabilité dans les résultats possibles.

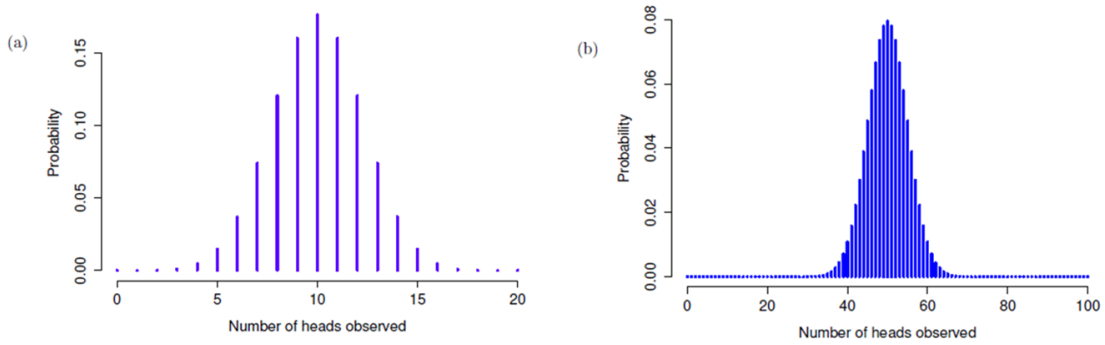


Figure 7-4 : Deux distributions binomiales, impliquant un scénario dans lequel je tire à pile ou face, donc la probabilité de succès sous-jacente est $\theta = 1/2$. Dans le graphique (a), nous supposons que je lance la pièce N = 20 fois. Dans le panneau (b) nous supposons que la pièce est lancée N = 100 fois.

La distribution normale

Bien que la distribution binomiale soit conceptuellement la distribution la plus simple à comprendre, ce n'est pas la plus importante. Cet honneur particulier revient à la **distribution normale**, également appelée « courbe en cloche » ou « distribution gaussienne ». Une distribution normale est décrite à l'aide de deux paramètres : la moyenne de la distribution μ et l'écart-type de la distribution σ.

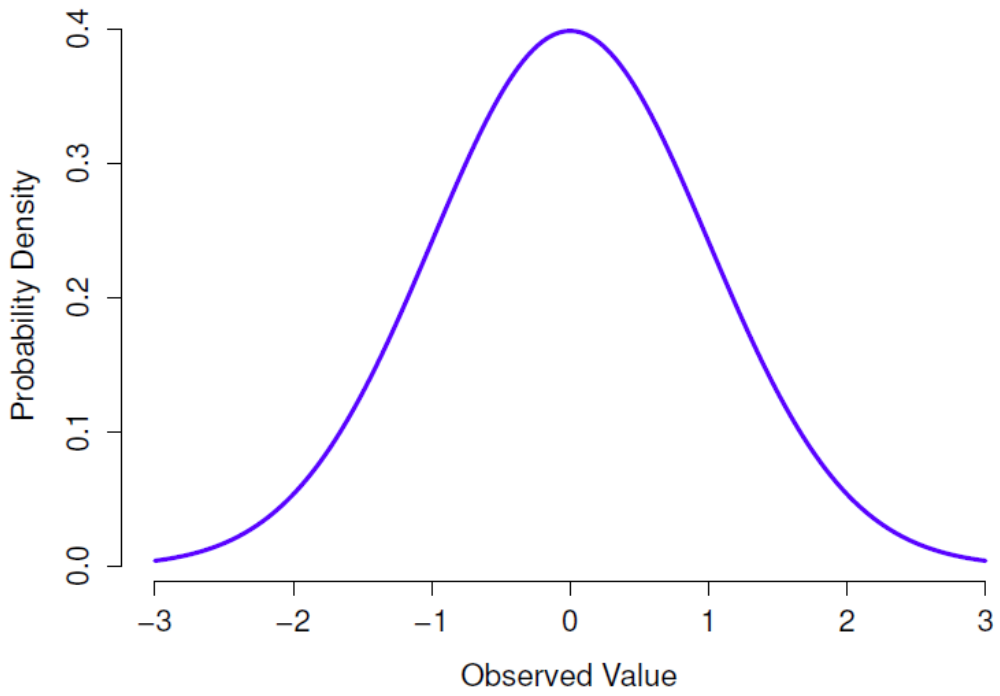


Figure 7-5 : Distribution normale avec moyenne μ = 0 et écart-type$\sigma = 1$. L'axe des x correspond à la valeur d'une variable, et l'axe des y nous indique la probabilité d'observer cette valeur. Notez cependant que l'axe des y est appelé « densité de probabilités » et non « Probability ». Il y a une particularité subtile et quelque peu frustrante des distributions continues qui fait que l'axe des y se comporte un peu bizarrement : la hauteur de la courbe ici n'est pas vraiment la probabilité d'observer une valeur x particulière. D'autre part, il est vrai que les hauteurs de la courbe vous indiquent quelles valeurs x sont les plus probables (les plus élevées !) (voir section 7.5.1 pour tous les détails agaçants).

La notation que nous utilisons parfois pour dire qu'une variable *X* est normalement distribuée est la suivante :

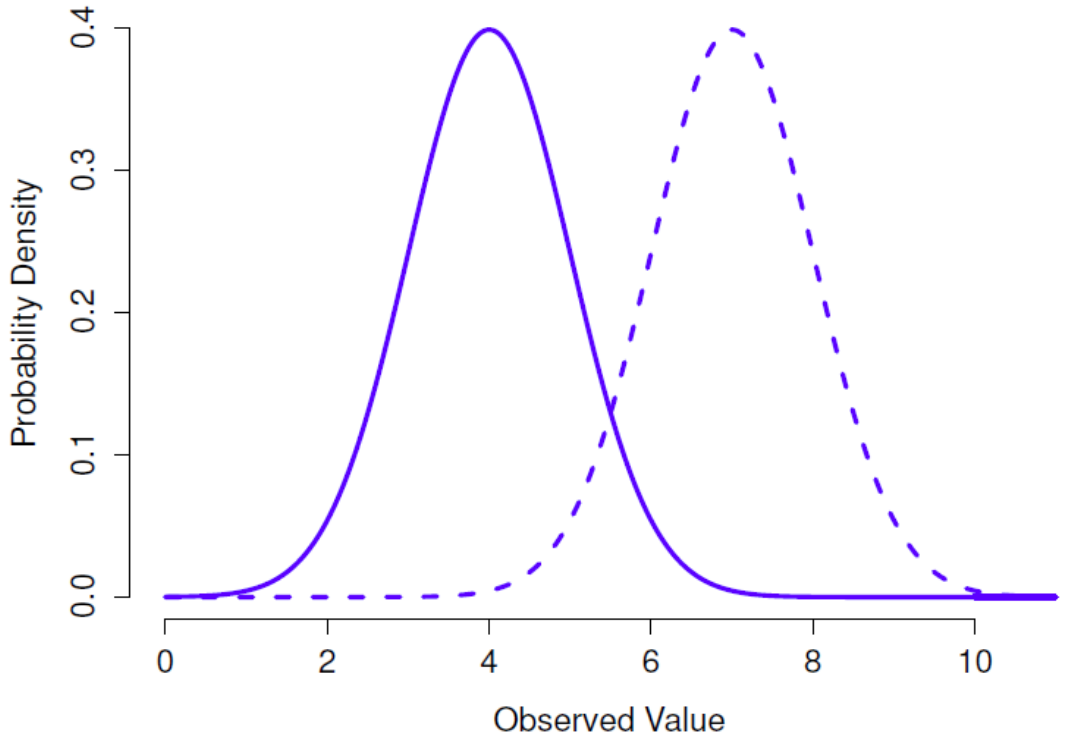
$$X \sim Normal(\mu, \sigma)$$

Bien sûr, c'est juste de la notation. Cela ne nous dit rien d'intéressant sur la distribution normale elle-même. Comme dans le cas de la distribution binomiale, j'ai inclus la formule de la distribution normale dans ce livre, parce que je pense qu'il est assez important que tous ceux qui apprennent les statistiques y jettent un coup d'œil, mais comme il s'agit d'un texte d'introduction, je ne veux pas m'y attarder, alors je l'ai mis de côté au Tableau 7-2.

Au lieu de se concentrer sur les mathématiques, essayons de comprendre ce que signifie le fait qu'une variable soit normalement distribuée. Pour ce faire, jetez un coup d'œil à la Figure 7-5 qui présente une distribution normale avec une moyenne μ = 0 et un écart-type $\sigma = 1$. Vous pouvez voir d'où vient le nom « courbe en cloche » ; elle ressemble un peu à une cloche. Remarquez que, contrairement aux graphiques que j'ai dessinés pour illustrer la distribution binomiale, l'image de la distribution normale de la Figure 7-5 montre une distribution lissée au lieu des barres d'un histogramme. Ce n'est pas un choix arbitraire, la

distribution normale est continue alors que la distribution binomiale est discrète. Par exemple, dans l'exemple du jet de dé de la dernière section, il était possible d'obtenir 3 ou 4 crânes, mais impossible d'obtenir 3,9 crânes. Les chiffres que j'ai mentionnés dans la section précédente reflètent ce fait. Dans la Figure 7-3, par exemple, il y a une barre située à *X = 3* et une autre à ** = *4 mais il n'y a rien entre les deux. Les quantités continues n'ont pas cette contrainte. Supposons, par exemple, qu'il s'agisse du temps qu'il fait. La température par une agréable journée de printemps peut être de 23 degrés, 24 degrés, 23,9 degrés, ou n'importe quoi entre les deux, puisque la température est une variable continue. Par conséquent, une distribution normale pourrait être tout à fait appropriée pour décrire les températures printanières.[42]

En gardant cela à l'esprit, voyons si nous ne pouvons pas avoir une intuition sur le fonctionnement de la distribution normale. Voyons d'abord ce qui se passe quand on joue avec les paramètres de la distribution. Pour cela, la Figure 7-6 présente les distributions normales qui ont des moyennes différentes mais ont le même écart-type.



[42] En pratique, la distribution normale est si pratique que les gens ont tendance à l'utiliser même lorsque la variable n'est pas réellement continue. Tant qu'il y a suffisamment de catégories (p. ex. réponses à un questionnaire selon l'échelle de Likert), il est assez courant d'utiliser la distribution normale comme approximation. Cela fonctionne beaucoup mieux en pratique que vous ne le pensez.

Figure 7-6 : Une illustration de ce qui se passe lorsque vous modifiez la moyenne d'une distribution normale. Dans les deux cas, l'écart-type est $\sigma = 1$. Comme on pouvait s'y attendre, les deux distributions ont la même forme, mais la ligne en pointillés est décalée vers la droite.

Comme on peut s'y attendre, toutes ces distributions ont la même « largeur ». La seule différence entre eux est qu'ils ont été déplacés vers la gauche ou vers la droite. Sur tous les autres points, ils sont identiques. Par contre, si nous augmentons l'écart-type tout en maintenant la moyenne constante, le pic de la distribution reste au même endroit mais la distribution s'élargit, comme vous pouvez le voir à la Figure 7-7. Notez, cependant, que lorsque nous élargissons la distribution, la hauteur du pic diminue.

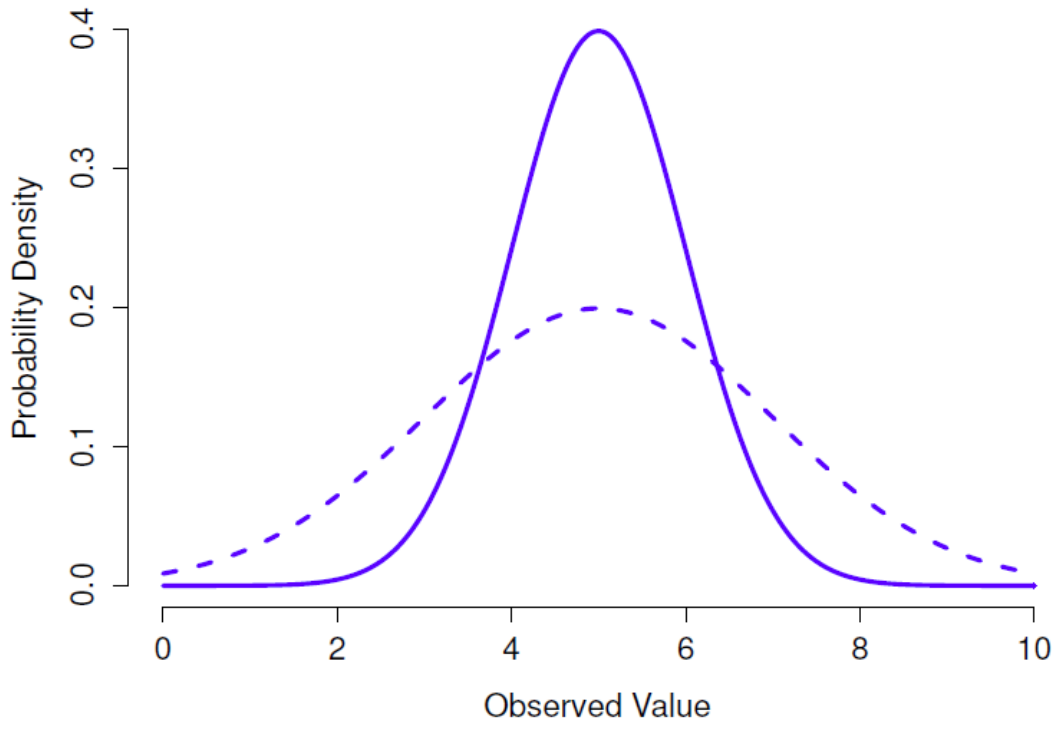


Figure 7-7 : Une illustration de ce qui se passe lorsque vous modifiez l'écart-type d'une distribution normale. Les deux distributions représentées dans cette figure ont une moyenne de μ = 5, mais elles ont des écarts-types différents. La ligne pleine correspond à une distribution avec un écart-type $\sigma = 1$, et la ligne pointillée montre une distribution avec un écart-type $\sigma = 2$. Par conséquent, les deux distributions sont « centrées » au même endroit, mais la ligne pointillée est plus large que la solide.

Ceci doit se produire, de la même manière que les hauteurs des barres que nous avons utilisées pour dessiner une distribution binomiale discrète doivent *totaliser* 1, l'*aire* totale *sous la courbe* pour la distribution normale doit être égale à 1. Avant de poursuivre, j'aimerais souligner une caractéristique importante de la distribution normale. Indépendamment de la moyenne réelle et de l'écart-type, 68,3 % de la superficie se situe à moins d'un écart-type de la moyenne. De même, 95,4 % de la distribution se situe à

l'intérieur de plus ou moins deux écarts-types de la moyenne et 99,7 % de la distribution se situe à l'intérieur de plus ou moins trois écarts-types. Cette idée est illustrée à la Figure 7-8.

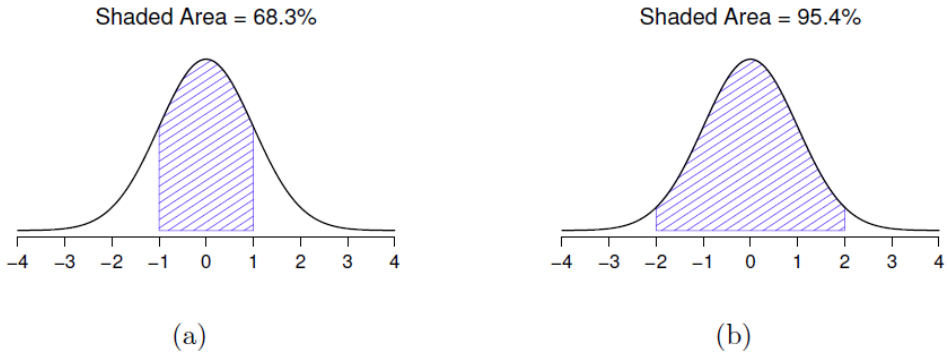


Figure 7-8 : L'aire sous la courbe indique la probabilité qu'une observation se situe dans une plage particulière. Les lignes pleines représentent les distributions normales avec une moyenne μ = 0 et un écart-type $\sigma = 1$. Les zones ombrées illustrent les « zones sous la courbe » pour deux cas importants. Dans le panel a, nous pouvons voir qu'il y a 68,3 % de chances qu'une observation se situe dans un écart-type de la moyenne. Dans le panel b, nous voyons qu'il y a 95,4 % de chances qu'une observation se situe dans les deux écarts types de la moyenne.

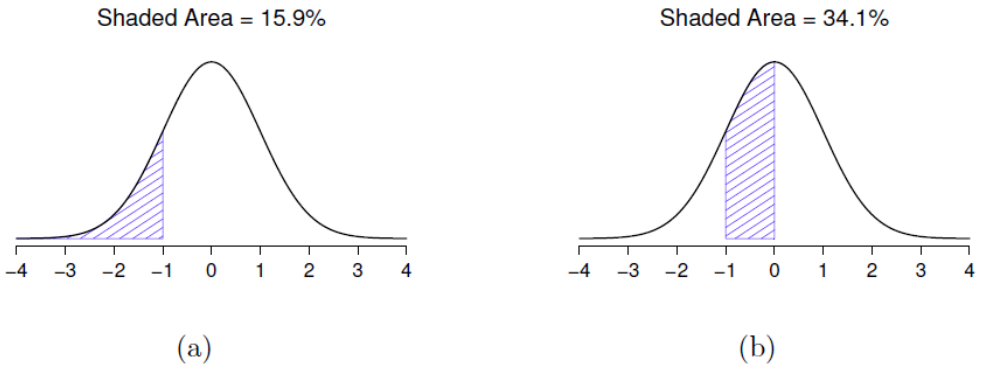


Figure 7-9 : Deux autres exemples de « l'aire sous l'idée de courbe ». Il y a 15,9 % de chances qu'une observation se situe un écart-type inférieur ou supérieure à la moyenne (panel a), et 34,1 % de chances que l'observation se situe quelque part entre un écart-type inférieur à la moyenne et la moyenne (panel b). Notez que si vous additionnez ces deux chiffres, vous obtenez 15,9% + 34,1% = 50%. Pour les données normalement distribuées, il y a 50 % de chances qu'une observation soit inférieure à la moyenne. Et bien sûr, cela implique aussi qu'il y a 50 % de chances qu'elle soit supérieure à la moyenne.

Densité de probabilité

Il y a quelque chose que j'ai essayé de cacher tout au long de ma discussion sur la distribution normale, quelque chose que certains manuels d'introduction omettent complètement. Ils ont peut-être raison de le faire. Cette "chose" que je cache est bizarre et

contre-intuitive, même si l'on s'en tient aux normes déformantes qui s'appliquent aux statistiques. Heureusement, ce n'est pas quelque chose que vous devez comprendre de manière approfondie pour faire des statistiques de base. C'est plutôt quelque chose qui commence à devenir important plus tard, quand vous aurez dépassé le stade de la base. Donc, si ça n'a pas de sens, ne vous inquiétez pas trop, mais assurez-vous d'en suivre l'essentiel.

Tout au long de ma discussion sur la distribution normale, il y a eu une ou deux choses qui ne se tiennent pas très bien. Vous avez peut-être remarqué que l'axe des y de ces figures est marqué « Densité de probabilité » plutôt que densité. Peut-être avez-vous remarqué que j'ai utilisé p(X) au lieu de P(X) lorsque j'ai donné la formule pour la distribution normale.

En fait, ce qui est présenté ici n'est pas vraiment une probabilité, c'est autre chose. Pour comprendre ce qu'est ce quelque chose, il faut passer un peu de temps à réfléchir à ce que *signifie* vraiment le fait de dire que *X* est une variable continue. Disons que nous parlons de la température extérieure. Le thermomètre me dit qu'il fait 23 degrés, mais je sais que ce n'est pas vraiment vrai. Il ne fait pas *exactement* 23 degrés. Il fait peut-être 23,1 degrés. Mais je sais que ce n'est pas vraiment vrai non plus parce qu'il pourrait faire 23,09 degrés. Mais je sais que… eh bien, vous avez compris l'idée. Ce qui est délicat avec les quantités réellement continues, c'est qu'on ne sait jamais vraiment ce qu'elles sont exactement.

Pensez maintenant à ce que cela implique lorsque nous parlons de probabilités. Supposons que la température maximale de demain soit échantillonnée à partir d'une distribution normale avec une moyenne de 23 et un écart-type 1. Quelle est la probabilité que la température atteigne *exactement* 23 degrés ? La réponse est « zéro », ou peut-être « un nombre si proche de zéro qu'il pourrait aussi bien être zéro ». Pourquoi en est-il ainsi ? C'est comme essayer de lancer une fléchette sur une cible infiniment petite. Peu importe à quel point vous visez bien, vous ne le toucherez jamais. Dans la vraie vie, vous n'obtiendrez jamais une valeur de 23 exactement. Ce sera toujours quelque chose comme 23,1 ou 22,99998 ou autre. En d'autres termes, il est tout à fait inutile de parler de la probabilité que la température soit exactement de 23 degrés. Cependant, dans le langage courant, si je vous disais qu'il faisait 23 degrés dehors et qu'il faisait 22,9998 degrés, vous ne me traiteriez probablement pas de menteur. Car dans le langage courant, « 23 degrés » veut dire en général quelque chose comme « entre 22,5 et 23,5 degrés ». Et bien qu'il ne semble pas avoir beaucoup de sens de s'interroger sur la probabilité que la température soit exactement de 23 degrés, il semble raisonnable de s'interroger sur la probabilité que la température se situe entre 22,5 et 23,5, ou entre 20 et 30, ou toute autre plage de températures.

Le but de cette discussion est de préciser que lorsqu'on parle de distributions continues, il n'est pas utile de parler de la probabilité d'une valeur précise. Cependant, ce dont nous *pouvons* parler, c'est de la probabilité que la valeur se situe à l'intérieur d'une fourchette particulière de valeurs. Pour connaître la probabilité associée à une plage particulière, il suffit de calculer « l'aire sous la courbe ». Nous avons déjà vu ce concept, dans la Figure 7-8, les zones ombragées illustrent les probabilités réelles (p. ex. dans la Figure 7-8a, elle montre la probabilité d'observer une valeur qui se situe à moins de 1 écart type de la moyenne).

Bien, donc cela explique une partie de l'histoire. J'ai expliqué un peu comment les distributions de probabilités continues devraient être interprétées (c'est-à-dire que l'aire sous la courbe est l'élément clé). Mais que signifie la formule pour p(x) que j'ai décrite plus tôt ? Évidemment, p(x) ne décrit pas une probabilité, mais qu'est-ce que c'est ? Le nom de cette quantité p(x) est une **densité de probabilité**, et dans les graphiques que nous avons dessinés, elle correspond à la *hauteur de la* courbe. Les densités elles-mêmes ne sont pas significatives en soi, mais elles sont « arrangées » pour s'assurer que l'*aire* sous la courbe puissent toujours être interprétée comme de véritables probabilités. Pour être honnête, c'est à peu près tout ce que vous avez vraiment besoin de savoir pour l'instant.[43]

Autres distributions utiles

La distribution normale est la distribution la plus utilisée par les statistiques (pour des raisons qui seront discutées sous peu), et la distribution binomiale est très utile à de nombreuses fins. Mais le monde des statistiques est rempli de distributions de probabilités, dont certaines que nous rencontrerons en passant. En particulier, les trois qui apparaîtront dans ce livre sont la distribution *t*, la distribution χ^2 et la distribution *F.* Je ne donnerai pas de formules pour aucune d'entre elles, ni n'en parlerai trop en détail, mais je vais vous montrer quelques images.

---

[43] Pour les lecteurs qui connaissent un peu le calcul, je vais donner une explication un peu plus précise. De la même manière que les probabilités sont des nombres non négatifs qui doivent s'additionner à 1, les densités de probabilité sont des nombres non négatifs qui doivent s'intégrer à 1 (où l'intégrale est prise sur toutes les valeurs possibles de *X*). Pour calculer la probabilité que *X* se situe entre *a* et *b*, nous calculons l'intégrale définie de la fonction de densité sur la plage correspondante, $\int_a^b p(x)\text{dx}$.Si vous ne vous souvenez pas ou si vous n'avez jamais appris le calcul, ne vous en faites pas. Ce n'est pas nécessaire pour ce livre.

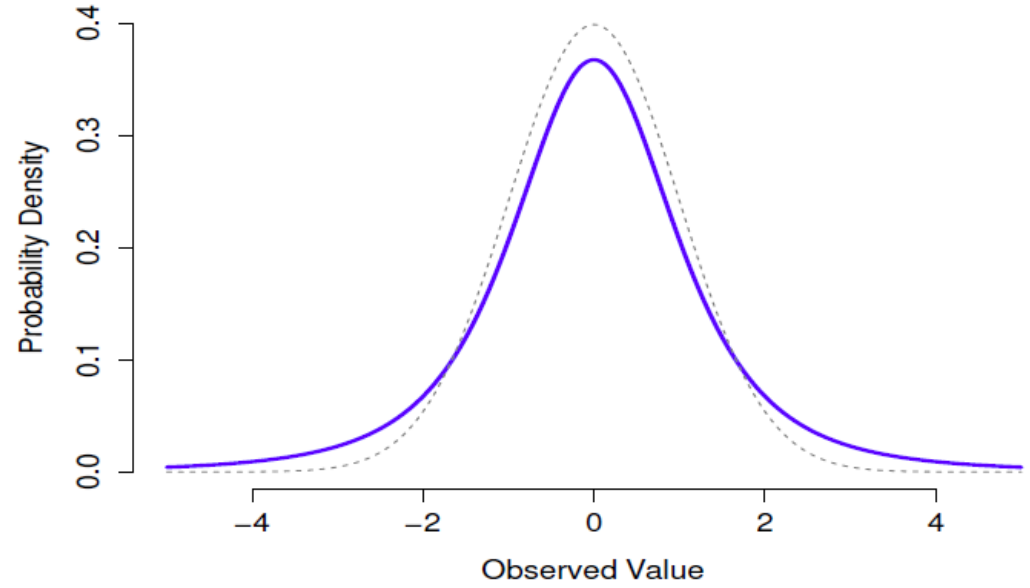


Figure 7-10 : Distribution t avec 3 degrés de liberté (ligne pleine). Cela ressemble à une distribution normale, mais ce n'est pas tout à fait la même chose. À des fins de comparaison, j'ai tracé une distribution normale standard sous la forme d'une ligne pointillée.

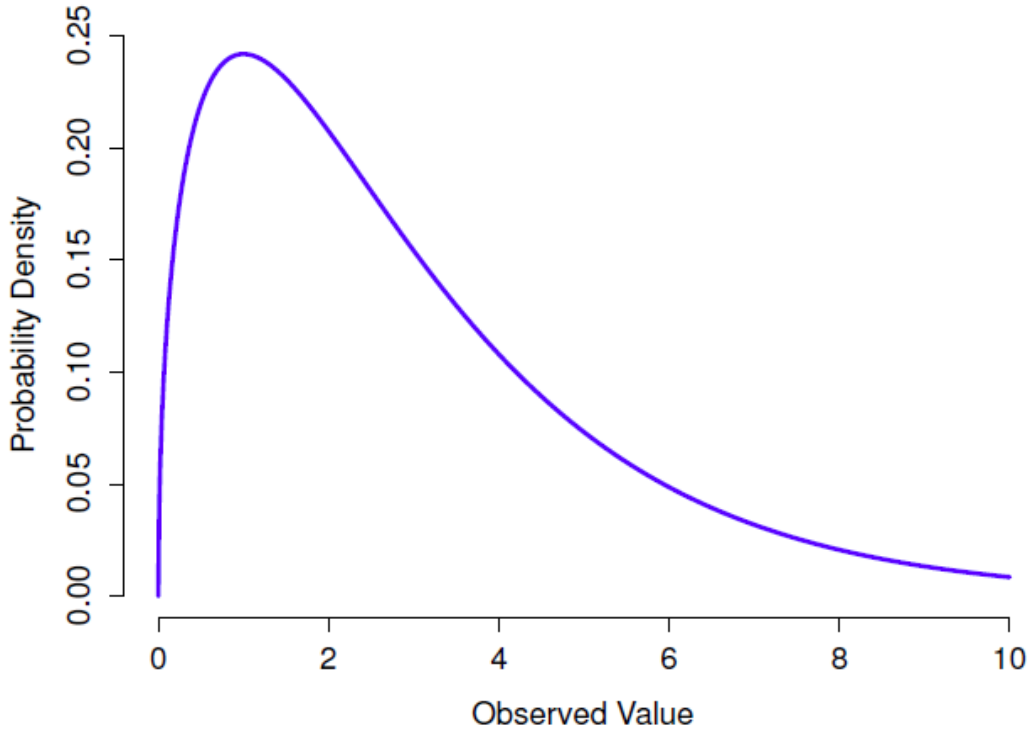


Figure 7-11 : Une distribution χ^2 avec 3 degrés de liberté. Notez que les valeurs observées doivent toujours être supérieures à zéro et que la distribution est assez asymétrique. Ce sont les principales caractéristiques d'une distribution du chi carré

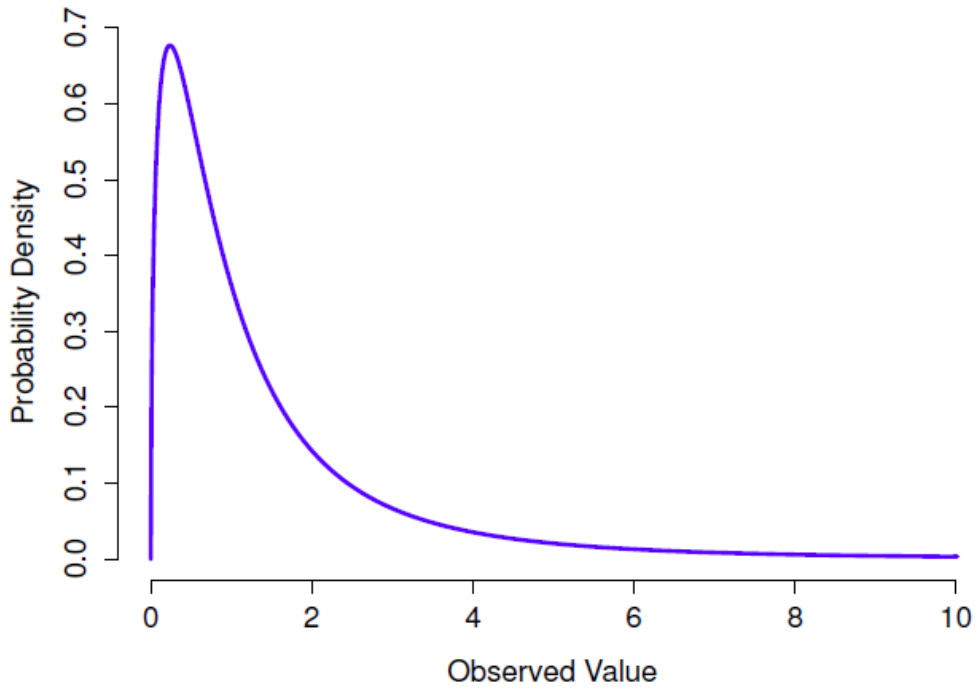


Figure 7-12 : Une distribution F avec 3 et 5 degrés de liberté. Qualitativement parlant, cela ressemble beaucoup à une distribution du chi carré, mais ce n'est pas tout à fait la même chose

- La **distribution** *t* est une distribution continue qui ressemble beaucoup à une distribution normale, voir Figure 7-10. Notez que les « queues » de la distribution *t* sont « plus lourdes » (c'est-à-dire qu'elles s'étendent plus vers l'extérieur) que les queues de la distribution normale). C'est la différence importante entre les deux. Cette distribution tend à se produire dans les situations où vous pensez que les données suivent une distribution normale, mais que vous ne connaissez pas la moyenne ou l'écart-type. Nous reviendrons sur cette distribution au chapitre 11.
- La **distribution** χ^2 est une autre distribution qui apparaît dans beaucoup d'endroits différents. La situation dans laquelle nous le verrons est celle de l'analyse des données catégoriques (chapitre 10), mais c'est l'une de ces choses qu'on rencontre un peu partout en fait. Quand vous creusez dans les mathématiques (et qui n'aime pas faire cela ?), il s'avère que la principale raison pour laquelle la distribution χ^2 apparaît partout est que si vous avez plusieurs variables qui sont normalement distribuées, que vous calculez le carré leurs valeurs et puis les additionner (une procédure appelée « faire la somme des carrés »), cette somme a une distribution χ^2. Vous seriez étonné de voir combien de fois ce fait s'avère utile. Quoi qu'il en soit, voici à quoi ressemble une distribution de χ^2 : Figure 7-11.

- La **distribution** *F* ressemble un peu à une distribution χ^2, et elle apparaît chaque fois que vous avez besoin de comparer deux distributions χ^2 entre elles. Certes, cela ne semble pas exactement quelque chose que toute personne saine d'esprit voudrait faire, mais cela s'avère très important dans l'analyse des données du monde réel. Rappelez-vous quand j'ai dit que χ^2 s'avère être la distribution clé quand on prend une « somme de carrés » ? Eh bien, ce que cela signifie, c'est que si vous voulez comparer deux « sommes de carrés » différents, vous parlez probablement de quelque chose qui a une distribution *F.* Bien sûr, je ne vous ai pas encore donné d'exemple de quelque chose qui implique une somme de carrés, mais je le ferai au chapitre 13. Et c'est là qu'on tombera sur la distribution *F.* Oh, et il y a une image à la Figure 7-12.

Bien, il est temps de terminer cette section. Nous avons vu trois nouvelles distributions : χ^2, *t* et *F*. Ce sont toutes des distributions continues, et elles sont toutes étroitement liées à la distribution normale. L'essentiel pour nous, c'est que vous saisissiez l'idée de base que ces distributions sont toutes profondément liées les unes aux autres, et à la distribution normale. Plus loin dans ce livre, nous allons rencontrer des données qui sont normalement distribuées, ou du moins supposées l'être. Ce que je veux que vous compreniez maintenant, c'est que, si vous supposez que vos données sont normalement distribuées, vous ne devriez pas être surpris de voir les distributions χ^2, *t* et *F* apparaître partout quand vous commencez à essayer de faire votre analyse de données.

Résumé

Dans ce chapitre, nous avons parlé de probabilité. Nous avons parlé de ce que la probabilité signifie et pourquoi les statisticiens ne s'entendent pas sur ce qu'elle signifie. Nous avons parlé des règles auxquelles les probabilités doivent obéir. Et nous avons introduit l'idée d'une distribution de probabilités et passé une bonne partie du chapitre à parler de certaines des distributions de probabilités les plus importantes avec lesquelles les statisticiens travaillent. La ventilation section par section ressemble à ceci :

- Théorie des probabilités et statistiques (section 7.1)
- Opinions fréquentistes et bayésiennes sur la probabilité (section 7.2)
- Notions de base de la théorie des probabilités (section 7.3)
- Distribution binomiale (section 7.4), distribution normale (section 7.5) et autres (section 7.6)

Comme vous pouvez vous y attendre, ce panorama n'est en aucun cas exhaustif. La théorie des probabilités est une importante branche des mathématiques à part entière, entièrement distincte de son application aux statistiques et à l'analyse des données. Ainsi, il existe des milliers de livres écrits sur le sujet et les universités offrent généralement de multiples cours entièrement consacrés à la théorie des probabilités. Même la tâche « plus simple » de documenter les distributions de probabilités standard est un grand sujet. J'ai décrit cinq distributions de probabilités standard dans ce chapitre, mais j'ai un livre de 45 chapitres intitulé « Statistical Distributions » (Evans, Barston, and Pollard 1983) qui contient *beaucoup* plus que cela. Heureusement pour vous, très peu sont nécessaires. Il est peu probable que vous ayez besoin de connaître des douzaines de distributions statistiques lorsque vous effectuez des analyses de données dans le monde réel, et vous n'en aurez

certainement pas besoin pour ce livre, mais cela ne fait jamais de mal de savoir qu'il y a d'autres possibilités.

Pour en revenir à ce dernier point, on a l'impression que tout ce chapitre n'est qu'une digression. Beaucoup d'étudiants des cours de psychologie de premier cycle en statistique lisent ce contenu très rapidement (je sais que le mien l'a fait), et même les cours les plus avancés « oublient » souvent de revoir les fondements fondamentaux du domaine. La plupart des psychologues universitaires ne connaîtraient pas la différence entre la probabilité et la densité et, jusqu'à tout récemment, très peu d'entre eux étaient au courant de la différence entre la probabilité bayésienne et la probabilité fréquentiste. Cependant, je pense qu'il est important de comprendre ces choses avant de passer aux applications. Par exemple, il y a beaucoup de règles sur ce que vous êtes « autorisé » à dire lorsque vous faites des inférences statistiques et beaucoup d'entre elles peuvent sembler arbitraires et étranges. Cependant, elles commencent à avoir du sens si vous comprenez qu'il y a cette distinction bayésienne/fréquentiste. De même, au chapitre 11, nous allons parler de ce qu'on appelle le *t-test*, et si vous voulez vraiment avoir une idée de la mécanique du *t-test,* il est vraiment utile d'avoir une idée de ce à quoi ressemble réellement une *distribution t.* Vous comprenez l'idée, j'espère.

Estimation de quantités inconnues à partir d'un échantillon

Au début du dernier chapitre, j'ai souligné la distinction cruciale entre *statistiques descriptives* et *statistiques inférentielles*. Comme nous l'avons vu au chapitre 4, le rôle des statistiques descriptives est de résumer de façon concise ce que nous savons. En revanche, le but des statistiques inférentielles est « d'apprendre ce que nous ne savons pas de ce que nous faisons ». Maintenant que nous avons une base en théorie des probabilités, nous sommes bien placés pour réfléchir au problème de l'inférence statistique. Quels genres de choses aimerions-nous apprendre ? Et comment les apprend-on ? Telles sont les questions qui sont au cœur des statistiques inférentielles, et elles sont traditionnellement divisées en deux « grandes idées » : l'estimation et la vérification d'hypothèses. Le but de ce chapitre est de présenter la première de ces grandes idées, la théorie de l'estimation, mais je vais d'abord parler de la théorie de l'échantillonnage parce que la théorie de l'estimation n'a de sens que si vous comprenez l'échantillonnage. Par conséquent, le présent chapitre se divise naturellement en deux parties : les sections 8.1 à 8.3 sont axées sur la théorie de l'échantillonnage, et les sections 8.4 et 8.5 utilisent la théorie de l'échantillonnage pour discuter de la façon dont les statisticiens envisagent l'estimation.

Échantillons, populations et échantillonnage

Dans le préambule de la quatrième partie, j'ai parlé de l'énigme de l'insertion et j'ai souligné le fait que *tout* apprentissage exige que l'on fasse des hypothèses. Accepter que c'est vrai, c'est notre première tâche que de formuler des hypothèses assez générales sur des données qui ont du sens. C'est là qu'intervient la **théorie de l'échantillonnage.** Si la théorie des probabilités est le fondement de toute théorie statistique, la théorie de l'échantillonnage est le cadre autour duquel vous pouvez construire le reste de la maison. La théorie de l'échantillonnage joue un rôle énorme en précisant les hypothèses sur lesquelles reposent

vos inférences statistiques. Pour parler de la façon dont les statisticiens perçoivent les inférences, nous devons être un peu plus explicites sur ce *dont* nous tirons des inférences (l'échantillon) et *sur* ce *à propos de quoi* nous tirons des inférences (la population).

Dans presque toutes les situations qui nous intéressent, les données dont nous disposons en tant que chercheurs sont un **échantillon** de données. Nous avons peut-être fait des expériences avec un certain nombre de participants, une société de sondage a peut-être téléphoné à un certain nombre de personnes pour leur poser des questions sur leurs intentions de vote, et ainsi de suite. De cette façon, l'ensemble des données dont nous disposons est fini et incomplet. Par exemple, une société de sondage n'a ni le temps ni l'argent nécessaires pour sonder tous les électeurs du pays. Dans notre discussion précédente sur les statistiques descriptives (chapitre 4), cet échantillon était la seule chose qui nous intéressait. Notre seul but était de trouver des moyens de décrire, de résumer et de représenter graphiquement cet échantillon. Cela est sur le point de changer.

Définir une population

Un échantillon est une chose concrète. Vous pouvez ouvrir un fichier de données et il y a les données de votre échantillon. Une **population**, par contre, est une idée plus abstraite. Il s'agit de l'ensemble de toutes les personnes possibles, ou de toutes les observations possibles, au sujet desquelles vous voulez tirer des conclusions et qui est généralement *beaucoup* plus grand que l'échantillon. Dans un monde idéal, le chercheur commencerait l'étude avec une idée claire de ce qu'est la population d'intérêt, puisque le processus de conception d'une étude et de vérification des hypothèses avec les données dépend de la population au sujet de laquelle vous voulez faire des affirmations.

Parfois, il est facile d'indiquer la population d'intérêt. Par exemple, dans l'exemple de la « société de sondage » qui a ouvert le chapitre, la population se composait de tous les électeurs inscrits au moment de l'étude, des millions de personnes. L'échantillon était constitué d'un ensemble de 1000 personnes qui appartiennent toutes à cette population. Dans la plupart des études, la situation est beaucoup moins simple. Dans une expérience psychologique typique, déterminer la population d'intérêt est un peu plus compliqué. Supposons que je mène une expérience à laquelle participent 100 étudiants de premier cycle. Mon but, en tant que cognitiviste, est d'essayer d'apprendre quelque chose sur le fonctionnement de l'esprit. De ce point de vue, lequel des éléments suivants correspondrait à la « population » :

- Tous les étudiants en psychologie de l'Université d'Adélaïde ?
- Les étudiants en psychologie de premier cycle en général, n'importe où dans le monde ?
- Des Australiens vivent actuellement ?
- Des Australiens du même âge que mon échantillon ?
- Quelqu'un de vivant ?
- Un être humain, passé, présent ou futur ?
- Tout organisme biologique ayant un degré d'intelligence suffisant et opérant dans un environnement terrestre ?

- Un être intelligent ?

Chacune de ces définitions définit un véritable groupe d'entités possédant un esprit, qui pourraient toutes m'intéresser en tant que cognitiviste, et savoir quelle devrait être la véritable population d'intérêt n'est pas du tout clair. Prenons un autre exemple, celui du jeu Wellesley-Croker dont nous avons discuté dans l'introduction. L'échantillon ici est une séquence spécifique de 12 victoires et 0 défaite pour Wellesley. Quelle est la population ?

- Tous les résultats jusqu'à ce que Wellesley et Croker arrivent à destination ?
- Tous les résultats si Wellesley et Croker avaient joué le jeu pour le reste de leur vie ?
- Tous les résultats si Wellseley et Croker vivaient éternellement et jouaient le jeu jusqu'à ce que le monde soit à court de collines ?
- Tous les résultats si nous créions un ensemble infini d'univers parallèles et que la paire Wellesely/Croker faisait des suppositions sur les 12 mêmes collines dans chaque univers ?

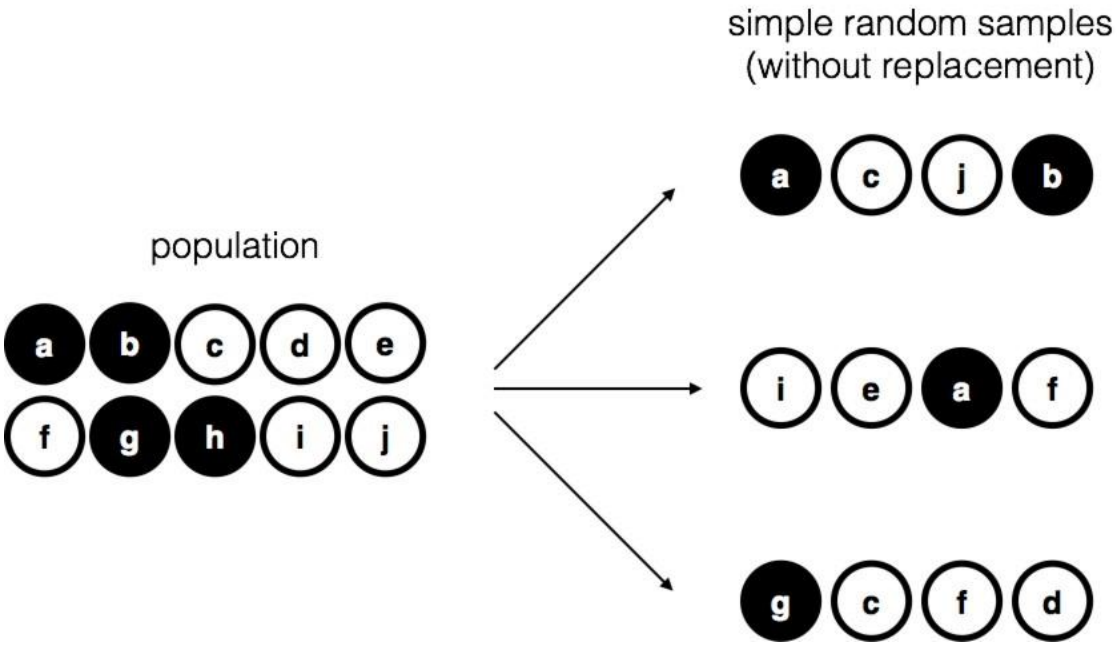


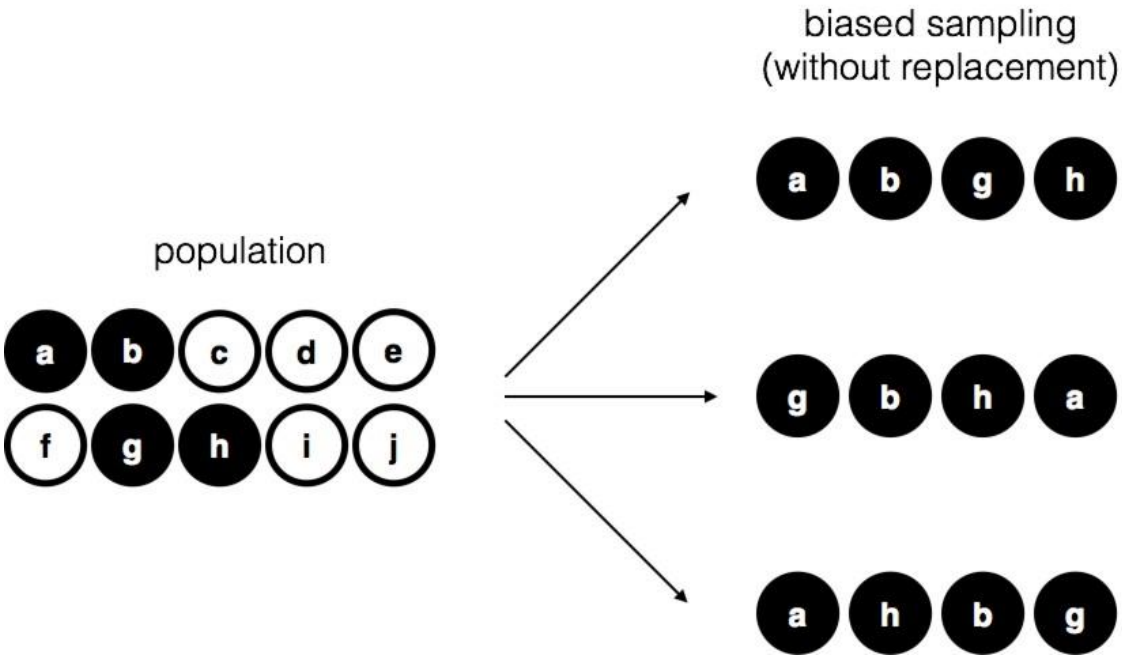
Figure 8-1 : Échantillonnage aléatoire simple sans remplacement à partir d'une population finie

Encore une fois, ce n'est pas évident de savoir quelle est la population.

Échantillons aléatoires simples

Quelle que soit ma définition de la population, le point critique est que l'échantillon est un sous-ensemble de la population et que notre but est d'utiliser notre connaissance de l'échantillon pour tirer des conclusions sur les propriétés de la population. La relation entre les deux dépend de la *procédure de* sélection de l'échantillon. Cette procédure est appelée **méthode d'échantillonnage** et il est important de comprendre pourquoi elle est importante.

Pour simplifier les choses, imaginons que nous ayons un sac contenant 10 jetons. Chaque jeton a une lettre unique imprimée sur lui afin que nous puissions distinguer les 10 jetons. Les jetons sont disponibles en deux couleurs, noir et blanc. Cet ensemble de jetons est la population d'intérêt et il est représenté graphiquement à gauche de la Figure 8-1. Comme vous pouvez le voir en regardant l'image, il y a 4 jetons noirs et 6 jetons blancs, mais bien sûr dans la vraie vie nous ne le saurions pas si nous ne regardons pas dans le sac. Imaginez maintenant que vous faites « l'expérience » suivante : vous secouez le sac, fermez les yeux et retirez 4 jetons sans en remettre aucune dans le sac. D'abord le jeton *a* (noire), puis le jeton *c* (blanche), puis *j* (blanche) et enfin *b* (noire). Si vous le souhaitez, vous pouvez ensuite remettre toutes les jetons dans le sac et répéter l'expérience, comme illustré à droite sur la Figure 8-1. Chaque fois que vous obtenez des résultats différents, mais la procédure est identique dans chaque cas. Le fait qu'une même procédure peut conduire à des résultats différents à chaque fois nous conduits à parler d'un processus *aléatoire.*[44] Cependant, parce que nous avons secoué le sac avant de retirer les jetons, il semble raisonnable de penser que chaque jeton a les mêmes chances d'être sélectionnée. Une procédure dans laquelle chaque membre de la population a les mêmes chances d'être sélectionné s'appelle un **simple échantillon aléatoire**. Le fait que nous *n'ayons pas* remis les jetons dans le sac après les avoir retirées signifie que vous ne pouvez pas observer la même chose deux fois, et dans de tels cas les observations sont les suivantes dont on dit qu'ils ont été échantillonnés **sans remise**.



[44] La définition mathématique correcte du hasard est extraordinairement technique et dépasse largement le cadre de ce livre. Nous ne serons pas techniques ici et dirons qu'un processus comporte un élément de hasard chaque fois qu'il est possible de répéter le processus et d'obtenir des réponses différentes à chaque fois.

Figure 8-2 : Échantillonnage biaisé sans remplacement à partir d'une population finie

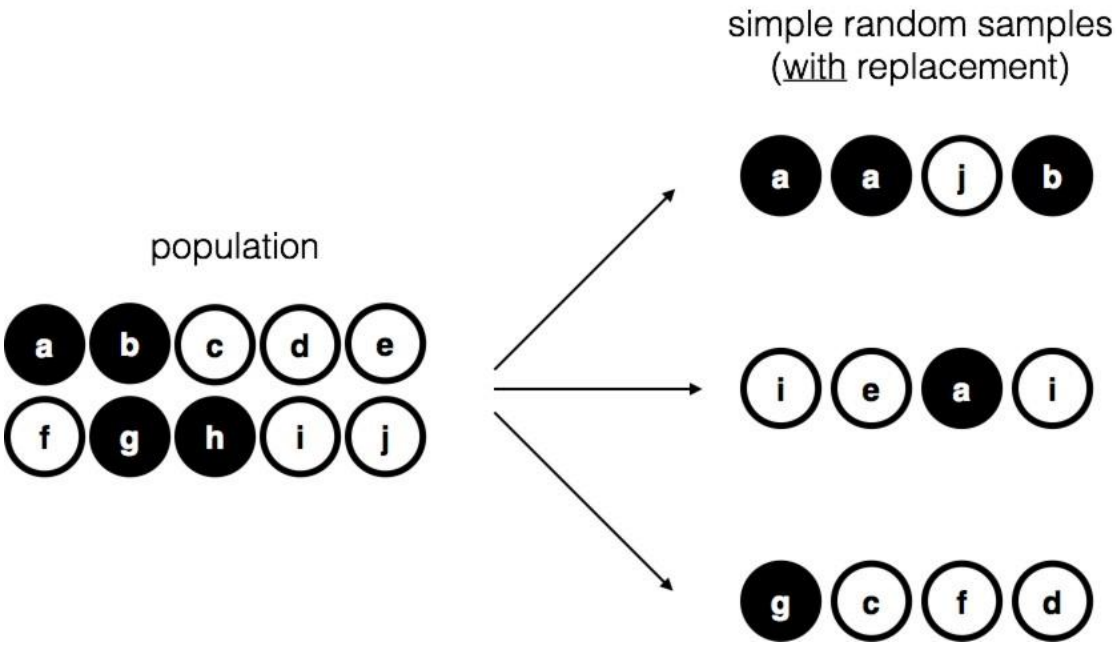


Figure 8-3 : Échantillonnage aléatoire simple avec remplacement dans une population finie

Pour vous assurer de bien comprendre l'importance de la procédure d'échantillonnage, envisagez une autre façon de procéder à l'expérience. Supposons que mon fils de 5 ans ait ouvert le sac et décidé de retirer quatre jetons noirs sans en remettre aucun dans le sac. Ce plan d'échantillonnage *biaisé* est illustré à la Figure 8-2. Considérons maintenant la valeur probante de voir 4 jetons noirs et 0 jetons blancs. Cela dépend clairement du plan d'échantillonnage. Si vous savez que le plan d'échantillonnage est biaisé pour ne sélectionner que des jetons noirs, alors un échantillon composé uniquement de jetons noirs ne vous dit pas grand-chose sur la population ! C'est la raison pour laquelle les statisticiens aiment beaucoup qu'un ensemble de données puisse être considéré comme un simple échantillon aléatoire, parce qu'il rend l'analyse des données *beaucoup* plus facile.

Une troisième procédure mérite d'être mentionnée. Cette fois-ci, nous fermons les yeux, secouons le sac et sortons un jeton. Cette fois, cependant, nous enregistrons l'observation et remettons le jeton dans le sac. Encore une fois, nous fermons les yeux, secouons le sac et sortons un jeton. Nous répétons ensuite cette procédure jusqu'à ce que nous ayons 4 jetons. Les ensembles de données générés de cette façon sont encore de simples échantillons aléatoires, mais parce que nous remettons les jetons dans le sac immédiatement après les avoir tirés, on parle d'un échantillon **avec remise**. La différence entre cette situation et la première est qu'il est possible d'observer le même membre de la population plusieurs fois, comme l'illustre la Figure 8-3.

D'après mon expérience, la plupart des expériences de psychologie ont tendance à être des échantillonnages sans remise, parce que la même personne n'est pas autorisée à participer deux fois à l'expérience. Toutefois, la plupart des théories statistiques reposent sur l'hypothèse que les données proviennent d'un simple échantillon aléatoire *avec* remise.

Dans la vie réelle, cela importe très rarement. Si la population d'intérêt est importante (p. ex. a plus de 10 entités !), la différence entre l'échantillonnage avec ou sans remise est trop faible pour être prise en compte. La différence entre les échantillons aléatoires simples et les échantillons biaisés, par contre, n'est pas une chose si facile à rejeter.

La plupart des échantillons ne sont pas de simples échantillons aléatoires

Comme vous pouvez le constater en regardant la liste des populations possibles que j'ai présentée ci-dessus, il est presque impossible d'obtenir un échantillon aléatoire simple de la plupart des populations d'intérêt. Quand je fais des expériences, je considérerais comme un petit miracle que mes participants soient un échantillon aléatoire d'étudiants en psychologie de premier cycle de l'université d'Adélaïde, même si c'est de loin la population la plus étroite à laquelle je voudrais généraliser. Une discussion approfondie d'autres types de plans d'échantillonnage dépasse la portée de ce livre, mais pour vous donner une idée de ce qui existe, je vais en énumérer quelques-uns des plus importants.

- *Échantillonnage stratifié*. Supposons que votre population soit (ou puisse être) divisée en plusieurs sous-populations ou *strates* différentes. Peut-être menez-vous une étude à plusieurs endroits différents, par exemple. Au lieu d'essayer d'échantillonner au hasard l'ensemble de la population, vous essayez plutôt de prélever un échantillon aléatoire distinct dans chacune des strates. L'échantillonnage stratifié est parfois plus facile à réaliser qu'un simple échantillonnage aléatoire, surtout lorsque la population est déjà divisée en strates distinctes. Il peut aussi être plus efficace qu'un simple échantillonnage aléatoire, surtout lorsque certaines sous-populations sont rares. Par exemple, lorsqu'on étudie la schizophrénie, il vaudrait beaucoup mieux diviser la population en deux[45] strates (schizophrène et non schizophrène), puis échantillonner un nombre égal de personnes de chaque groupe. Si vous choisissiez des personnes au hasard, vous obtiendriez si peu de schizophrènes dans l'échantillon que votre étude serait inutile. Ce type particulier d'échantillonnage stratifié est appelé *sur échantillonnage* parce qu'il constitue une tentative délibérée de surreprésenter des groupes rares.
- *L'échantillonnage en boule de neige est* une technique particulièrement utile lorsqu'il s'agit d'échantillonner une population « cachée » ou difficile d'accès et est particulièrement courante en sciences sociales. Supposons, par exemple, que les chercheurs souhaitent mener un sondage d'opinion auprès des personnes transgenres. L'équipe de recherche pourrait n'avoir que les coordonnées de quelques personnes transgenres, de sorte que l'enquête commence par leur demander de participer (étape 1). A la fin de l'enquête, les participants sont invités à fournir les coordonnées d'autres personnes qui pourraient souhaiter participer. Au cours de l'étape 2, ces nouveaux contacts font l'objet d'une enquête. Le processus se poursuit jusqu'à ce que les chercheurs aient suffisamment de données.

---

[45] Rien dans la vie n'est aussi simple. Il n'y a pas une division évidente des gens en catégories binaires comme "schizophrène" et "pas schizophrène". Mais ce n'est pas un texte de psychologie clinique, alors pardonnez-moi quelques simplifications ici et là.

Le grand avantage de l'échantillonnage en boule de neige est qu'il permet d'obtenir des données dans des situations qu'il serait impossible d'obtenir autrement. Du point de vue statistique, le principal inconvénient est que l'échantillon est particulièrement non aléatoire de sorte qu'il est difficile à traiter. Du côté de la vie réelle, l'inconvénient est que la procédure peut être contraire à l'éthique si elle n'est pas bien gérée, car les populations cachées le sont souvent pour une raison. J'ai choisi les personnes transgenres comme exemple ici pour mettre en évidence cette question. Si vous n'étiez pas prudent, vous pourriez finir par sortir avec des gens qui ne veulent pas être révélés (très mauvaise approche), et même si vous ne faites pas cette erreur, il peut toujours être intrusif d'utiliser les réseaux sociaux des sujets pour les étudier. Il est certainement très difficile d'obtenir le consentement éclairé des gens *avant de* les contacter, mais dans de nombreux cas, le simple fait de les contacter et de leur dire « Nous voulons vous étudier » peut être blessant. Les réseaux sociaux sont des choses complexes, et ce n'est pas parce que vous pouvez les utiliser pour obtenir des données que vous devriez toujours le faire.

- *L'échantillonnage de commodité* est plus ou moins ce à quoi il ressemble. Les échantillons sont choisis d'une manière qui convient au chercheur et non au hasard dans la population d'intérêt. L'échantillonnage en boule de neige est un type d'échantillonnage de commodité, mais il en existe beaucoup d'autres. Un exemple courant en psychologie est celui des études qui s'appuient sur des étudiants de premier cycle en psychologie. Ces échantillons sont généralement non aléatoires à deux égards. Premièrement, le fait de se fier aux étudiants de premier cycle en psychologie signifie automatiquement que vos données sont limitées à une seule sous-population. Deuxièmement, les étudiants choisissent habituellement les études auxquelles ils participent, de sorte que l'échantillon est un sous-ensemble auto-sélectionné d'étudiants en psychologie et non un sous-ensemble choisi au hasard. Dans la vie réelle, la plupart des études sont des échantillons de commodité d'une façon ou d'une autre. Il s'agit parfois d'une limite importante, mais pas toujours.

Quelle importance cela a-t-il si vous n'avez pas un échantillon aléatoire simple ?

Bien, la collecte de données du monde réel n'a pas tendance à impliquer des échantillons aléatoires simples et agréables. C'est important ? Un peu de réflexion devrait vous faire comprendre qu'il *peut être* important que vos données ne soient pas un simple échantillon aléatoire. Pensez simplement à la différence entre les Figure 8-1 et Figure 8-2. Cependant, ce n'est pas aussi grave que ça en a l'air. Certains types d'échantillons biaisés ne posent aucun problème. Par exemple, lorsque vous utilisez une technique d'échantillonnage stratifié, vous *savez en* fait quel est le biais parce que vous l'avez créé délibérément, souvent pour *accroître l'*efficacité de votre étude, et il existe des techniques statistiques que vous pouvez utiliser pour corriger les biais que vous avez présentés (non couverts dans ce livre !). Donc, dans ces situations, ce n'est pas un problème.

De façon plus générale, cependant, il est important de se rappeler que l'échantillonnage aléatoire est un moyen d'atteindre une fin, et non une fin en soi. Supposons que vous vous êtes fiés à un échantillon de commodité et que, par conséquent, vous pouvez supposer qu'il est biaisé. Un biais dans votre méthode d'échantillonnage ne pose problème que s'il vous amène à tirer des conclusions erronées. De ce point de vue, je dirais que nous n'avons pas

besoin que l'échantillon soit généré au hasard à *tous les* égards, nous avons seulement besoin qu'il soit aléatoire en ce qui concerne le phénomène pertinent sur le plan psychologique qui nous intéresse. Supposons que je fasse une étude sur la capacité de la mémoire de travail. Dans l'étude 1, j'ai en fait la capacité d'échantillonner au hasard tous les êtres humains actuellement vivants, à une exception près : Je ne peux tester que les gens nés le lundi. Dans l'étude 2, je suis en mesure d'échantillonner au hasard la population australienne. Je veux généraliser mes résultats à la population de tous les humains vivants. Quelle étude est la meilleure ? La réponse, évidemment, est l'étude 1. Pourquoi ? Parce que nous n'avons aucune raison de penser qu'être « né un lundi » a un rapport intéressant avec la capacité de mémoire de travail. En revanche, je peux penser à plusieurs raisons pour lesquelles « être Australien » peut avoir de l'importance. L'Australie est un pays riche et industrialisé avec un système éducatif très développé. Les personnes qui ont grandi dans ce système auront vécu des expériences de vie plus semblables à celles des personnes qui ont conçu les tests de capacité de mémoire de travail. Cette expérience commune pourrait facilement se traduire par des croyances similaires sur la façon de passer un test, une supposition commune sur le fonctionnement de l'expérimentation psychologique, et ainsi de suite. Ces choses pourraient avoir de l'importance. Par exemple, le style « passation de test » pourrait avoir appris aux participants australiens à se concentrer exclusivement sur des sujets assez abstraits, beaucoup plus que les personnes qui n'ont pas grandi dans un environnement similaire. Ce pourrait donc induire en erreur sur ce qu'est la capacité de la mémoire de travail. Cela pourrait donc conduire à une image trompeuse de ce qu'est la capacité de mémoire de travail.

Il y a deux points cachés dans cette discussion. Premièrement, lorsque vous concevez vos propres études, il est important de penser à la population à laquelle vous tenez et d'essayer d'échantillonner d'une manière qui convient à cette population. En pratique, vous êtes habituellement obligé de vous contenter d'un « échantillon de convenance » (p. ex. les professeurs de psychologie échantillonnent les étudiants en psychologie parce que c'est la façon la moins coûteuse de recueillir des données, et nos coffres ne débordent pas vraiment d'or), mais si c'est le cas, vous devriez au moins prendre le temps de penser aux dangers que cette pratique pourrait représenter. Deuxièmement, si vous allez critiquer l'étude de quelqu'un d'autre parce qu'il a utilisé un échantillon de convenance plutôt que d'échantillonner laborieusement au hasard toute la population humaine, ayez au moins la courtoisie d'offrir une théorie précise sur la *façon dont* cela a pu fausser les résultats.

Paramètres de population et statistiques des échantillons

Si l'on met de côté les épineuses questions méthodologiques associées à l'obtention d'un échantillon aléatoire, examinons une question légèrement différente. Jusqu'à présent, nous avons parlé des populations comme le ferait un scientifique. Pour un psychologue, une population peut être un groupe de personnes. Pour un écologiste, une population peut être un groupe d'ours. Dans la plupart des cas, les populations auxquelles s'intéressent les scientifiques sont des choses concrètes qui existent réellement dans le monde réel. Les statisticiens, cependant, sont un drôle de lot. D'une part, ils *s'*intéressent aux données du monde réel et à la vraie science de la même façon que les scientifiques. D'autre part, ils opèrent également dans le domaine de l'abstraction pure, comme le font les mathématiciens. Par conséquent, la théorie statistique tend à être un peu abstraite

concernant la façon dont une population est définie. De la même manière que les chercheurs en psychologie opérationnalisent nos idées théoriques abstraites en termes de mesures concrètes (Section 2.1), les statisticiens opérationnalisent le concept de « population » en termes d'objets mathématiques avec lesquels ils savent comment travailler. Vous avez déjà rencontré ces objets au chapitre 7. C'est ce qu'on appelle des distributions de probabilités.

L'idée est assez simple. Disons qu'on parle de QI. Pour un psychologue, la population d'intérêt est un groupe d'humains qui ont un QI. Un statisticien « simplifie » cela en définissant de façon opérationnelle la population comme étant la distribution de probabilité représentée à la Figure 8-4a. Les tests de QI sont conçus de manière à ce que le QI moyen soit de 100, que l'écart-type des scores de QI soit de 15 et que la distribution des scores de QI soit normale. Ces valeurs sont appelées **paramètres de population** parce qu'elles représentent les caractéristiques de l'ensemble de la population. C'est-à-dire que nous disons que la moyenne de la population μ est de 100 et que l'écart-type de la population σ est de 15.

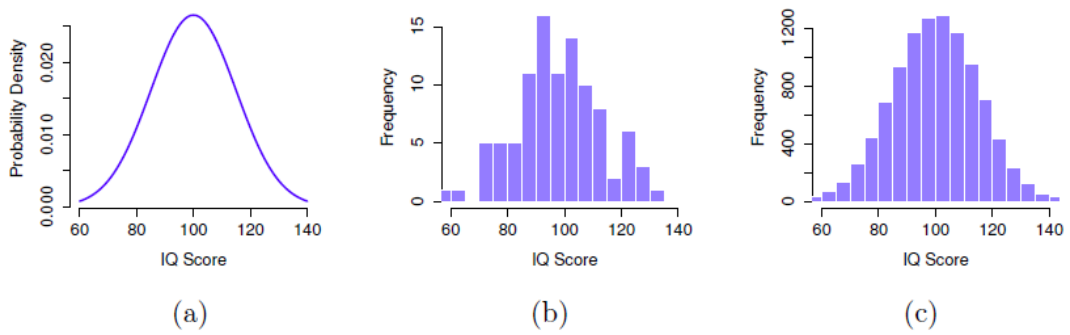


Figure 8-4 : La distribution de la population des scores de QI (panel a) et deux échantillons tirés au hasard. Dans le panel b, nous avons un échantillon de 100 observations, et dans le panel c, nous avons un échantillon de 10 000 observations.

Supposons que je fasse une expérience. Je sélectionne 100 personnes au hasard et je leur fais passer un test de QI, Figure 8-4 : Répartition des scores de QI dans la population (panel a) et deux échantillons tirés au hasard. Dans le panel b, nous avons un échantillon de 100 observations, et dans le panel c, nous avons un échantillon de 10 000 observations me donnant un simple échantillon aléatoire de la population. Mon échantillon consisterait en une collection de chiffres comme celle-ci :

```
106 101 98 80 74 ... 107 72 100
```

Chacun de ces scores de QI est échantillonné à partir d'une distribution normale avec une moyenne de 100 et un écart-type de 15. Donc, si je trace un histogramme de l'échantillon, j'obtiens quelque chose comme celui illustré à la Figure 8-4b. Comme vous pouvez le voir, l'histogramme est à *peu près de* la bonne forme, mais c'est une approximation très grossière de la distribution réelle de la population présentée à la Figure 8-4a. Lorsque je calcule la

moyenne de mon échantillon, j'obtiens un nombre assez proche de la moyenne de la population (100) mais pas identique. Dans ce cas, il s'avère que les personnes de mon échantillon ont un QI moyen de 98,5 et que l'écart-type de leur QI est de 15,9. Ces **statistiques d'échantillon** sont des propriétés de mon ensemble de données, et bien qu'elles soient assez semblables aux valeurs réelles de la population, elles ne sont pas les mêmes. En général, les statistiques d'échantillon sont les choses que vous pouvez calculer à partir de votre ensemble de données et les paramètres de population sont les choses que vous voulez apprendre. Plus loin dans ce chapitre, je parlerai de la façon dont vous pouvez estimer les paramètres de population à l'aide de vos statistiques d'échantillonnage (section 8.4) et de la façon de déterminer dans quelle mesure vous pouvez faire confiance à vos estimations (section 8.5), mais avant d'y arriver, il y a quelques autres notions à propos de la théorie d'échantillonnage que vous devez connaître.

La loi des grands nombres

Dans la section précédente, je vous ai montré les résultats d'une expérience fictive de QI avec un échantillon de N « 100. Les résultats étaient quelque peu encourageants puisque la moyenne réelle de la population est de 100 et que la moyenne de l'échantillon de 98,5 est une approximation assez raisonnable de celle-ci. Dans de nombreuses études scientifiques, ce niveau de précision est parfaitement acceptable, mais dans d'autres situations, il faut être beaucoup plus précis. Si nous voulons que nos statistiques d'échantillonnage soient beaucoup plus proches des paramètres de la population, que pouvons-nous faire à ce sujet ?

La réponse évidente est de recueillir plus de données. Supposons que nous ayons mené une expérience beaucoup plus vaste, cette fois en mesurant le QI de 10 000 personnes. Nous pouvons simuler les résultats de cette expérience en utilisant Jamovi. Le fichier IQsim.omv est un fichier de données Jamovi. Dans ce fichier, j'ai généré 10 000 nombres aléatoires échantillonnés à partir d'une distribution normale pour une population avec moyenne = 100 et sd = 15. Ceci a été fait en calculant une nouvelle variable à l'aide de la fonction = NORM(100,15). Un histogramme et un graphique de densité montrent que cet échantillon plus grand est une bien meilleure approximation de la distribution réelle de la population que le plus petit. Cela se reflète dans les statistiques de l'échantillon. Le QI moyen de l'échantillon le plus important est de 99,68 et l'écart-type est de 14,90. Ces valeurs sont maintenant très proches de la population réelle. Voir la Figure 8-5.

Je me sens un peu bête de dire cela, mais ce que je veux que vous reteniez de tout cela, c'est que les gros échantillons vous donnent généralement de meilleurs renseignements. Je me sens bête de le dire parce que c'est tellement évident que ça ne devrait pas avoir besoin d'être dit. En fait, c'est tellement évident que lorsque Jacob Bernoulli, l'un des fondateurs de la théorie des probabilités, a formalisé cette idée en 1713, il a été un peu bêta sur ce point. Voici comment il a décrit le fait que nous partageons tous cette intuition :

Car même le plus stupide des hommes, par un instinct de nature, seul et sans instruction (ce qui est remarquable), est convaincu que plus on a fait d'observations, moins on risque de s'éloigner de son but (voir Stigler (1986), p65).

D'accord, le passage semble un peu condescendant (pour ne pas dire sexiste), mais son point principal est correct. Il semble vraiment évident que plus de données vous donneront de meilleures réponses. La question est : pourquoi en est-il ainsi ? Il n'est pas surprenant

que cette intuition que nous partageons tous s'avère correcte, et les statisticiens l'appellent la **loi des grands nombres**. La loi des grands nombres est une loi mathématique qui s'applique à de nombreuses statistiques d'échantillons différents, mais la façon la plus simple d'y penser est d'adopter une loi sur les moyennes. La moyenne de l'échantillon est l'exemple le plus évident d'une statistique qui repose sur le calcul de la moyenne (parce que c'est ce qu'est la moyenne… une moyenne), alors voyons cela. Lorsqu'on l'applique à la moyenne de l'échantillon, la loi des grands nombres indique qu'à mesure que l'échantillon s'élargit, la moyenne de l'échantillon tend à se rapprocher de la moyenne réelle de la population. Ou, pour le dire un peu plus précisément, lorsque la taille de l'échantillon «approche» de l'infini (écrit $N \to \infty$), la moyenne de l'échantillon approche la moyenne de la population ($X \to \mu$).[46]

Je n'ai pas l'intention de vous démontrer que la loi des grands nombres est vraie, mais c'est l'un des outils les plus importants de la théorie statistique. La loi des grands nombres est la chose que nous pouvons utiliser pour justifier notre croyance que la collecte de plus en plus de données nous mènera finalement à la vérité. Pour n'importe quel ensemble de données particulier, les statistiques d'échantillon que nous calculons à partir de celui-ci seront erronées, mais la loi des grands nombres nous dit que si nous continuons à recueillir plus de données, ces statistiques d'échantillon auront tendance à se rapprocher de plus en plus des véritables paramètres démographiques.

---

[46] Techniquement, la loi des grands nombres s'applique à tout échantillon statistique qui peut être décrit comme une moyenne de quantités indépendantes. C'est certainement vrai pour la moyenne de l'échantillon. Cependant, il est également possible d'écrire beaucoup d'autres exemples de statistiques sous forme de moyennes d'une forme ou d'une autre. La variance d'un échantillon, par exemple, peut être réécrite comme une sorte de moyenne et est donc soumise à la loi des grands nombres. La valeur minimale d'un échantillon, cependant, ne peut pas être écrite comme une moyenne de quoi que ce soit et n'est donc pas régie par la loi des grands nombres.

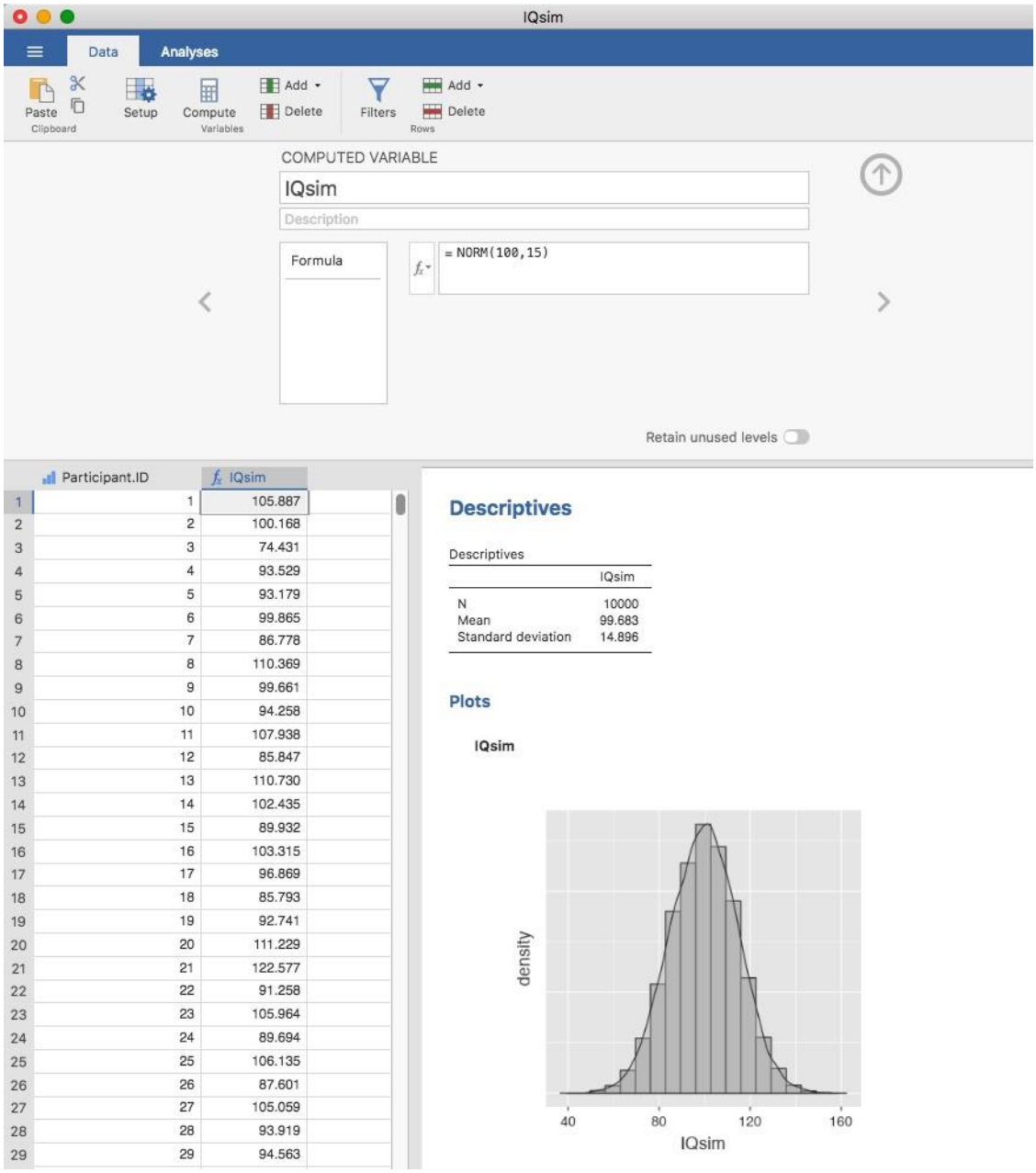


Figure 8-5 : Un échantillon aléatoire tiré d'une distribution normale à l'aide de Jamovi

Les distributions d'échantillonnage et le théorème de la limite centrale

La loi des grands nombres est un outil très puissant, mais elle ne suffira pas à répondre à toutes nos questions. Il ne nous donne entre autres qu'une « garantie à long terme ». À long terme, si nous étions en mesure de recueillir une quantité infinie de données, la loi des grands nombres garantirait que nos statistiques d'échantillonnage seraient exactes. Mais comme John Maynard Keynes l'a fait valoir en économie, une garantie à long terme n'est guère utile dans la vie réelle.

[Le] long terme est un guide trompeur de l'actualité. À long terme, nous sommes tous morts. Les économistes se fixent une tâche trop facile, trop inutile, s'ils ne peuvent nous dire, en période de tempête, que lorsque la tempête est passée depuis longtemps, l'océan est à nouveau plat. (Keynes (2009), p. 80)

Comme en économie, il en va de même en psychologie et en statistique. Il ne suffit pas de savoir que nous parviendrons *éventuellement à la* bonne réponse lors du calcul de la moyenne de l'échantillon. Savoir qu'un ensemble de données infiniment grand me dira la valeur exacte de la moyenne de la population est peu réconfortant lorsque mon ensemble de données *réelles* a une taille d'échantillon de $N = 100$. Dans la vraie vie, il faut donc savoir quelque chose sur le comportement de la moyenne de l'échantillon lorsqu'elle est calculée à partir d'un ensemble de données plus modeste !

Distribution d'échantillonnage de la moyenne

Dans cette optique, abandonnons l'idée que nos études auront des échantillons de 10 000 personnes et considérons plutôt qu'il s'agit d'une expérience très modeste. Cette fois-ci, nous allons échantillonner $N = 5$ personnes et mesurer leur QI. Comme avant, je peux simuler cette expérience avec la fonction Jamovi = NORM(100,15), mais je n'ai besoin que de 5 ID de participants cette fois-ci, pas 10 000. Ce sont les cinq nombres que Jamovi a générés :

```
90 82 94 99 110
```

Le QI moyen dans cet échantillon s'avère être exactement 95. Il n'est donc pas surprenant que ce soit beaucoup moins précis que l'expérience précédente. Imaginez maintenant que j'ai décidé de **reproduire** l'expérience. C'est-à-dire que je répète la procédure le plus fidèlement possible et que j'échantillonne au hasard 5 nouvelles personnes et mesure leur QI. Encore une fois, Jamovi me permet de simuler les résultats de cette procédure, et génère ces cinq nombres :

```
78 88 111 111 117
```

Cette fois, le QI moyen de mon échantillon est de 101. Si je répète l'expérience 10 fois, j'obtiens les résultats indiqués dans le Tableau 8-1, et comme vous pouvez le constater, la moyenne de l'échantillon varie d'une réplication à l'autre.

Supposons maintenant que j'ai décidé de continuer dans cette voie, en reproduisant encore et encore cette expérience des « cinq scores de QI ». Chaque fois que je répète l'expérience, je note l'exemple du Tableau 8-1 : Dix répétitions de l'expérience de QI, chacune avec une taille d'échantillon de $N = 5$. Je rapporte la moyenne de l'échantillon

Tableau 8-1 : 10 répétition de l'expérience sur les QI avec un échantillon de taille N=5

	Person 1	Person 2	Person 3	Person 4	Person 5	Sample Mean
Replication 1	90	82	94	99	110	95.0
Replication 2	78	88	111	111	117	101.0
Replication 3	111	122	91	98	86	101.6
Replication 4	98	96	119	99	107	103.8
Replication 5	105	113	103	103	98	104.4
Replication 6	81	89	93	85	114	92.4
Replication 7	100	93	108	98	133	106.4
Replication 8	107	100	105	117	85	102.8
Replication 9	86	119	108	73	116	100.4
Replication 10	95	126	112	120	76	105.8

Avec le temps, j'accumulerais un nouvel ensemble de données, dans lequel chaque expérience génère un seul point de données. Les 10 premières observations de mon ensemble de données sont les moyennes de l'échantillon énumérées au Tableau 8-1, de sorte que mon ensemble de données commence comme ceci :

```
95.0 101.0 101.6 103.8 104.4 ...
```

Et si je continuais comme ça pendant 10 000 répétitions, et que je faisais un histogramme. C'est exactement ce que j'ai fait, et vous pouvez voir les résultats à la Figure 8-6. Comme l'illustre cette figure, la moyenne de 5 QI se situe habituellement entre 90 et 110. Mais plus important encore, ce qu'il met en évidence, c'est que si nous répétons une expérience encore et encore, nous nous retrouvons avec une *distribution de* moyennes d'échantillons ! Cette distribution a un nom spécial dans les statistiques, elle s'appelle la **distribution d'échantillonnage de la moyenne**.

Les distributions d'échantillonnage sont une autre idée théorique importante en statistique, et elles sont cruciales pour comprendre le comportement des petits échantillons. Par exemple, lorsque j'ai fait la toute première expérience des « cinq scores de QI », la moyenne de l'échantillon s'est avérée être de 95. Ce que la distribution d'échantillonnage de la Figure 8-6 nous indique, cependant, c'est que l'expérience des « cinq scores de QI » n'est pas très précise. Si je répète l'expérience, la distribution d'échantillonnage me dit que je peux m'attendre à voir une moyenne d'échantillon entre 80 et 120.

Il existe des distributions d'échantillonnage pour n'importe quelle statistique d'échantillonnage !

Une chose à garder à l'esprit lorsque vous songez aux distributions d'échantillonnage est que *toute* statistique d'échantillonnage que vous pourriez vouloir calculer à une distribution d'échantillonnage. Par exemple, supposons que chaque fois que je répète l'expérience des « cinq scores de QI », je note le plus grand score de QI de l'expérience. Cela me donnerait un ensemble de données qui a commencé comme ça :

```
110 117 122 119 113...
```

Faire cela encore et encore me donnerait une distribution d'échantillonnage très différente, à savoir la *distribution d'échantillonnage du maximum*.

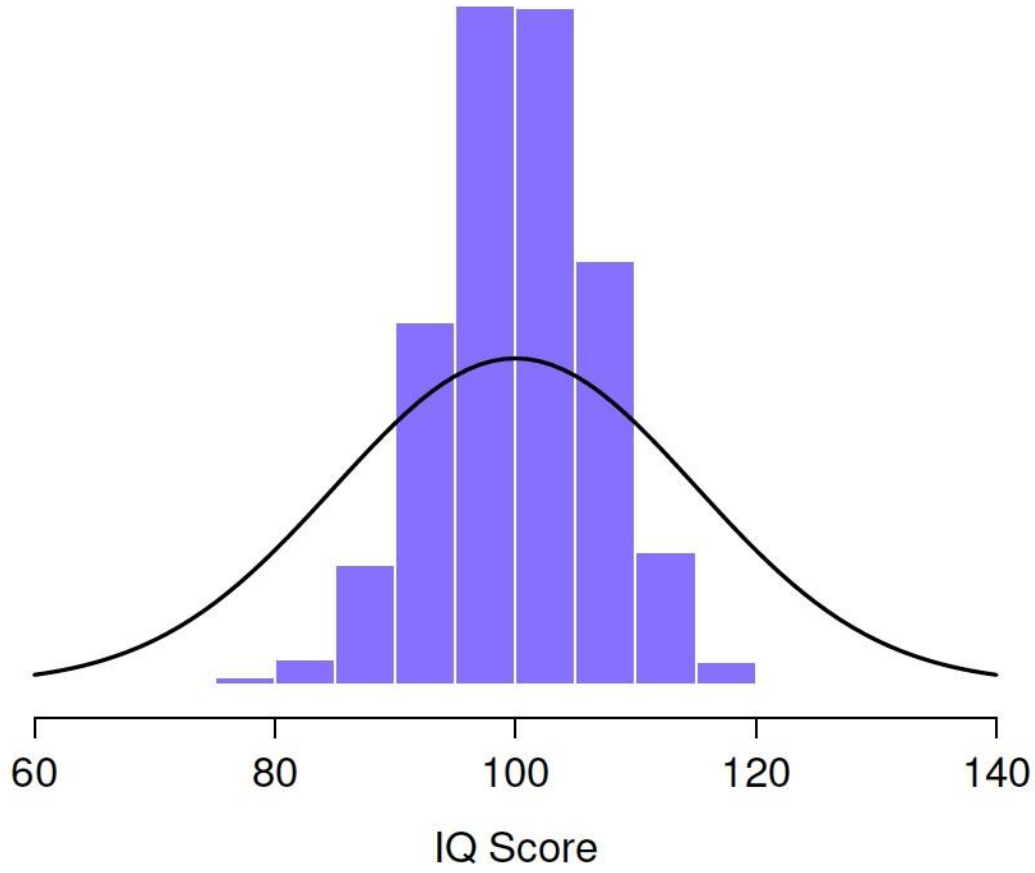


Figure 8-6 : Distribution d'échantillonnage de la moyenne de « l'expérience des cinq scores de QI ». Si vous échantillonnez 5 personnes au hasard et calculez leur QI moyen, vous obtiendrez presque certainement un nombre entre 80 et 120, même s'il y a beaucoup d'individus qui ont un QI supérieur à 120 ou inférieur à 80. À titre de comparaison, la ligne noire représente la distribution des scores de QI dans la population.

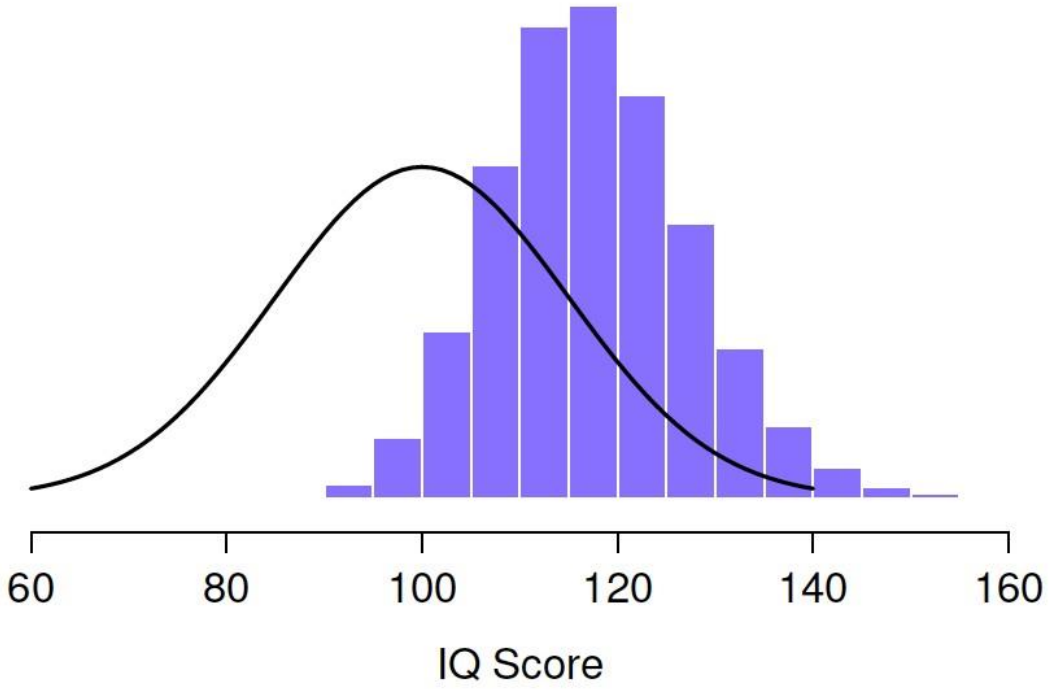


Figure 8-7 : La distribution d'échantillonnage du maximum pour l'«expérience des cinq scores de QI». Si vous échantillonnez 5 personnes au hasard et choisissez celle qui a le QI le plus élevé, vous verrez probablement quelqu'un avec un QI entre 100 et 140.

La distribution d'échantillonnage du maximum de 5 QI est illustrée à la Figure 8-7. Il n'est pas surprenant que si vous choisissez 5 personnes au hasard et que vous trouvez ensuite la personne ayant le QI le plus élevé, elles auront un QI supérieur à la moyenne. La plupart du temps, vous vous retrouverez avec quelqu'un dont le QI se situe entre 100 et 140.

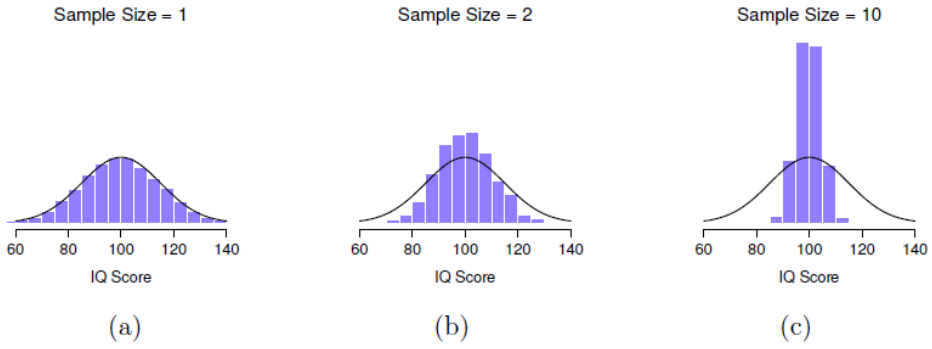


Figure 8-8 : Une illustration de la façon dont la distribution d'échantillonnage de la moyenne dépend de la taille de l'échantillon. Dans chaque panel, j'ai généré 10 000 échantillons de données sur le QI et calculé le QI moyen observé dans chacun de ces ensembles de données. Les histogrammes de ces placettes montrent la distribution de ces

moyennes (c.-à-d. la distribution d'échantillonnage de la moyenne). Chaque score de QI individuel a été tiré d'une distribution normale avec une moyenne de 100 et un écart-type de 15, qui est représenté par la ligne noire pleine. Dans le panel a, chaque ensemble de données ne contenait qu'une seule observation, de sorte que la moyenne de chaque échantillon n'est que le QI d'une personne. Par conséquent, la distribution d'échantillonnage de la moyenne est bien sûr identique à la distribution de la population des scores de QI. Cependant, lorsque nous augmentons la taille de l'échantillon à 2, la moyenne d'un échantillon tend à être plus proche de la moyenne de la population que du quotient intellectuel d'une personne, de sorte que l'histogramme (c.-à-d. la distribution d'échantillonnage) est un peu plus étroit que la distribution dans la population. Lorsque nous portons la taille de l'échantillon à 10 (panel c), nous constatons que la distribution des moyennes de l'échantillon tend à être assez étroitement concentrée autour de la moyenne réelle de la population.

Le théorème de la limite centrale

Pour l'instant, j'espère que vous avez une bonne idée de ce que sont les distributions d'échantillonnage, et en particulier de la distribution d'échantillonnage de la moyenne. Dans cette section, je veux parler de la façon dont la distribution d'échantillonnage de la moyenne change en fonction de la taille de l'échantillon. Intuitivement, vous connaissez déjà une partie de la réponse. Si vous n'avez que quelques observations, la moyenne de l'échantillon est susceptible d'être assez imprécise. Si vous répliquez une petite expérience et recalculez la moyenne, vous obtiendrez une réponse très différente. En d'autres termes, la distribution d'échantillonnage est assez large. Si vous répliquez une grande expérience et recalculez la moyenne de l'échantillon, vous obtiendrez probablement la même réponse que la dernière fois, donc la distribution d'échantillonnage sera très étroite. Vous pouvez le voir visuellement à la Figure 8-8, qui montre que plus la taille de l'échantillon est grande, plus la distribution d'échantillonnage est étroite. Nous pouvons quantifier cet effet en calculant l'écart-type de la distribution d'échantillonnage, qu'on appelle l'**erreur type**. L'erreur-type d'une statistique est souvent appelée SE, et comme nous nous intéressons habituellement à l'erreur-type de la *moyenne de l'*échantillon, nous utilisons souvent l'acronyme SEM. Comme vous pouvez le voir, rien qu'en regardant l'image, plus la taille de l'échantillon *N* augmente, plus le SEM diminue.

Bien, c'est une partie de l'histoire. Cependant, il y a quelque chose que j'ai négligé jusqu'ici. Tous mes exemples jusqu'ici ont été basés sur les expériences de « IQ scores », et parce que les scores de QI sont à peu près normalement distribués, j'ai supposé que la distribution de la population est normale. Et si ce n'est pas normal ? Qu'arrive-t-il à la distribution d'échantillonnage de la moyenne ? Ce qui est remarquable, c'est que, quelle que soit la forme de votre distribution de population, Lorsque N augmente la distribution d'échantillonnage de la moyenne commence à ressembler davantage à une distribution normale. Pour vous donner une idée, j'ai fait quelques simulations. Pour ce faire, j'ai commencé avec la distribution « cumulée" montrée dans l'histogramme de la Figure 8-9. Comme vous pouvez le voir en comparant l'histogramme de forme triangulaire à la courbe en cloche tracée par la ligne noire, la distribution de la population ne ressemble pas du tout à une distribution normale. Ensuite, j'ai simulé les résultats d'un grand nombre d'expériences. Dans chaque expérience, j'ai prélevé N = 2 échantillons de cette distribution,

puis j'ai calculé la moyenne de l'échantillon. La Figure 8-9b présente l'histogramme de ces moyennes d'échantillonnage (c.-à-d. la distribution d'échantillonnage de la moyenne pour N = 2). Cette fois, l'histogramme produit une distribution en cloche. Ce n'est toujours pas normal, mais c'est beaucoup plus près de la ligne noire que la distribution de la population à la Figure 8-9a. Lorsque j'augmente la taille de l'échantillon à N = 4, la distribution d'échantillonnage de la moyenne est très proche de la normale (Figure 8-9c), et au moment où nous atteignons une taille d'échantillon de N = 8, elle est presque parfaitement normale. En d'autres termes, tant que la taille de votre échantillon n'est pas minuscule, la distribution d'échantillonnage de la moyenne sera à peu près normale, peu importe à quoi ressemble la distribution de votre population !

Sur la base de ces chiffres, il semble que nous ayons des preuves pour toutes les allégations suivantes concernant la distribution d'échantillonnage de la moyenne.

- La moyenne de la distribution d'échantillonnage est la même que la moyenne de la population.
- L'écart-type de la distribution d'échantillonnage (c.-à-d. l'erreur type) diminue à mesure que la taille de l'échantillon augmente.
- La forme de la distribution d'échantillonnage devient normale à mesure que la taille de l'échantillon augmente.

En fait, non seulement ces affirmations sont vraies, mais il existe un théorème très célèbre en statistique qui prouve les trois, il est connu sous le nom de **théorème de la limite centrale**. Entre autres choses, le théorème de la limite centrale nous dit que si la distribution de la population a une moyenne μ et un écart-type σ, alors la distribution d'échantillonnage de la moyenne a aussi une moyenne μ et l'erreur type de la moyenne est

$$SEM = \frac{\sigma}{\sqrt{N}}$$

Comme nous divisons l'écart-type de la population σ par la racine carrée de la taille de l'échantillon N, le SEM diminue à mesure que la taille de l'échantillon augmente. Il nous indique également que la forme de la distribution d'échantillonnage devient normale.[47]

Ce résultat est utile pour toutes sortes de choses. Il nous dit pourquoi les grandes expériences sont plus fiables que les petites, et parce qu'il nous donne une formule explicite

---

[47] Comme d'habitude, je suis un peu négligent. Le théorème de la limite centrale est un peu plus général que ne l'implique cette section. Comme la plupart des textes d'introduction aux statistiques, j'ai discuté d'une situation où le théorème de la limite centrale s'applique : lorsque vous prenez une moyenne sur un grand nombre d'événements indépendants tirés de la même distribution. Cependant, le théorème de la limite centrale est beaucoup plus large que cela. Il y a toute une classe de choses appelées "U-statistiques" par exemple, qui satisfont toutes le théorème de la limite centrale et sont donc normalement distribuées pour les échantillons de grande taille. La moyenne est l'une de ces statistiques, mais ce n'est pas la seule.

pour l'erreur type, il nous dit à *quel point* une grande expérience est *beaucoup* plus fiable. Il nous dit pourquoi la distribution normale est, bien sûr, *normale*.

Dans les expériences réelles, bon nombre des choses que nous voulons mesurer sont en fait des moyennes de d'ensemble de quantités différentes (p. ex. l'intelligence « générale », telle que mesurée par le QI, est une moyenne d'un grand nombre de compétences et d'aptitudes « spécifiques »), et lorsque cela se produit, la quantité moyenne devrait suivre une distribution normale. En raison de cette loi mathématique, la distribution normale apparaît encore et encore dans les données réelles.

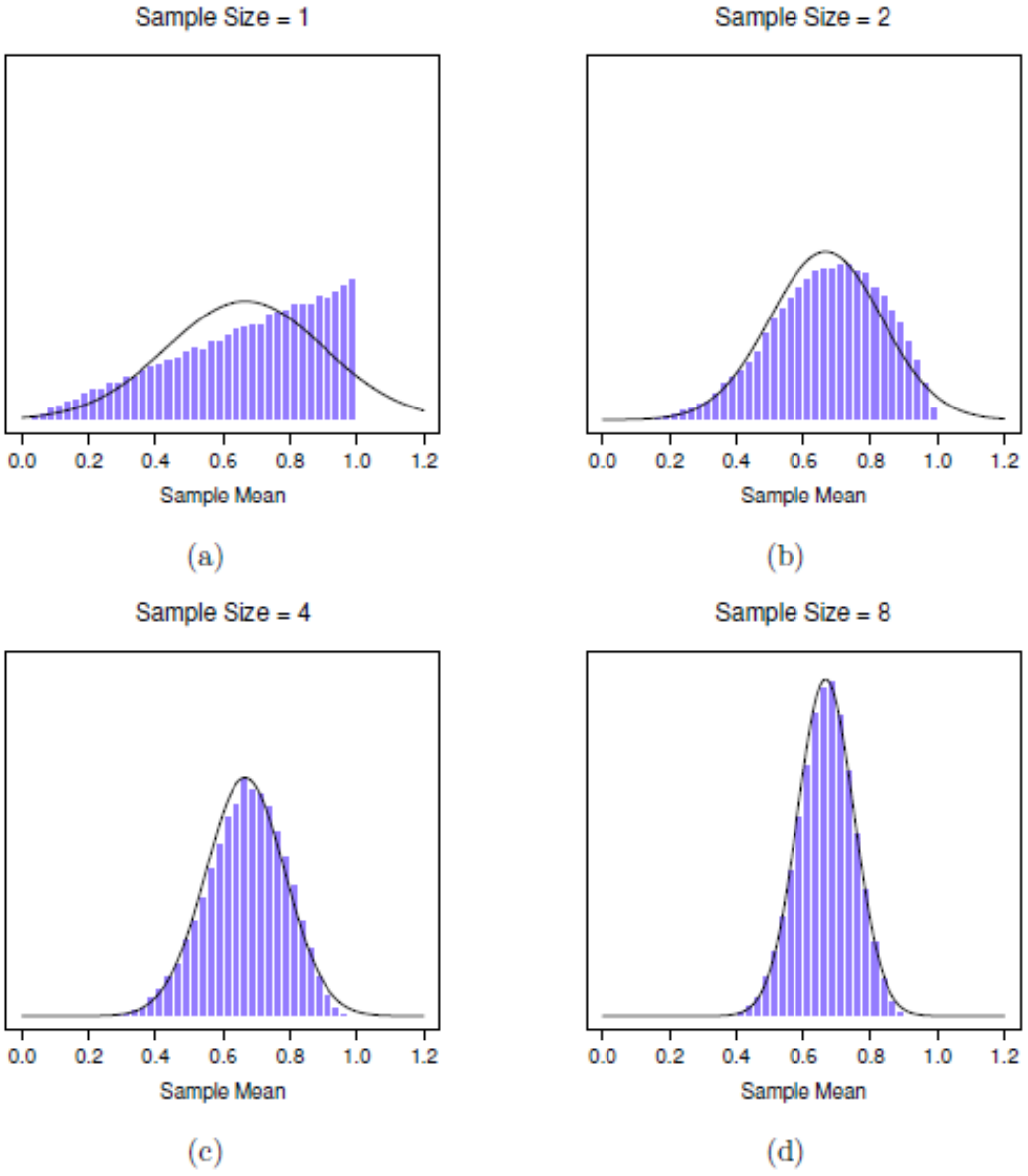


Figure 8-9 : Démonstration du théorème de la limite centrale. Dans le panel a, nous avons une distribution non normale de la population, et les panels b-d montrent la distribution d'échantillonnage de la moyenne pour les échantillons de taille 2,4 et 8 pour les données

tirées de la distribution du panel a. Comme vous pouvez le voir, même si la distribution originale de la population est non normale, la distribution d'échantillonnage de la moyenne devient assez proche de la normale lorsque vous avez un échantillon même de quatre observations.

Estimation des paramètres de population

Dans tous les exemples de QI présentés dans les sections précédentes, nous connaissions à l'avance les paramètres de la population. Comme tous les étudiants de premier cycle reçoivent leur tout premier cours sur la mesure de l'intelligence, les scores de QI sont *définis* comme ayant une moyenne de 100 et un écart-type de 15. Cependant, c'est un peu un mensonge. Comment savoir si les scores de QI ont une moyenne réelle de 100 dans la population ? Eh bien, nous le savons parce que les personnes qui ont conçu les tests les ont administrés à de très gros échantillons et qu'elles ont ensuite « truqué » les règles de notation pour que leur échantillon ait une moyenne de 100. Ce n'est pas une mauvaise chose bien sûr, c'est une partie importante de la conception d'une mesure psychologique. Cependant, il est important de garder à l'esprit que cette moyenne théorique de 100 ne s'applique qu'à la population que les concepteurs des tests ont utilisée pour concevoir les tests. Les bons concepteurs de tests s'efforceront en fait de fournir des « normes de test » qui peuvent s'appliquer à un grand nombre de populations différentes (par exemple, différents groupes d'âge, nationalités, etc.).

C'est très pratique, mais bien sûr, presque tous les projets de recherche d'intérêt impliquent l'examen d'une population de personnes différentes de celles qui sont utilisées dans les normes d'essai. Supposons, par exemple, que vous vouliez mesurer l'effet de l'intoxication saturnine à faible concentration sur le fonctionnement cognitif à Port Pirie, une ville industrielle d'Australie-Méridionale où se trouve une fonderie de plomb. Peut-être décidez-vous de comparer les scores de QI des habitants de Port Pirie à ceux d'un échantillon comparable de Whyalla, une ville industrielle d'Australie-Méridionale dotée d'une raffinerie d'acier.[48] Peu importe la ville à laquelle vous pensez, il n'est pas logique de *supposer*

---

[48] Veuillez noter que si cette question vous intéressait vraiment, vous devriez être beaucoup plus prudent que moi. Vous ne pouvez pas simplement comparer les scores de QI de Whyalla à Port Pirie et supposer que les différences sont dues au saturnisme. Même s'il est vrai que les seules différences entre les deux villes correspondaient à des raffineries différentes (et ce n'est pas le cas, loin s'en faut), il faut tenir compte du fait que les gens croient déjà que la pollution au plomb provoque des déficits cognitifs. Si vous vous rappelez le chapitre 2, cela signifie qu'il y a des effets de demande différents pour l'échantillon de Port Pirie et pour celui de Whyalla. En d'autres termes, vous pourriez vous retrouver avec une différence de groupe illusoire dans vos données, causée par le fait que les gens pensent qu'il y a une différence réelle. Je trouve peu plausible de penser que les habitants de Port Pirie ne seraient pas bien au courant de ce que vous essayez de faire si un groupe de chercheurs se présentait à Port Pirie avec des blouses de laboratoire et des tests de QI, et encore moins plausible de penser que beaucoup de gens vous en voudraient beaucoup de le faire. Ces gens ne seront pas aussi coopératifs dans les tests. D'autres personnes à Port Pirie pourraient être plus motivées à réussir parce qu'elles ne veulent pas que leur ville natale ait

simplement que la vraie population a un QI moyen de 100. Personne n'a, à ma connaissance, produit de données de normalisation sensées qui peuvent être automatiquement appliquées aux villes industrielles d'Australie-Méridionale. Nous allons devoir **estimer les** paramètres de la population à partir d'un échantillon de données. Alors, comment fait-on ça ?

Estimation de la moyenne de la population

Supposons que nous allions à Port Pirie et que 100 habitants de la région aient la gentillesse de passer un test de QI. Le QI moyen de ces personnes s'avère être $\overline{X}$=98,5. Quel est donc le vrai QI moyen pour l'ensemble de la population de Port Pirie ? De toute évidence, nous ne connaissons pas la réponse à cette question. Cela pourrait être 97,2, mais cela pourrait aussi être 103,5. Notre échantillonnage n'est pas exhaustif et nous ne pouvons donc pas donner une réponse définitive. Néanmoins, si j'étais forcé, sous la menace d'une arme à feu, de donner une « meilleure estimation », je dirais 98,5. C'est l'essence même de l'estimation statistique : donner la meilleure estimation possible.

Dans cet exemple, l'estimation du paramètre inconnu d'une population est simple. Je calcule la moyenne de l'échantillon et je m'en sers comme **estimation de la moyenne de la population**. C'est assez simple, et dans la prochaine section, je vais expliquer la justification statistique de cette réponse intuitive. Toutefois, pour l'instant, ce que je veux faire, c'est m'assurer que vous reconnaissez que la statistique de l'échantillon et l'estimation du paramètre population sont des choses conceptuellement différentes. Un échantillon statistique est une description de vos données, alors que l'estimation n'est qu'une estimation de la population. Dans cet esprit, les statisticiens utilisent souvent des notations différentes pour s'y référer. Par exemple, si la moyenne réelle de la population est notée μ, alors nous utiliserons $\hat{\mu}$ pour nous référer à notre estimation de la moyenne de la population. Par contre, la moyenne de l'échantillon est notée $\overline{X}$ ou parfois m. Cependant, dans les échantillons aléatoires simples, l'estimation de la moyenne de la population est identique à la moyenne de l'échantillon. Si j'observe une moyenne d'échantillon de $\overline{X}$ = 98,5 alors mon estimation de la moyenne de population est aussi $\hat{\mu} = 98.5$. Pour aider à garder la notation claire, voici un tableau pratique :

Symbole	Qu'est-ce que c'est ?	On sait ce que c'est ?
$\overline{X}$ | Moyenne de l'éch	antillon | Oui, calculée à	partir des données brutes
μ	Moyenne réelle de la population	Presque jamais connue avec certitude

---

mauvaise mine. Les effets de motivation qui s'appliqueraient dans Whyalla sont susceptibles d'être plus faibles, parce que les gens n'ont pas de concept « d'intoxication au minerai de fer » de la même manière qu'ils ont un concept « d'intoxication au plomb ». La psychologie est difficile.

$\hat{\mu}$	Estimation de la moyenne de la population	Oui, identique à la moyenne de l'échantillon dans les échantillons aléatoires simples

Estimation de l'écart-type de la population

Jusqu'à présent, l'estimation semble assez simple, et vous vous demandez peut-être pourquoi je vous ai forcé à lire tous ces trucs sur la théorie de l'échantillonnage. Dans le cas de la moyenne, notre estimation du paramètre de population (c.-à-d. $\hat{\mu}$) s'est avérée identique à celle de l'échantillon statistique correspondant (c.-à-d. $\overline{X}$). Cependant, ce n'est pas toujours vrai. Pour voir cela, réfléchissons à la façon de construire une **estimation de l'écart-type de la population**, que nous indiquerons à $\hat{\sigma}$. Que devons-nous utiliser comme estimation dans ce cas ? Votre première pensée pourrait être que nous pourrions faire la même chose que pour l'estimation de la moyenne, et utiliser simplement la statistique de l'échantillon comme estimation. C'est presque la bonne chose à faire, mais pas tout à fait.

Voilà pourquoi. Supposons que j'ai un échantillon qui contient une seule observation. Pour cet exemple, il est utile de considérer un échantillon où vous n'avez aucune intuition sur ce que pourraient être les vraies valeurs de la population, alors utilisons quelque chose de complètement fictif. Supposons que l'observation en question mesure la *cromulence de* mes chaussures. Il s'avère que mes chaussures ont une cromulence de 20. Voilà mon échantillon :

20

C'est un échantillon parfaitement légitime, même s'il a une taille d'échantillon de $N = 1$. Il a une moyenne d'échantillon de 20 et parce que chaque observation dans cet échantillon est égale à la moyenne de l'échantillon (évidemment !) il a un écart type d'échantillon de 0. Comme une description de l'*échantillon* cela semble tout à fait juste, l'échantillon contient une seule observation et donc aucune variation observée dans l'échantillon. Un écart-type d'échantillon de $s = 0$ est ici la bonne réponse. Mais en tant qu'estimation de l'écart-type de la *population,* cela parait complètement fou, ne croyez-vous pas ? Certes, vous et moi ne savons rien du tout de ce qu'est la « cromulence », mais nous savons quelque chose des données. La seule raison pour laquelle nous ne voyons aucune variabilité dans l'*échantillon* est que l'échantillon est trop petit pour afficher une variation ! Donc, si vous avez un échantillon de taille $N = 1$, vous avez l'*impression que* la bonne réponse est simplement de dire « aucune idée du tout ».

Remarquez que vous *n'*avez *pas la* même intuition lorsqu'il s'agit de la moyenne de l'échantillon et de la moyenne de la population. S'il est forcé de faire une meilleure estimation de la population, cela signifie qu'il n'est pas complètement insensé de deviner que la moyenne de la population est de 20. Bien sûr, vous ne vous sentiriez probablement pas très confiant dans cette supposition parce que vous n'avez qu'une seule observation sur laquelle travailler, mais c'est quand même la meilleure supposition que vous pouvez faire.

Étendons un peu cet exemple. Supposons maintenant que je fasse une deuxième observation. Mon ensemble de données contient maintenant $N = 2$ observations de la cromulence des chaussures, et l'échantillon complet ressemble maintenant à ceci :

20, 22

Cette fois-ci, notre échantillon est *juste* assez grand pour nous permettre d'observer une certaine variabilité : deux observations est le nombre minimum nécessaire pour qu'une variabilité puisse être observée ! Pour notre nouvel ensemble de données, la moyenne de l'échantillon est

$$\bar{X}$$

= 21, et l'écart-type de l'échantillon est $s = 1$. Encore une fois, pour ce qui est de la moyenne de la population, la meilleure estimation que nous puissions faire est la moyenne de l'échantillon. Si nous devions deviner, nous devinerions probablement que la cromulence moyenne de la population est de 21. Qu'en est-il de l'écart-type ? C'est un peu plus compliqué. L'écart-type de l'échantillon n'est basé que sur deux observations, et si vous êtes comme moi, vous avez probablement l'intuition que, avec seulement deux observations, nous n'avons pas donné à la population « assez de chance » pour nous révéler sa véritable variabilité. Ce n'est pas seulement que nous soupçonnons que l'estimation est *erronée*, après tout, avec seulement deux observations, nous nous attendons à ce qu'elle le soit dans une certaine mesure. L'inquiétude est que l'erreur est *systématique*. Plus précisément, nous soupçonnons que l'écart-type de l'échantillon est probablement inférieur à celui de la population.

Cette intuition semble juste, mais ce serait bien de le démontrer d'une manière ou d'une autre. Il existe en fait des preuves mathématiques qui confirment cette intuition, mais à moins d'avoir le bon bagage mathématique, elles n'aident pas beaucoup. Je vais plutôt simuler les résultats de quelques expériences. Dans cet esprit, revenons à nos études sur le QI. Supposons que le QI moyen de la population réelle est de 100 et que l'écart-type est de 15. Je vais d'abord faire une expérience dans laquelle je mesure $N = 2$ scores de QI et je vais calculer l'écart-type de l'échantillon. Si je le fais encore et encore, et que je trace un histogramme de ces écarts-types d'échantillon, ce que j'ai, c'est la *distribution d'échantillonnage de l'écart type*. J'ai tracé cette distribution dans la Figure 8-10. Même si l'écart type réel de la population est de 15, la moyenne des écarts types de l'*échantillon* n'est que de 8,5. Remarquez qu'il s'agit d'un résultat très différent de celui que nous avons obtenu à la Figure 8-8b lorsque nous avons tracé la distribution d'échantillonnage de la moyenne, où la moyenne de la population est de 100 et la moyenne des moyennes de l'échantillon est également de 100.

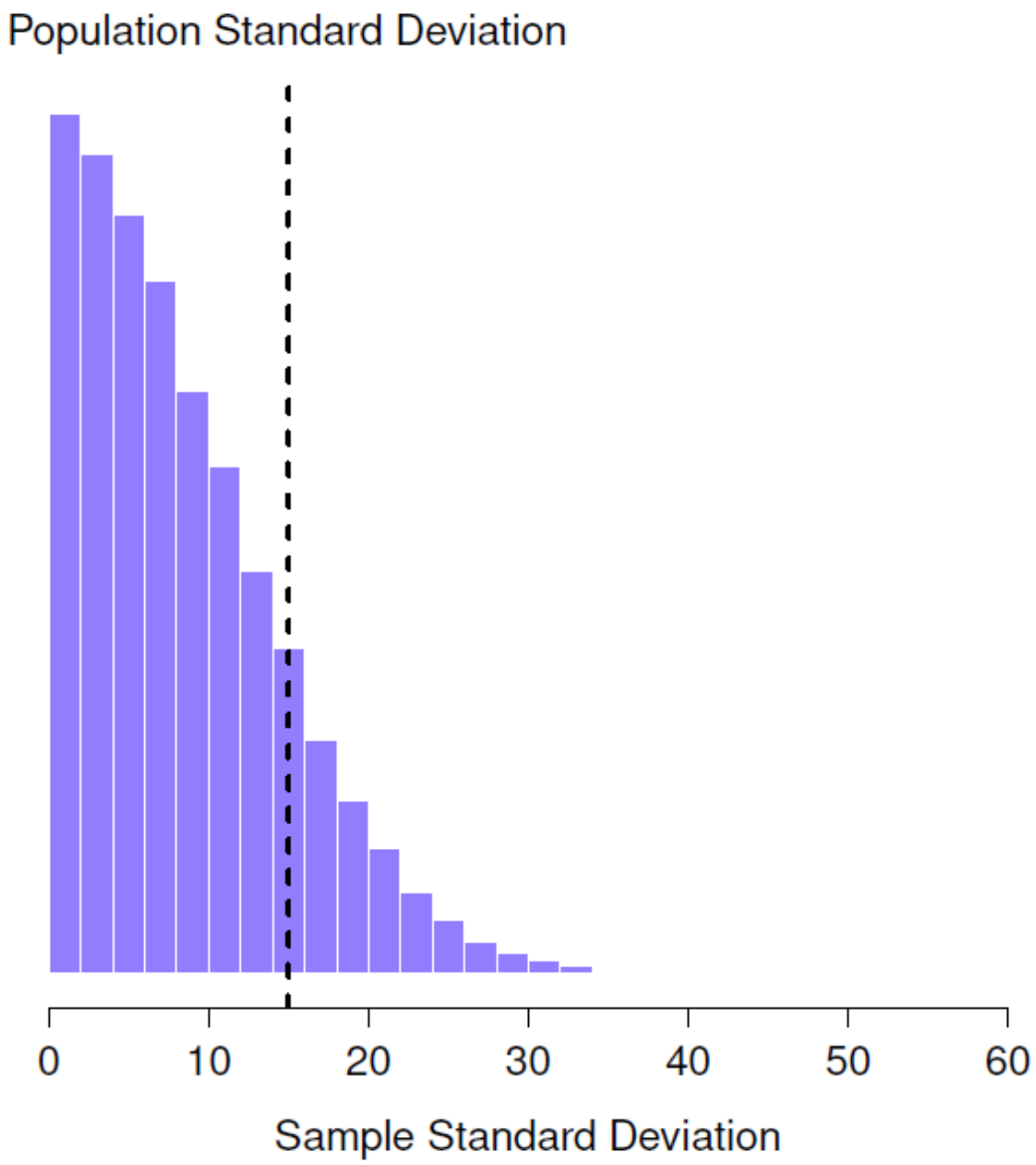


Figure 8-10 : Distribution d'échantillonnage de l'écart-type de l'échantillon pour une expérience à «deux scores de QI». L'écart-type réel de la population est de 15 (ligne pointillée), mais comme vous pouvez le voir sur l'histogramme, la grande majorité des expériences produiront un écart-type de l'échantillon beaucoup plus petit que celui-ci. En moyenne, cette expérience produirait un écart-type d'échantillon de seulement 8,5, bien en dessous de la valeur réelle ! En d'autres termes, l'écart-type de l'échantillon est une estimation biaisée de l'écart-type de la population.

Maintenant, étendons la simulation. Au lieu de nous limiter à la situation où $N = 2$, répétons l'exercice pour les tailles d'échantillon de 1 à 10. Si nous traçons la moyenne et l'écart-type moyen de l'échantillon en fonction de la taille de l'échantillon, nous obtenons les résultats présentés à la Figure 8-11. Sur le côté gauche (panneau a) j'ai tracé la moyenne des

moyennes des échantillons et sur le côté droit (panneau b) j'ai tracé l'écart type moyen. Les deux graphiques sont très différents : *en moyenne, la moyenne* de l'échantillon moyen est égale à la moyenne de la population. Il s'agit d'un **estimateur non biaisé**, ce qui explique essentiellement pourquoi votre meilleure estimation de la moyenne de la population est la moyenne de l'échantillon[49]. Le graphique de droite est très différent : en moyenne, l'écart-type *s de l'*échantillon est *inférieur* à l'écart-type de la population σ. C'est un **estimateur biaisé**. En d'autres termes, si nous voulons faire une « meilleure estimation » $\hat{\sigma}$ de la valeur de l'écart-type de la population σ nous devons nous assurer que notre estimation est un peu plus grande que l'écart-type *s de l'*échantillon.

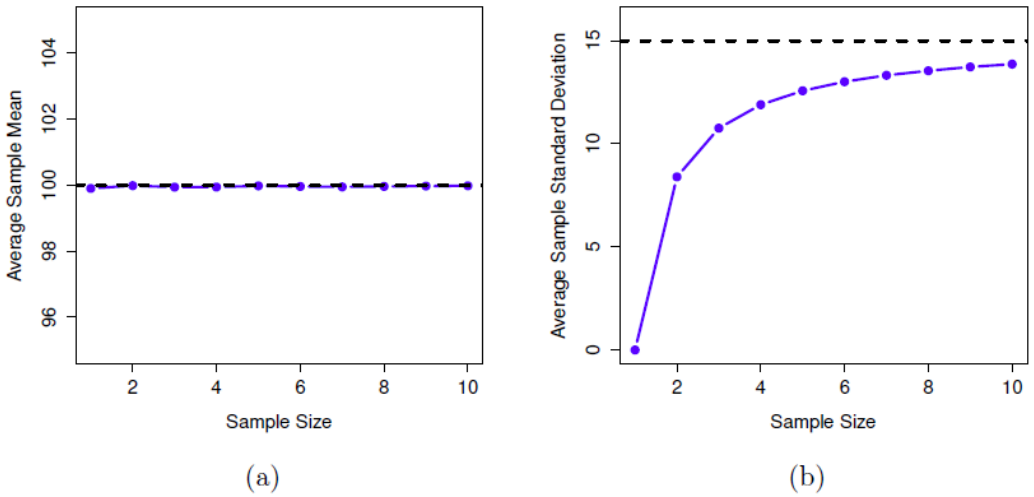


Figure 8-11 : Illustration du fait que la moyenne de l'échantillon est un estimateur non biaisé de la moyenne de la population (panel a), mais que l'écart-type de l'échantillon est biaisé (panel b). Pour la figure, j'ai généré 10 000 ensembles de données simulées avec 1 observation chacun, 10 000 autres avec 2 observations, et ainsi de suite jusqu'à une taille d'échantillon de 10. Chaque ensemble de données était constitué de fausses données sur le QI, c'est-à-dire que les données étaient normalement distribuées avec une moyenne de population réelle de 100 et un écart type de 15. En moyenne, la moyenne de l'échantillon est de 100, quelle que soit la taille de l'échantillon (panel a). Cependant, les écarts-types des échantillons s'avèrent systématiquement trop faibles (panel b), en particulier pour les petites tailles d'échantillons.

La solution à ce biais systématique s'avère très simple. Voici comment ça marche. Avant d'aborder l'écart-type, examinons la variance. Si vous vous souvenez de la section 4.2, la

---

[49] Je dois noter que je cache quelque chose ici. L'impartialité est une caractéristique souhaitable pour un estimateur, mais il y a d'autres choses qui comptent en plus du biais. Cependant, ce n'est pas l'objet de ce livre d'en discuter en détail. Je veux simplement attirer votre attention sur le fait qu'il y a là une certaine complexité cachée.

variance de l'échantillon est définie comme étant la moyenne des écarts quadratiques de la moyenne de l'échantillon. C'est à dire :

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

La variance d'échantillon s^2 est un estimateur biaisé de la variance de la population σ^2 Mais il s'avère que nous n'avons qu'à faire un tout petit ajustement pour transformer cela en un estimateur non biaisé. Tout ce que nous avons à faire est de diviser par *N-1* plutôt que par *N*. Si nous faisons cela, nous obtenons la formule suivante :

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

Il s'agit d'un estimateur non biaisé de la variance de la population σ. En outre, cela répond enfin à la question que nous avons soulevée à la section 4.2. Pourquoi Jamovi nous a-t-il donné des réponses légèrement différentes pour la variance ? C'est parce que Jamovi calcule

$$\hat{\sigma}^2$$

pas *s2*, voilà pourquoi. Il en va de même pour l'écart-type. Si nous divisons par *N* - 1 au lieu de *N*, notre estimation de l'écart-type de la population devient :

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

et lorsque nous utilisons la fonction d'écart-type intégrée de Jamovi, ce qu'il fait est de calculer $\hat{\sigma}$, pas *s*. *a*[50]

---

[50] D'accord, je cache quelque chose d'autre ici. De façon bizarre et contre-intuitive, puisque $\hat{\sigma}^2$ est un estimateur non biaisé de σ^2, on pourrait supposer que prendre la racine carrée serait bien et que $\hat{\sigma}$ serait un estimateur non biaisé de σ. Bizarrement, ça ne l'est pas. Il y a en fait un biais subtil et minuscule dans $\hat{\sigma}$. C'est tout simplement bizarre : $\hat{\sigma}^2$ est une estimation non biaisée de la variance de la population σ^2, mais lorsque vous prenez la racine carrée, il s'avère que $\hat{\sigma}$. est un estimateur biaisé de l'écart-type de la population σ. Bizarre, bizarre, bizarre, pas vrai ? Alors, pourquoi $\hat{\sigma}$. est-il biaisé ? La réponse technique est « parce que les transformations non linéaires (par exemple, la racine carrée) ne correspondent pas aux attentes », mais cela ressemble à du charabia pour tous ceux qui n'ont pas suivi de cours en statistique mathématique. Heureusement, cela n'a pas d'importance en pratique. Le biais est faible, et dans la vraie vie tout le monde utilise $\hat{\sigma}$. et ça marche très bien. Parfois, les mathématiques sont tout simplement ennuyeuses.

Un dernier point. Dans la pratique, beaucoup de gens ont tendance à se référer à $\hat{\sigma}$. (c.-à-d. la formule où nous divisons par *N*-1) comme écart-type de l'*échantillon.* Techniquement, c'est incorrect. L'écart-type de l'*échantillon* devrait être égal à *s* (c.-à-d. la formule où nous divisons par *N*). Ce n'est pas la même chose, que ce soit sur le plan conceptuel ou numérique. L'une est une propriété de l'échantillon, l'autre est une caractéristique estimée de la population. Cependant, dans presque toutes les applications de la vie réelle, ce qui nous préoccupe réellement, c'est l'estimation du paramètre de population, et donc les gens rapportent toujours $\hat{\sigma}$ plutôt que *s*. C'est le bon chiffre à rapporter, bien sûr. C'est juste que les gens ont tendance à être un peu imprécis au sujet de la terminologie lorsqu'ils la rédigent, parce que « l'écart-type de l'échantillon » est plus court que « l'écart-type estimé de la population ». Ce n'est pas grand-chose, et en pratique, je fais la même chose que tout le monde. Néanmoins, je pense qu'il est important de garder les deux *concepts* séparés. Ce n'est jamais une bonne idée de confondre les « propriétés connues de votre échantillon » avec les « suppositions sur la population dont il provient ». Dès que vous commencez à penser que *s* et $\hat{\sigma}$ sont la même chose, vous commencez à faire exactement cela.

Pour terminer cette section, voici quelques autres tableaux qui vous aideront à clarifier les choses.

Symbole	Qu'est-ce que c'est ?	On sait ce que c'est ?
s	Écart-type de l'échantillon	Oui, calculé à partir des données brutes
σ	Écart-type de la population	Presque jamais connu avec certitude
$\hat{\sigma}$	Estimation de l'écart-type de la population	Oui, mais ce n'est pas la même chose que l'écart-type de l'échantillon

Symbole	Qu'est-ce que c'est ?	On sait ce que c'est ?
s^2	Variance de l'échantillon	Oui, calculé à partir des données brutes
σ^2	Variation de la population	Presque jamais connu avec certitude
$\hat{\sigma}^2$	Estimation de la variation de la population	Oui, mais ce n'est pas la même chose que la variance de l'échantillon

Estimation d'un intervalle de confiance

Les statistiques, c'est de ne jamais avoir à dire qu'on est certain - Origine inconnue[51]

[51] Cette citation apparaît sur un grand nombre de t-shirts et de sites web, et est même mentionnée dans quelques articles académiques.(voir http://www. mais je n'ai jamais trouvé la source originale.

Jusqu'à présent, dans ce chapitre, j'ai exposé les principes de base de la théorie de l'échantillonnage sur lesquels se fondent les statisticiens pour deviner les paramètres de la population à partir d'un échantillon de données. Comme l'illustre cette discussion, l'une des raisons pour lesquelles nous avons besoin de toute cette théorie de l'échantillonnage est que chaque ensemble de données nous laisse avec une certaine incertitude, de sorte que nos estimations ne seront jamais parfaitement exactes. Ce qui manque dans cette discussion, c'est une tentative de *quantifier* le degré d'incertitude qui s'attache à notre estimation. Il ne suffit pas de deviner que, disons, le QI moyen des étudiants en psychologie de premier cycle est de 115 (oui, je viens d'inventer ce chiffre). Nous voulons aussi pouvoir dire quelque chose qui exprime le degré de certitude que nous avons à son propos. Par exemple, il serait bien de pouvoir dire qu'il y a 95% de chances que la vraie moyenne se situe entre 109 et 121. Le nom pour ceci est un **intervalle de confiance** pour la moyenne.

Armé d'une compréhension des distributions d'échantillonnage, l'établissement d'un intervalle de confiance pour la moyenne est en fait assez facile. Voici comment ça marche. Supposons que la vraie moyenne de la population est μ et que l'écart-type est σ. Je viens de terminer mon étude qui a N participants, et le QI moyen parmi ces participants est $\overline{X}$. Notre analyse du théorème de la limite centrale (section 8.3.3) nous a appris que la distribution d'échantillonnage de la moyenne est approximativement normale. Nous savons également, d'après notre analyse de la distribution normale (section 7.5), qu'il y a 95 % de chances qu'une quantité normalement distribuée se situe à l'intérieur d'environ deux écarts-types de la moyenne réelle.

Pour être plus précis, la réponse la plus correcte est qu'il y a 95 % de chances qu'une quantité normalement distribuée se situe à l'intérieur de 1,96 écart type de la moyenne réelle. Ensuite, rappelez-vous que l'écart-type de la distribution d'échantillonnage est désigné sous le nom d'erreur-type, et que l'erreur-type de la moyenne est écrite sous le nom de SEM. Lorsque nous mettons tous ces éléments ensemble, nous apprenons qu'il y a une probabilité de 95 % que la moyenne de l'échantillon $\overline{X}$ que nous avons effectivement observée se situe à l'intérieur de 1,96 erreur type de la moyenne de la population.

Mathématiquement, nous écrivons ceci comme :

$$\mu - (1{,}96 * SEM) \leq \overline{X} \leq \mu + (1{,}96 * SEM)$$

où le SEM est égal à $\sigma/\sqrt{N}$ et nous pouvons être sûrs à 95% que c'est vrai. Cependant, cela ne répond pas à la question qui nous intéresse. L'équation ci-dessus nous indique ce à quoi nous devons nous attendre au sujet de la moyenne de l'échantillon étant donné que nous connaissons les paramètres de la population. Ce que nous *voulons,* c'est que ce travail se fasse dans l'autre sens. Nous voulons savoir ce que nous devons croire des paramètres de la population, étant donné que nous avons observé un échantillon particulier. Cependant, ce n'est pas trop difficile à faire. En utilisant un peu d'algèbre de lycée, une façon sournoise de réécrire notre équation est comme ceci :

$$\overline{X} - (1{,}96 * SEM) \leq \mu\overline{X} + (1{,}96 * SEM)$$

Ce qui est révélateur, c'est que la plage de valeurs a une probabilité de 95 % de contenir la moyenne de la population μ. Nous appelons cette plage un **intervalle de confiance à 95 %**,

appelé CI95. Bref, tant que *N* est suffisamment grand (assez grand pour que l'on croie que la distribution d'échantillonnage de la moyenne est normale), nous pouvons écrire cette formule comme étant notre formule pour l'intervalle de confiance à 95 % :

$$\text{CI}_{95} = \overline{X} \pm \left(1{,}96 \times \frac{\sigma}{\sqrt{N}}\right)$$

Bien sûr, il n'y a rien de spécial avec le chiffre 1,96. C'est tout simplement le multiplicateur qu'il vous faut utiliser si vous voulez un intervalle de confiance à 95 %. Si j'avais voulu un intervalle de confiance de 70%, j'aurais utilisé 1,04 comme chiffre magique plutôt que 1,96.

Une légère erreur dans la formule

Comme d'habitude, j'ai menti. La formule que j'ai donnée ci-dessus pour l'intervalle de confiance à 95 % est à peu près exacte, mais j'ai passé sous silence un détail important de la discussion. Notez que ma formule exige que vous utilisiez l'erreur type de la moyenne, SEM, ce qui vous oblige à utiliser l'écart-type de la population réelle σ. Pourtant, à la section 8.4, j'ai souligné le fait que nous ne *connaissons* pas réellement les vrais paramètres de population. Comme nous ne connaissons pas la valeur réelle de σ, nous devons plutôt utiliser une estimation de l'écart-type de la population ($\hat{\sigma}$). C'est assez simple à faire, mais cela a pour conséquence que nous devons utiliser les percentiles de la *distribution t* plutôt que la distribution normale pour calculer notre nombre magique, et la réponse dépend de la taille de l'échantillon. Quand *N* est très grand, on obtient à peu près la même valeur en utilisant la *distribution t* ou la distribution normale : 1,96. Mais lorsque *N* est petit, nous obtenons un nombre beaucoup plus grand lorsque nous utilisons la distribution *t* : 2,26.

Il n'y a rien de trop mystérieux dans ce qui se passe ici. Des valeurs plus élevées signifient que l'intervalle de confiance est plus large, ce qui indique que nous sommes plus incertains quant à la valeur réelle de μ. Lorsque nous utilisons la distribution *t au* lieu de la distribution normale, nous obtenons des nombres plus grands, ce qui indique que nous avons plus d'incertitude. Et pourquoi avons-nous cette incertitude supplémentaire ? Eh bien, parce que notre estimation de l'écart-type de la population $\hat{\sigma}$ pourrait être fausse ! Si c'est faux, cela signifie que nous sommes un peu moins sûrs de ce à quoi ressemble réellement notre distribution d'échantillonnage de la moyenne, et cette incertitude finit par se refléter dans un intervalle de confiance plus large.

Interpréter un intervalle de confiance

Le plus difficile dans les intervalles de confiance, c'est de comprendre ce qu'ils *signifient*. Chaque fois que les gens rencontrent pour la première fois des intervalles de confiance, le premier instinct est presque toujours de dire « qu'il y a une probabilité de 95% que la vraie moyenne se trouve à l'intérieur de l'intervalle de confiance ». C'est simple et cela semble saisir l'idée de bon sens de ce que cela signifie de dire que je suis « confiant à 95% ». Malheureusement, ce n'est pas tout à fait juste. La définition intuitive repose en grande partie sur vos *croyances* personnelles quant à la valeur de la moyenne de la population. Je dis que je suis confiant à 95 p. 100 parce que ce sont mes croyances. Dans la vie de tous les jours, c'est tout à fait normal, mais si vous vous souvenez de la section 7.2, vous remarquerez que parler de croyances et de confiance personnelles est une idée bayésienne.

Cependant, les intervalles de confiance *ne* sont *pas* des outils bayésiens. Comme tout le reste dans ce chapitre, les intervalles de confiance sont des outils *fréquentistes*, et si vous utilisez des méthodes fréquentistes, il n'est pas approprié de leur attacher une interprétation bayésienne. Si vous utilisez des méthodes fréquentistes, vous devez adopter des interprétations fréquentistes !

Bien, donc si ce n'est pas la bonne réponse, qu'est-ce que c'est ? Souvenez-vous de ce qu'on a dit à propos de la probabilité fréquentiste. La seule façon de parler de « probabilité » est de parler d'une séquence d'événements et de compter les fréquences des différents types d'événements. De ce point de vue, l'interprétation d'un intervalle de confiance à 95 % doit avoir quelque chose à voir avec la réplication. Plus précisément, si nous répétions l'expérience à maintes reprises et calculions un intervalle de confiance de 95 % pour chaque répétition, alors 95 % de ces *intervalles* contiendraient la vraie moyenne. De façon plus générale, 95 % de tous les intervalles de confiance construits à l'aide de cette procédure devraient contenir la moyenne réelle de la population. Cette idée est illustrée à la Figure 8-12, qui montre 50 intervalles de confiance construits pour une expérience de « mesure de 10 QI » (figure supérieure) et 50 autres intervalles de confiance pour une expérience de « mesure de 25 QI » (figure inférieure). Un peu fortuitement, sur les 100 répétitions que j'ai simulées, il s'est avéré que 95 d'entre elles exactement contenaient la vraie moyenne.

La différence critique ici est que la revendication bayésienne parle de probabilité au sujet de la moyenne de la population (c.-à-d. qu'elle fait référence à notre incertitude au sujet de la moyenne de la population), ce qui n'est pas permis selon l'interprétation fréquentiste de la probabilité parce que vous ne pouvez « reproduire » une population ! Dans la le pont de vue fréquentiste, la moyenne de population est fixe et aucune affirmation probabiliste ne peut être faite à ce sujet. Les intervalles de confiance, cependant, sont reproductibles pour que nous puissions répéter les expériences. Par conséquent, un fréquentiste est autorisé à parler de la probabilité de l'*intervalle de confiance* contienne la vraie moyenne, mais il n'est pas permis de parler de la probabilité que la *vraie moyenne de* la *population* (qui n'est pas un événement répétable) se situe dans l'intervalle de confiance.

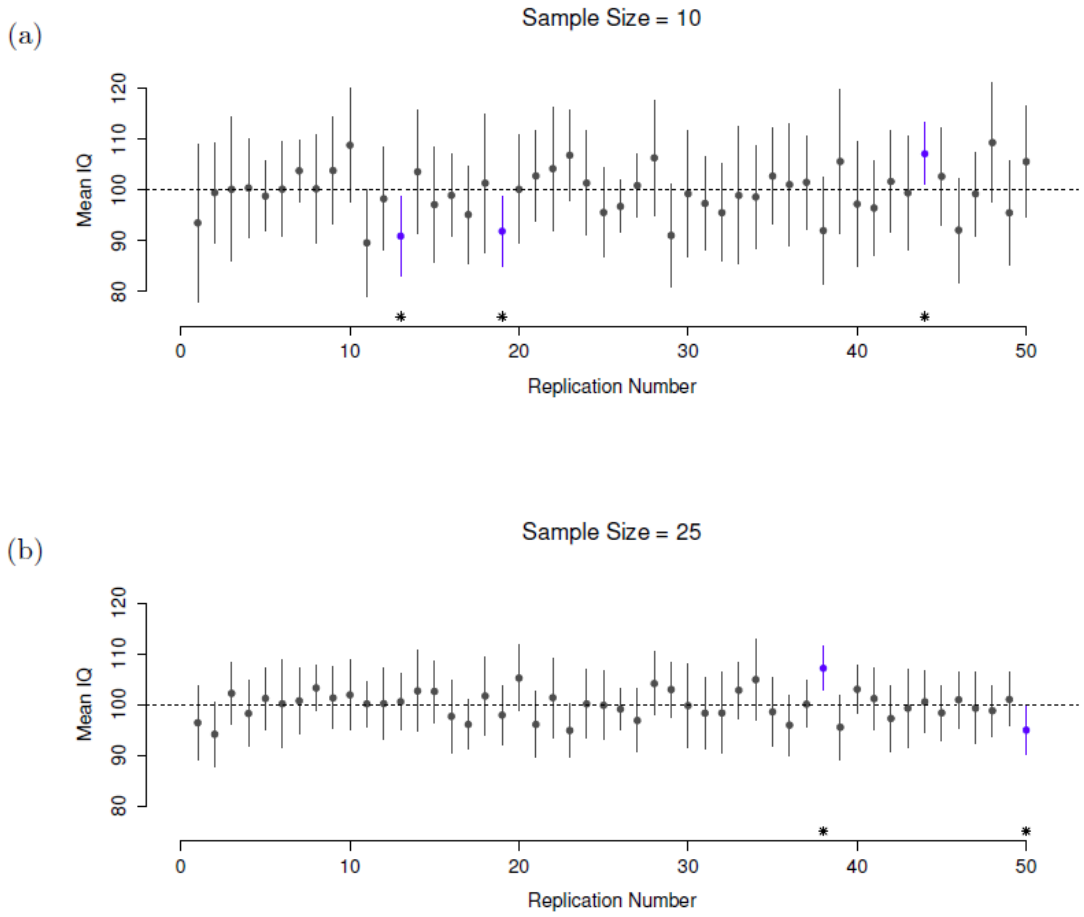


Figure 8-12 Intervalles de confiance à 95 %. La partie supérieure (panneau a) montre 50 répétitions simulées d'une expérience dans laquelle nous mesurons les QI de 10 personnes. Le point marque l'emplacement de la moyenne de l'échantillon et la ligne indique l'intervalle de confiance à 95 %. Au total, 47 des 50 intervalles de confiance contiennent la vraie moyenne (c.-à-d. 100), mais pas les trois intervalles marqués d'un astérisque. Le graphique du bas (panneau b) montre une simulation similaire, mais cette fois nous simulons des réplications d'une expérience qui mesure les QI de 25 personnes.

Je sais que cela semble un peu pédant, mais c'est important. C'est important parce que la différence d'interprétation conduit à une différence mathématique. Il existe une alternative bayésienne aux intervalles de confiance, appelés *intervalles crédibles*. Dans la plupart des cas, les intervalles crédibles sont assez semblables aux intervalles de confiance, mais dans d'autres cas, ils sont très différents. Comme promis, cependant, je parlerai plus en détail de la perspective bayésienne au chapitre 15.

Calcul des intervalles de confiance dans Jamovi

Pour autant que je puisse le dire, Jamovi n'inclut pas (encore) de méthode simple pour calculer les intervalles de confiance de la moyenne dans le cadre de la fonctionnalité

« Descriptives ». Mais les « Descriptives » ont une case à cocher pour la SEM., vous pouvez donc l'utiliser pour calculer l'intervalle de confiance inférieur de 95% comme :

Moyenne - (1,96 * SEM),

et l'intervalle de confiance supérieur à 95 % comme :

Moyenne + (1,96 * SEM)

Les intervalles de confiance à 95 % sont la norme de facto en psychologie. Ainsi, par exemple, si je charge le fichier IQsim.omv, vérifie la moyenne et SEM sous « Descriptives », je peux calculer l'intervalle de confiance associé au QI moyen simulé :

IC 95 % inférieur = 99,68 - (1,96 * 0,15) = 99,39 IC supérieur 95 % = 99,68 + (1,96 * 0,15) = 99,98

Ainsi, dans nos données simulées sur un grand échantillon avec N = 10 000, le QI moyen est de 99,68 avec un IC à 95 % de 99,39 à 99,98. J'espère que c'est assez clair. Ainsi, bien qu'il n'existe pas actuellement de moyen simple d'obtenir de Jamovi qu'il calcule l'intervalle de confiance dans le cadre des options de la variable « Descriptives », si nous le voulions, nous pourrions assez facilement le calculer à la main.

De même, lorsqu'il s'agit de tracer les intervalles de confiance dans Jamovi, cela n'est pas (encore) disponible dans le cadre des options « Descriptives ». Cependant, lorsque nous nous familiariserons avec des tests statistiques spécifiques, par exemple au chapitre 13, nous verrons que nous pouvons tracer des intervalles de confiance dans le cadre de l'analyse des données. C'est plutôt cool, on vous montrera alors comment faire cela plus tard.

Résumé

Dans ce chapitre, j'ai couvert deux sujets principaux. La première moitié du chapitre traite de la théorie de l'échantillonnage, et la seconde de la façon dont nous pouvons utiliser la théorie de l'échantillonnage pour construire des estimations des paramètres de la population. La répartition des sections ressemble à ceci :

- Idées de base sur les échantillons, l'échantillonnage et les populations (Section 8.1)
- Théorie statistique de l'échantillonnage : la loi des grands nombres (section 8.2), les distributions d'échantillonnage et le théorème de la limite centrale (section 8.3).
- Estimation des moyennes et des écarts-types (section 8.4)
- Estimation d'un intervalle de confiance (section 8.5)

Comme toujours, il y a beaucoup de sujets liés à l'échantillonnage et à l'estimation qui ne sont pas abordés dans ce chapitre, mais pour un cours d'introduction à la psychologie, c'est assez complet je pense. Pour la plupart des chercheurs appliqués, vous n'aurez pas besoin de beaucoup plus de théorie que cela. Une grande question que je n'ai pas abordée dans ce chapitre est ce que vous faites quand vous n'avez pas un échantillon aléatoire simple. Il y a beaucoup de théorie statistique sur laquelle on peut s'appuyer pour gérer cette situation, mais cela dépasse largement le cadre de ce livre.

Tests d'hypothèses

Le processus d'induction est le processus d'adoption de la loi la plus simple qui peut être faite pour s'harmoniser avec notre expérience. Cependant, ce processus n'a pas de fondement logique, mais seulement psychologique. Il est clair qu'il n'y a aucune raison de croire que le cours des événements le plus simple se produira réellement. C'est une hypothèse que le soleil se lèvera demain : et cela signifie que nous ne savons pas s'il se lèvera. - Ludwig Wittgenstein[52]

Dans le dernier chapitre, j'ai discuté des notions sous-jacentes à l'estimation, qui est l'une des deux « grandes notions » dans les statistiques inférentielles. Il est maintenant temps de nous pencher sur l'autre grande idée, à savoir la *vérification des hypothèses*. Dans sa forme la plus abstraite, la vérification d'hypothèses est vraiment une idée très simple. Le chercheur a une théorie sur le monde et veut déterminer si les données appuient ou non cette théorie. Cependant, les détails sont compliqués et la plupart des gens trouvent que la théorie de la vérification des hypothèses est la partie la plus frustrante des statistiques. La structure du chapitre est la suivante. Tout d'abord, je décrirai en détail le fonctionnement des tests d'hypothèses à l'aide d'un simple exemple de fonctionnement pour vous montrer comment un test d'hypothèse est « construit ». J'essaierai d'éviter d'être trop dogmatique en le faisant, et je me concentrerai plutôt sur la logique sous-jacente de la procédure de test.[53] Ensuite, je passerai un peu de temps à parler des différents dogmes, règles et hérésies qui entourent la théorie de la vérification des hypothèses.

Une ménagerie d'hypothèses

Finalement, nous finissons tous par succomber à la folie. Pour moi, ce jour arrivera quand je serai enfin promu professeur titulaire. En sécurité dans ma tour d'ivoire, heureusement protégée par la tenure, je pourrai enfin m'affranchir de mes sens (pour ainsi dire) et m'adonner à cette ligne de recherche psychologique la plus improductive, la recherche de la perception extrasensorielle (PES).[54]

---

[52] La citation provient du texte de Wittgenstein (1922), Tractatus Logico-Philosphicus.

[53] Une note technique. La description ci-dessous diffère subtilement de la description standard donnée dans un grand nombre de textes d'introduction. La théorie orthodoxe de la vérification des hypothèses nulles a émergé des travaux de Sir Ronald Fisher et Jerzy Neyman au début du 20e siècle ; mais Fisher et Neyman avaient en fait des points de vue très différents sur la façon dont elle devrait fonctionner. Le traitement standard des tests d'hypothèse que la plupart des textes utilisent est un hybride des deux approches. Le traitement est ici un peu plus proche de celui de Neyman que de l'orthodoxie, surtout en ce qui concerne la signification de la valeur p.

[54] Je présente mes excuses à tous ceux qui croient vraiment à ce genre de choses, mais d'après ce que j'ai lu dans la documentation sur la PES, il n'est tout simplement pas raisonnable de penser que c'est réel. Pour être juste, cependant, certaines des études sont rigoureusement conçues, alors c'est en fait un domaine intéressant pour réfléchir à la conception de la recherche en psychologie. Et bien sûr, notre pays est assez libre pour que

Supposons que ce jour glorieux soit arrivé. Ma première étude est une étude simple dans laquelle je cherche à tester si la clairvoyance existe. Chaque participant s'assoit à une table et se voit remettre une carte par un expérimentateur. La carte est noire d'un côté et blanche de l'autre. L'expérimentateur enlève la carte et la dépose sur une table dans une pièce adjacente. La carte est placée côté noir vers le haut ou côté blanc vers le haut complètement au hasard, la randomisation n'ayant lieu qu'après que l'expérimentateur ait quitté la salle avec le participant. Un deuxième expérimentateur arrive et demande au participant de quel côté de la carte est maintenant tournée vers le haut. C'est une expérience ponctuelle. Chaque personne ne voit qu'une seule carte et ne donne qu'une seule réponse, et à aucun moment le participant n'est réellement en contact avec quelqu'un qui connaît la bonne réponse. Mon ensemble de données est donc très simple. J'ai posé la question à N personnes et un certain nombre X d'entre elles ont donné la bonne réponse. Pour rendre les choses concrètes, supposons que j'ai testé $N = 100$ personnes et que $X = 62$ d'entre elles ont obtenu la bonne réponse. Un nombre étonnamment élevé, certes, mais est-il assez important pour que je me sente en sécurité en affirmant que j'ai trouvé des preuves pour l'ESP ? C'est dans ce cas qu'il est utile de vérifier les hypothèses. Cependant, avant de parler de la façon de *vérifier* les hypothèses, nous devons être clairs sur ce que nous entendons par hypothèses.

Hypothèses de recherche versus hypothèses statistiques

La première distinction que vous devez garder à l'esprit se situe entre les hypothèses de recherche et les hypothèses statistiques. Dans mon étude PES, mon objectif scientifique global est de démontrer que la clairvoyance existe. Dans cette situation, j'ai un objectif de recherche clair : j'espère découvrir des preuves de la PES. Dans d'autres situations, je pourrais même être beaucoup plus neutre que cela, je pourrais dire alors que mon but de recherche est de déterminer si la clairvoyance existe ou non. Peu importe la façon dont je veux me présenter, le point fondamental que j'essaie de faire valoir ici, c'est qu'une hypothèse de recherche implique la formulation d'une affirmation scientifique valable et vérifiable. Si vous êtes psychologue, vos hypothèses de recherche portent essentiellement sur les *constructions psychologiques*. L'une ou l'autre des **hypothèses** suivantes pourrait être considérée comme une **hypothèse de recherche** :

- *Écouter de la musique réduit votre capacité à prêter attention à d'autres choses.* Il s'agit d'une affirmation sur la relation causale entre deux concepts psychologiquement significatifs (écouter de la musique et faire attention aux choses), c'est donc une hypothèse de recherche parfaitement raisonnable.
- *L'intelligence est liée à la personnalité*. Comme la dernière, il s'agit d'une affirmation relationnelle sur deux constructions psychologiques (intelligence et personnalité), mais l'affirmation est plus faible : corrélationnelle et non causale.
- *L'intelligence est la vitesse de traitement de l'information*. Cette hypothèse a un tout autre caractère. Ce n'est pas du tout une affirmation relationnelle. C'est une affirmation

---

vous puissiez passer votre temps et vos efforts à me prouver que j'ai tort si vous le souhaitez, mais je ne pense pas que ce soit une utilisation terriblement pragmatique de votre intellect.

ontologique sur le caractère fondamental de l'intelligence (et je suis presque sûr que c'est faux). Cela vaut la peine de s'étendre sur ce point en fait. Il est généralement plus facile de penser à la façon de construire des expériences pour vérifier des hypothèses de recherche de la forme « est-ce que X affecte Y ? » que pour répondre à des affirmations comme « qu'est-ce que X ? » Et dans la pratique, ce qui se passe habituellement, c'est que vous trouvez des moyens de tester les affirmations relationnelles qui découlent de vos affirmations ontologiques. Par exemple, si je crois que l'intelligence *est la* vitesse de traitement de l'information dans le cerveau, mes expériences consisteront souvent à rechercher des relations entre les mesures de l'intelligence et celles de la vitesse de traitement. Par conséquent, la plupart des questions quotidiennes de recherche ont tendance à être de nature relationnelle, mais elles sont presque toujours motivées par des questions ontologiques plus profondes sur l'état de la nature.

Notez qu'en pratique, mes hypothèses de recherche pourraient se chevaucher beaucoup. Mon but ultime dans l'expérience de PES pourrait être de tester une revendication ontologique comme « la PES existe », mais je pourrais me limiter opérationnellement à une hypothèse plus étroite comme « Certaines personnes peuvent « voir » des objets d'une manière clairvoyante ». Cela dit, il y a certaines choses qui ne comptent pas vraiment comme des hypothèses de recherche appropriées et ayant un sens :

- *L'amour est un champ de bataille*. C'est trop vague pour être testable. Bien qu'il soit acceptable qu'une hypothèse de recherche ait un certain degré d'imprécision, il doit être possible d'opérationnaliser vos idées théoriques. Je ne suis peut-être pas assez créatif pour le voir, mais je ne vois pas comment cela peut être converti en un plan de recherche concret. Si c'est vrai, ce n'est pas une hypothèse de recherche scientifique, c'est une chanson pop. Ça ne veut pas dire que ce n'est pas intéressant. Beaucoup de questions profondes que se posent les humains entrent dans cette catégorie. Peut-être qu'un jour la science sera capable de construire des théories vérifiables de l'amour, ou de tester si Dieu existe, et ainsi de suite. Mais pour l'instant, nous ne pouvons pas, et je ne parierais pas sur une approche scientifique satisfaisante de l'un ou de l'autre.
- *La première règle du club de tautologie est la première règle du club de tautologie*. Il ne s'agit pas d'une affirmation de fond de quelque nature que ce soit. C'est vrai par définition. Aucun état de nature concevable ne pourrait être incompatible avec cette affirmation. Nous disons qu'il s'agit d'une hypothèse infalsifiable et qu'en tant que telle, elle ne relève pas du domaine de la science. Quoi que vous fassiez d'autre en science, vos affirmations doivent avoir la possibilité d'être réfutées.
- *Plus de gens dans mon expérience diront « oui » que « non »*. Celle-ci ne constitue pas une hypothèse de recherche parce qu'il s'agit d'une affirmation au sujet de l'ensemble des données, et non au sujet de la psychologie (à moins, bien sûr, que votre vraie question de recherche soit de savoir si les gens ont une sorte de biais du « oui »). En fait, cette hypothèse commence à ressembler davantage à une hypothèse statistique qu'à une hypothèse de recherche.

Comme vous pouvez le constater, les hypothèses de recherche peuvent parfois être quelque peu confuses et, en fin de compte, ce sont des affirmations *scientifiques*. Les **hypothèses statistiques** ne sont ni l'une ni l'autre de ces choses. Les hypothèses statistiques doivent

être mathématiquement précises et correspondre à des affirmations précises sur les caractéristiques du mécanisme de production des données (c.-à-d. la « population »). Malgré tout, l'intention est que les hypothèses statistiques aient une relation claire avec les hypothèses de recherche qui vous tiennent à cœur ! Par exemple, dans mon étude sur la PES, mon hypothèse de recherche est que certaines personnes sont capables de voir à travers les murs ou autre. Ce que je veux faire, c'est « mapper » ceci sur une affirmation sur la façon dont les données ont été générées. Réfléchissons donc à ce que pourrait être cette déclaration. La quantité qui m'intéresse dans l'expérience est *P(correct)*, la probabilité vraie mais inconnue avec laquelle les participants à mon expérience répondent correctement à la question. Utilisons la lettre grecque θ (thêta) pour faire référence à cette probabilité. Voici quatre hypothèses statistiques différentes :

- Si la PES n'existe pas et si mon expérience est bien conçue, mes participants ne font que deviner. Je m'attends donc à ce qu'ils réussissent la moitié du temps et donc mon hypothèse statistique est que la vraie probabilité de choisir correctement est $\theta = 0{,}5$.
- Supposons également qu'il existe une PES et que les participants puissent voir la carte. Si c'est vrai que les gens feront mieux que le hasard et l'hypothèse statistique est que $\theta > 0{,}5$.
- Une troisième possibilité est que la PES existe, mais les couleurs sont toutes inversées et les gens ne s'en rendent pas compte (ok, c'est dingue, mais on ne sait jamais). Si c'est comme ça que ça marche, on s'attendrait à ce que la performance des gens soit *inférieure au* hasard. Cela correspondrait à une hypothèse statistique selon laquelle $\theta < 0.5$
- Enfin, supposons que la PES existe mais que je ne sais pas si les gens voient la bonne ou la mauvaise couleur. Dans ce cas, la seule affirmation que je pourrais faire au sujet des données serait que la probabilité de faire la bonne réponse *n'est pas* égale à 0,5. Ceci correspond à l'hypothèse statistique que $\theta \neq 0.5$.

Ce sont tous des exemples légitimes d'hypothèses statistiques parce qu'il s'agit d'énoncés concernant un paramètre de population et qu'ils sont liés de façon significative à mon expérience.

Ce qui ressort clairement de cette discussion, je l'espère, c'est que lorsqu'on tente d'élaborer un test d'hypothèse statistique, le chercheur a en fait deux hypothèses bien distinctes à considérer. D'abord, il a une hypothèse de recherche (une affirmation sur la psychologie), et cela correspond ensuite à une hypothèse statistique (une affirmation sur la population génératrice de données). Dans mon exemple de PES, il pourrait s'agir de :

Hypothèse de **recherche de** Dani : « La PES existe »
Hypothèse **statistique** de Dani : $\theta \neq 0.5$

Le point clé à comprendre est celui-ci. *Un test d'hypothèse statistique est un test de l'hypothèse statistique et non de l'hypothèse de recherche*. Si votre étude est mal conçue, le lien entre votre hypothèse de recherche et votre hypothèse statistique est rompu. Pour donner un exemple stupide, supposons que mon étude sur la PES ait été menée dans une situation où le participant peut réellement voir la carte se refléter dans une fenêtre. Si cela

se produit, je serais en mesure de trouver des preuves très solides que $\theta \neq 0.5$, mais cela ne nous dirait rien sur la question de savoir si « la PES existe ».

Hypothèses nulles et hypothèses alternatives

Pour l'instant, tout va bien. J'ai une hypothèse de recherche qui correspond à ce que je veux croire au sujet du monde, et je peux l'appliquer à une hypothèse statistique qui correspond à ce que je veux croire sur la façon dont les données ont été générées. C'est à ce stade que les choses deviennent quelque peu contre-intuitives pour beaucoup de gens. Parce que ce que je m'apprête à faire, c'est inventer une nouvelle hypothèse statistique (l'hypothèse « nulle », *H0*) qui correspond exactement à l'opposé de ce que je veux croire, puis me concentrer exclusivement sur celle-ci presque au détriment de ce qui m'intéresse réellement (qu'on appelle maintenant l'hypothèse « alternative », *H1*). Dans notre exemple de PES, l'hypothèse nulle est que $\theta = 0.5$, puisque c'est ce à quoi on pourrait s'attendre si la PES n'existait pas. Mon espoir, bien sûr, est que PES soit totalement vraie et donc l'*alternative* à cette hypothèse nulle est $\theta \neq 0.5$ aussi. Fondamentalement, ce que nous faisons ici est de diviser les valeurs possibles de θ en deux groupes : ces valeurs dont j'espère vraiment qu'elles ne sont pas vraies (la valeur nulle), et les valeurs dont je serais heureux si elles s'avéraient exactes (l'alternative). Ce faisant, il est important de comprendre que le but d'un test d'hypothèse n'est pas de montrer que l'hypothèse alternative est (probablement) vraie. Le but est de montrer que l'hypothèse nulle est (probablement) fausse. La plupart des gens trouvent ça plutôt bizarre.

La meilleure façon d'y penser, d'après mon expérience, est d'imaginer qu'un test d'hypothèse est un procès criminel[55], *le procès de l'hypothèse nulle*. L'hypothèse nulle est l'accusé, le chercheur est le procureur et le test statistique lui-même est le juge. Tout comme dans un procès criminel, il y a présomption d'innocence. L'hypothèse nulle est *considérée comme* vraie à moins que vous, le chercheur, ne puissiez prouver hors de tout doute raisonnable qu'elle est fausse. Vous êtes libre de concevoir votre expérience comme bon vous semble (dans les limites du raisonnable, évidemment !) et votre objectif est de maximiser les chances que les données aboutissent à une condamnation pour le crime d'être faux. Le piège, c'est que le test statistique établit les règles du procès et que ces règles sont conçues pour protéger l'hypothèse nulle, notamment pour s'assurer que si l'hypothèse nulle est vraie, les chances d'une fausse condamnation sont faibles. C'est très important. Après tout, l'hypothèse nulle n'a pas d'avocat, et étant donné que le chercheur essaie désespérément de prouver qu'elle est fausse, il faut la protéger.

Deux types d'erreurs

Avant d'entrer dans les détails sur la façon dont un test statistique est construit, il est utile de comprendre la philosophie qui le sous-tend. J'y ai fait allusion en soulignant la similitude

---

[55] Cette analogie ne fonctionne que si vous êtes issu d'un système juridique accusatoire comme celui du Royaume-Uni, des États-Unis et de l'Australie. D'après ce que j'ai compris, le système inquisitoire français est très différent.

entre un test d'hypothèse nulle et un procès criminel, mais je dois maintenant être explicite. Idéalement, nous aimerions construire notre test de manière à ne jamais faire d'erreurs. Malheureusement, comme le monde est en désordre, ce n'est jamais possible. Parfois, on est vraiment malchanceux. Par exemple, supposez que vous tirez une pièce de monnaie 10 fois de suite et qu'elle tombe sur face 10 fois de suite. Cela semble être une preuve très solide pour conclure que la pièce est biaisée, mais bien sûr, il y a une chance sur 1024 que cela se produise même si la pièce était totalement juste. En d'autres termes, dans la vraie vie, nous devons *toujours* accepter qu'il y a une chance que nous ayons commis une erreur. Par conséquent, l'objectif des tests d'hypothèses statistiques n'est pas d'*éliminer les* erreurs, mais de les *minimiser*.

A ce stade, nous devons être un peu plus précis sur ce que nous entendons par « erreurs ». D'abord, disons ce qui est évident. Soit que l'hypothèse nulle est vraie, soit qu'elle est fausse, et notre test retiendra ou rejettera l'hypothèse nulle.[56] Ainsi, comme l'illustre le tableau ci-dessous, après avoir effectué le test et fait notre choix, l'une des quatre situations suivantes aurait pu se produire :

	retenir H_0	rejeter H_0
H_0 est vrai	décision juste	erreur (type I)
H_0 est faux	erreur (type II)	décision juste

Par conséquent, il y a *deux* types d'erreurs différents. Si nous rejetons une hypothèse nulle qui est en fait vraie, alors nous avons fait une **erreur de type I**. Par contre, si nous retenons l'hypothèse nulle alors qu'elle est en fait fausse, nous avons fait une **erreur de type II**.

---

[56] Un aparté concernant le langage que vous utilisez pour parler de vérification d'hypothèses. Tout d'abord, un terme que vous devriez vraiment éviter, c'est le mot « prouver ». Un test statistique ne prouve pas vraiment qu'une hypothèse est vraie ou fausse. La preuve implique la certitude et, comme le dit l'adage, les statistiques signifient qu'il n'est jamais nécessaire de dire que l'on est certain. Sur ce point, presque tout le monde sera d'accord. Cependant, au-delà de cela, il y a beaucoup de confusion. Certains prétendent que vous n'avez le droit de faire que des affirmations comme « rejeter l'hypothèse nulle », « omettre de rejeter la l'hypothèse nulle », ou peut-être « retenir l'hypothèse nulle ». Selon ce point de vue, on ne peut pas dire des choses comme « accepter l'hypothèse alternative » ou « accepter l'hypothèse nulle ». Personnellement, je pense que c'est trop sévère. À mon avis, cela confond la mise à l'épreuve d'hypothèses nulles avec la vision falsifiante du processus scientifique de Karl Popper. Bien qu'il y ait des similitudes entre la falsification et la vérification d'hypothèses nulles, elles ne sont pas équivalentes. Cependant, bien que je pense personnellement que c'est bien de parler d'accepter une hypothèse (à condition que « l'acceptation » ne signifie pas nécessairement que c'est vrai, surtout dans le cas de l'hypothèse nulle), beaucoup de gens ne seront pas d'accord. Surtout gardez à l'esprit que cette bizarrerie particulière existe pour ne pas être pris au dépourvu lors de la rédaction de vos propres résultats.

Vous vous souvenez quand j'ai dit que les tests statistiques étaient un peu comme un procès criminel ? Eh bien, je le pensais vraiment. Un procès criminel exige que vous établissiez « hors de tout doute raisonnable » que l'accusé l'a commis. Toutes les règles de preuve sont (du moins en théorie) conçues pour s'assurer qu'il n'y a (presque) aucune chance de condamner à tort un suspect innocent. Le procès vise à protéger les droits de l'accusé, comme l'a dit le juriste anglais William Blackstone, « il vaut mieux que dix coupables s'échappent que celui qu'un innocent qui souffre ». En d'autres termes, un procès criminel ne traite pas les deux types d'erreur de la même façon. Punir l'innocent est considéré comme bien pire que de libérer le coupable. Un test statistique est à peu près la même chose. Le principe de conception le plus important du test est de *contrôler la* probabilité d'une erreur de type I, pour la maintenir en dessous d'une probabilité fixe. Cette probabilité, qui est dénommée α, est appelée le **niveau de signification** du test. Et je le répète encore une fois, parce que c'est tellement central dans l'ensemble du dispositif, un test d'hypothèse est dit avoir un niveau de signification α si le taux d'erreur de type I n'est pas supérieur à α

Qu'en est-il alors du taux d'erreur de type II ? Eh bien, nous aimerions aussi les garder sous contrôle, et nous indiquons cette probabilité par β. Cependant, il est beaucoup plus courant de se référer à la **puissance** du test, c'est-à-dire la probabilité avec laquelle nous rejetons une hypothèse nulle quand elle est vraiment fausse, qui est $1 - \beta$. Pour aider à garder cela en tête, voici de nouveau le même tableau mais avec les chiffres pertinents ajoutés :

	retenir H_0	rejeter H_0
H_0 est vrai	$1 - \alpha$ (probabilité de conservation correcte)	α (taux d'erreur de type I)
H_0 est faux	β (taux d'erreur de type II)	$1 - \beta$ (puissance du test)

Un test d'hypothèse « puissant » est un test qui a une petite valeur de β, tout en gardant α fixé au (petit) niveau souhaité. Par convention, les scientifiques utilisent trois niveaux différents de α : .05, .01 et .001. Notez l'asymétrie ici ; les tests sont conçus pour s'assurer que le niveau de α est maintenu petit mais il n'y a aucune garantie correspondante concernant β. Nous aimerions certainement que le taux d'erreur de type II soit petit et nous essayons de concevoir des tests qui le gardent petit, mais ceci est généralement secondaire au besoin écrasant de contrôler le taux d'erreur du type I. Comme l'aurait dit Blackstone s'il avait été statisticien, il est « préférable de retenir 10 fausses hypothèses nulles que d'en rejeter une seule vraie ». Pour être honnête, je ne sais pas si je suis d'accord avec cette philosophie. Il y a des situations où je pense que c'est logique, et d'autres où ce n'est pas le cas, mais ce n'est ni ici ni là. C'est ainsi que les tests sont construits.

Statistiques des tests et distributions d'échantillonnage

À ce stade, nous devons commencer à parler plus précisément de la façon dont un test d'hypothèse est construit. Pour ce faire, revenons à l'exemple de la PES. Ignorons les données réelles que nous avons obtenues, pour l'instant, et pensons à la structure de l'expérience. Quels que soient les chiffres réels, la *forme des* données est que X sur N personnes ont correctement identifié la couleur de la carte cachée. De plus, supposons pour

l'instant que l'hypothèse nulle soit vraie, que la PES n'existe pas et que la vraie probabilité que quelqu'un choisisse la bonne couleur soit exactement $\theta = 0.5$. À quoi devraient ressembler les données ? Évidemment, on s'attendrait à ce que la proportion de personnes qui donnent la bonne réponse soit assez près de 50 p. 100. Ou, pour exprimer cela en termes plus mathématiques, nous dirions que *X/N* est d'environ 0,5. Bien sûr, on ne s'attendrait pas à ce que cette fraction soit *exactement de* 0,5. Si, par exemple, on testait *N* = 100 personnes et que *X* = 53 d'entre elles avaient la bonne réponse, on serait probablement forcé d'admettre que les données sont assez cohérentes avec l'hypothèse nulle. D'un autre côté, si *X* = 99 de nos participants répondaient correctement, nous serions assez confiants que l'hypothèse nulle est fausse. De même, si seulement *X* = 3 personnes obtenaient la bonne réponse, nous serions tout aussi confiants que l'hypothèse nulle est fausse. Soyons un peu plus techniques à ce sujet. Nous avons une quantité *X* que nous pouvons calculer en regardant nos données. Après avoir examiné la valeur de *X,* nous décidons s'il faut croire que l'hypothèse nulle est correcte ou rejeter l'hypothèse nulle en faveur de l'alternative. Le nom de cette valeur que nous calculons pour guider nos choix est une **statistique de test**.

Après avoir choisi une statistique de test, l'étape suivante consiste à indiquer précisément quelles valeurs de la statistique de test entraîneraient le rejet de l'hypothèse nulle et quelles valeurs nous inciteraient à la conserver. Pour ce faire, nous devons déterminer quelle serait la **distribution d'échantillonnage de la statistique de test** si l'hypothèse nulle était réellement vraie (nous avons parlé des distributions d'échantillonnage plus tôt dans la section 8.3.1). Pourquoi en avons-nous besoin ? Parce que cette distribution nous dit exactement quelles valeurs de *X* notre hypothèse nulle nous amènerait à attendre. Et, par conséquent, nous pouvons utiliser cette distribution comme outil pour évaluer dans quelle mesure l'hypothèse nulle correspond à nos données.

Comment détermine-t-on réellement la distribution d'échantillonnage de la statistique du test ? Pour beaucoup de tests d'hypothèse, cette étape est en fait assez compliquée, et plus tard dans le livre, vous me verrez être un peu évasif à ce sujet pour certains tests (pour certains d'entre eux, je ne les comprends même pas moi-même). Cependant, c'est parfois très facile. Et, heureusement pour nous, notre exemple de PES nous fournit l'un des cas les plus faciles. Notre paramètre de population θ n'est que la probabilité globale que les gens répondent correctement lorsqu'on leur pose la question, et notre statistique de test *X* est le *nombre* de personnes qui l'ont fait sur un échantillon de *N*.

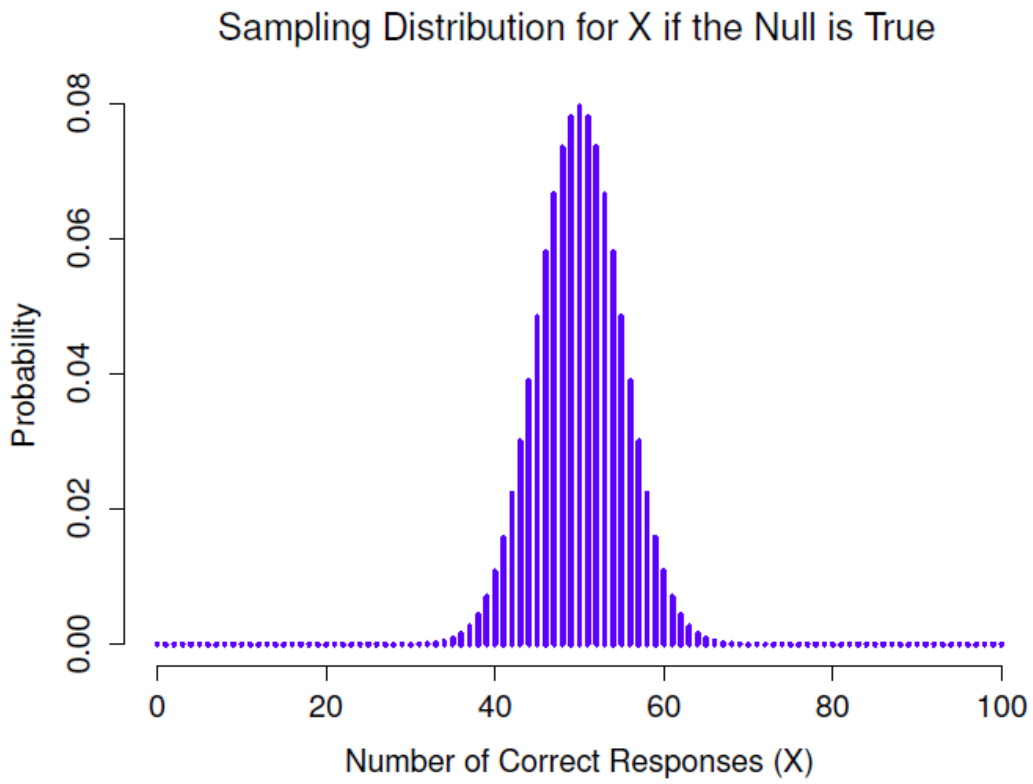


Figure 9-1 : La distribution d'échantillonnage pour notre statistique de test X lorsque l'hypothèse nulle est vraie. Pour notre scénario ESP, il s'agit d'une distribution binomiale. Comme on pouvait s'y attendre, puisque l'hypothèse nulle indique que la probabilité d'une réponse correcte est $\theta = 0{,}5$, la distribution d'échantillonnage indique que la valeur la plus probable est 50 (sur 100) bonnes réponses. La plus grande partie de la masse de probabilité se situe entre 40 et 60.

Nous avons déjà vu une distribution comme celle-ci à la section 7.4, et c'est exactement la même que la distribution binomiale ! Ainsi, pour utiliser la notation et la terminologie que j'ai présentées dans cette section, nous dirions que l'hypothèse nulle prédit que X est distribué de façon binomiale, ce qui est écrit

$$X \sim Binomiale\ (\theta, N)$$

Puisque l'hypothèse nulle indique que $\theta = 0{,}5$ et que notre expérience compte $N = 100$ personnes, nous avons la distribution d'échantillonnage dont nous avons besoin. Cette distribution d'échantillonnage est illustrée à la Figure 9-1. Sans surprise, l'hypothèse nulle dit que $X = 50$ est le résultat le plus probable, et qu'il est presque certain que nous verrons entre 40 et 60 bonnes réponses.

Prendre des décisions

Bon, nous sommes très près d'en avoir fini. Nous avons construit une statistique de test (X) et nous avons choisi cette statistique de test de telle manière que nous sommes assez

confiants que si *X* est proche de *N/2* alors nous devrions conserver l'hypothèse nulle, sinon nous devrions la rejeter. La question qui demeure est la suivante. Quelles valeurs exactes de la statistique du test faut-il associer à l'hypothèse nulle, et quelles valeurs vont exactement avec l'hypothèse alternative ? Dans mon étude ESP, par exemple, j'ai observé une valeur de *X*=62. Quelle décision devrais-je prendre ? Devrais-je choisir de croire l'hypothèse nulle ou l'hypothèse alternative ?

Régions critiques et valeurs critiques

Pour répondre à cette question, nous devons introduire le concept de **région critique** pour la statistique de test *X*. La région critique du test correspond aux valeurs de *X* qui nous amèneraient à rejeter une hypothèse nulle (c'est pourquoi la région critique est aussi parfois appelée région de rejet). Comment trouver cette région critique ? Eh bien, considérons ce que nous savons :

- *X* doit être très grand ou très petit pour rejeter l'hypothèse nulle.
- Si l'hypothèse nulle est vraie, la distribution d'échantillonnage de *X* est Binomiale (0,5, N).
- Si $\alpha = 0.5$, la région critique doit couvrir 5% de cette distribution d'échantillonnage.

Il est important que vous compreniez bien ce dernier point. La région critique correspond aux valeurs de *X* pour lesquelles on rejetterait l'hypothèse nulle, et la distribution d'échantillonnage en question décrit la probabilité d'obtenir une valeur particulière de *X* si l'hypothèse nulle était effectivement vraie. Supposons maintenant que nous choisissions une région critique qui couvre 20% de la distribution d'échantillonnage, et supposons que l'hypothèse nulle soit vraie. Quelle serait la probabilité d'un rejet incorrect de l'annulation ? La réponse est bien sûr 20 %. Et, par conséquent, nous aurions construit un test ayant un niveau α de 0,2. Si nous voulons $\alpha = 0.5$, la région critique ne *peut* couvrir que 5% de la distribution d'échantillonnage de notre statistique de test.

Il s'avère que ces trois choses résolvent le problème de façon unique. Notre région critique se compose des *valeurs* les plus *extrêmes*, connues sous le nom de **queues de la distribution**. C'est ce qu'illustre la Figure 9-2. Si nous voulons $\alpha = 0.5$ alors nos régions critiques correspondent à $X \leq 40$ et $X \geq 60$.[57] C'est-à-dire que, si le nombre de personnes qui disent «vrai» se situe entre 41 et 59, alors nous devrions retenir l'hypothèse nulle. Si le nombre est compris entre 0 et 40, ou entre 60 et 100, alors nous devrions rejeter l'hypothèse nulle. Les nombres 40 et 60 sont souvent appelés **valeurs critiques** car ils définissent les limites de la région critique.

---

[57] Strictement parlant, le test que je viens de construire a $\alpha = .057$, ce qui est un peu trop généreux. Cependant, si j'avais choisi 39 et 61 comme limites pour la région critique, alors la région critique ne couvre que 3,5 % de la distribution. Je me suis dit qu'il était plus logique d'utiliser 40 et 60 comme valeurs critiques et d'être prêt à tolérer un taux d'erreur de type I de 5,7%, puisque c'est le plus près possible d'une valeur de $\alpha = 0.5$.

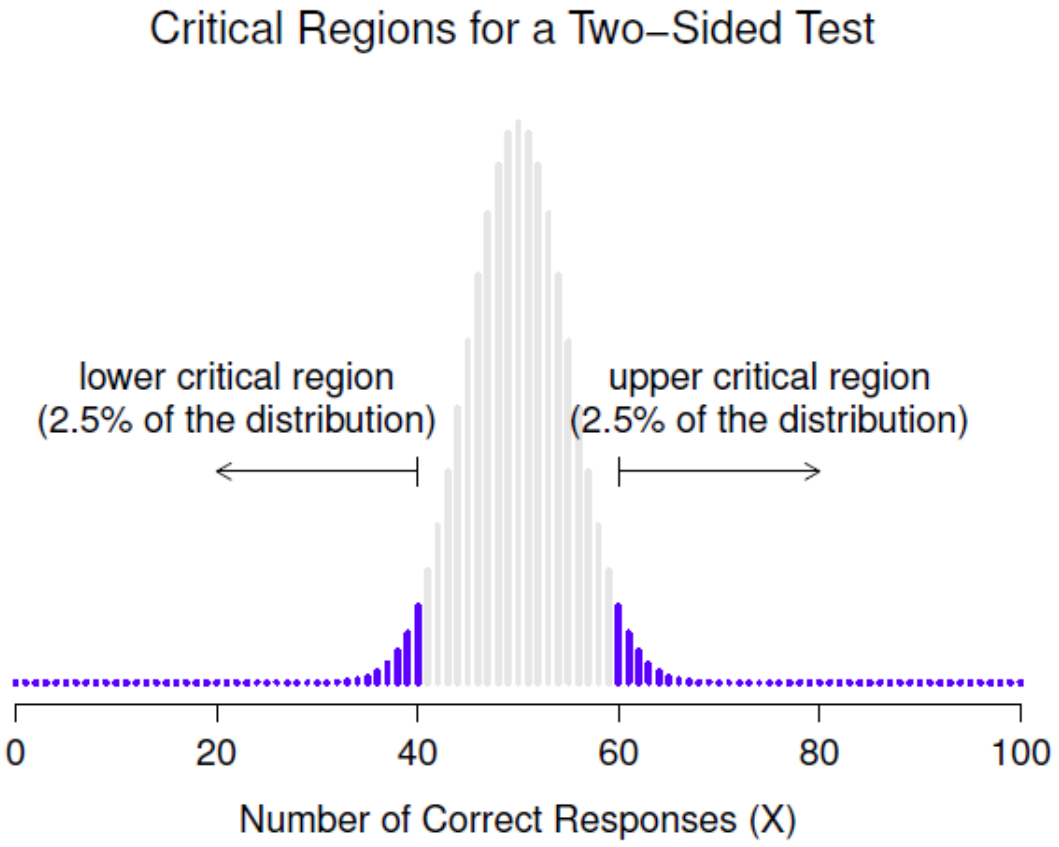


Figure 9-2 : La région critique associée au test d'hypothèse pour l'étude ESP, pour un test d'hypothèse avec un niveau de signification de $\alpha = .05$. Le graphique montre la distribution d'échantillonnage de X sous l'hypothèse nulle (c.-à-d. identique à la Figure 9-1). Les barres grises correspondent aux valeurs de X pour lesquelles on retiendrait l'hypothèse nulle. Les barres bleues (plus foncées) indiquent la région critique, ces valeurs de X pour lesquelles nous rejetterions la valeur nulle. Étant donné que l'hypothèse alternative est bilatérale (c.-à-d. qu'elle permet à la fois $\theta < 0{,}5$ et $\theta > 0{,}5$), la région critique couvre les deux queues de la distribution. Pour assurer un niveau de .05 à α, nous devons nous assurer que chacune des deux régions englobe 2,5 % de la distribution de l'échantillonnage.

À ce stade, notre test d'hypothèse est pratiquement terminé :

1. nous choisissons un niveau α (p. ex. $\alpha = .05$) ;
2. calculons des statistiques de test (p. ex., X) qui permettent de bien comparer H_0 à H_1 (d'un point de vue qui ai su sens) ;
3. déterminons la distribution d'échantillonnage de la statistique de test en supposant que l'hypothèse nulle est vraie (dans ce cas, c'est la binomiale) ; puis
4. calculons la région critique qui produit un niveau approprié de α (0-40 et 60-100).

Il ne nous reste plus qu'à calculer la valeur de la statistique du test pour les données réelles (par exemple, $X = 62$) et la comparer aux valeurs critiques pour prendre notre décision.

Puisque 62 est supérieur à la valeur critique de 60, nous rejetterions l'hypothèse nulle. Ou, pour le formuler un peu différemment, nous disons que le test a produit un résultat statistiquement **significatif.**

Note sur la « signification » statistique

Comme d'autres techniques occultes de divination, la méthode statistique a un jargon privé délibérément inventé pour cacher ses méthodes aux non-praticiens. - Attribué à G. O. Ashley[58]

Une très brève digression s'impose à ce stade, en ce qui concerne le mot « significatif ». Le concept de signification statistique est en fait très simple, mais il porte un nom très malheureux. Si les données nous permettent de rejeter l'hypothèse nulle, nous disons que « le résultat est *statistiquement significatif* », ce qui est souvent réduit à « le résultat est significatif ». Cette terminologie est plutôt ancienne et remonte à une époque où « significatif » signifiait simplement quelque chose comme « signifié », plutôt que sa signification moderne qui est beaucoup plus proche de « important ». En conséquence, beaucoup de lecteurs modernes sont très confus lorsqu'ils commencent à apprendre les statistiques parce qu'ils pensent qu'un « résultat significatif » doit être un résultat important. Ça ne veut pas dire ça du tout. Tout ce que « statistiquement significatif » signifie, c'est que les données nous ont permis de rejeter une hypothèse nulle. La question de savoir si le résultat est réellement important dans le monde réel est une question très différente et dépend de toutes sortes d'autres choses.

La différence entre les tests unilatéraux et bilatéraux

Il y a encore une chose que je tiens à souligner au sujet du test d'hypothèse que je viens de construire. Si on prend un moment pour réfléchir aux hypothèses statistiques que j'ai utilisées, nous remarquons que l'hypothèse alternative couvre *à la fois* la possibilité que $\theta < .5$ et la possibilité que $\theta > .5$.

$H_0: \theta = .5$ $H_1: \theta \neq .5$

C'est logique si on pense vraiment que la PES pourrait produire soit une performance supérieure au hasard, soit une performance inférieure au hasard (et il y a des gens qui pensent cela). En langage statistique, il s'agit d'un exemple de **test bilatéral**. On l'appelle ainsi parce que l'hypothèse alternative couvre la zone des deux « côtés » de l'hypothèse nulle, et par conséquent la région critique du test couvre les deux queues de la distribution d'échantillonnage (2,5% de chaque côté si $\alpha = 0.5$), comme illustré précédemment dans la Figure 9-2.

Cependant, ce n'est pas la seule possibilité. Je ne suis prêt à croire en la PES que si elle produit de meilleures performances que le hasard. Si c'est le cas, alors mon hypothèse alternative ne couvrirait que la possibilité que $\theta > .5$, et par conséquent l'hypothèse nulle devient $\theta < .5$

---

[58] Internet semble assez convaincu qu'Ashley a dit cela, bien que je n'arrive pas à trouver quelqu'un prêt à donner une source pour cette affirmation

$H_0: \theta \leq .5$ $H_1: \theta$ .5

Lorsque cela se produit, nous avons ce qu'on appelle **un test unilatéral** et la région critique ne couvre qu'une seule queue de la distribution d'échantillonnage. C'est ce qu'illustre la Figure 9-3.

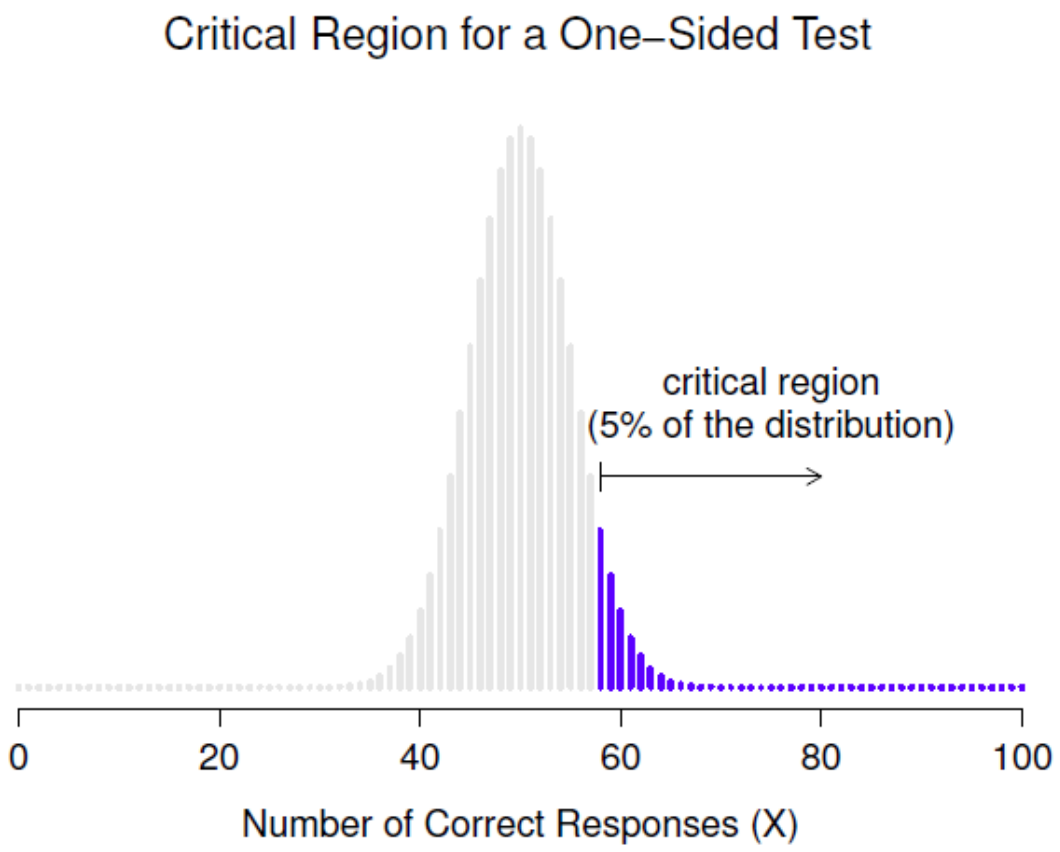


Figure 9-3 : La région critique pour un essai unilatéral. Dans ce cas, l'hypothèse alternative est que $\theta > .5$ de sorte que nous ne rejetterions que l'hypothèse nulle pour les grandes valeurs de X. Par conséquent, la région critique ne couvre que la partie supérieure de la queue de la distribution d'échantillonnage, en particulier les 5 % supérieurs de la distribution. Comparez cela à la version bilatérale de la Figure 9.2.

La valeur *p* d'un test

Dans un sens, notre test d'hypothèse est complet. Nous avons construit une statistique de test, calculé sa distribution d'échantillonnage si l'hypothèse nulle est vraie, et ensuite construit la région critique pour le test. Néanmoins, j'ai en fait omis le nombre le plus important de tous, **la valeur** *p*. C'est à ce sujet que nous passons maintenant. Il y a deux façons quelque peu différentes d'interpréter une valeur *p*, l'une proposée par Sir Ronald Fisher et l'autre par Jerzy Neyman. Les deux versions sont légitimes, bien qu'elles reflètent des façons très différentes de concevoir les tests d'hypothèse. La plupart des manuels d'introduction tendent à ne donner que la version de Fisher, mais je pense que c'est un peu

dommage. À mon avis, la version de Neyman est plus propre et reflète mieux la logique du test de l'hypothèse nulle. Mais vous n'êtes peut-être pas d'accord, alors j'ai inclus les deux. Je vais commencer par la version de Neyman.

Une vision plus douce de la prise de décision

L'un des problèmes de la procédure de vérification des hypothèses que j'ai décrite est qu'elle ne fait aucune distinction entre un résultat qui est « à peine significatif » et ceux qui sont « très significatifs ». Par exemple, dans mon étude sur le PES, les données que j'ai obtenues tombaient tout juste de à l'intérieur de la région critique, alors j'ai obtenu un effet significatif, mais c'était assez limite. Par contre, supposons que je fasse une étude dans laquelle $X = 97$ de mes N=100 participants ont obtenu la bonne réponse. Ce serait évidemment significatif aussi, mais ma marge est beaucoup plus grande, de sorte qu'il n'y a vraiment aucune ambiguïté à ce sujet. La procédure que j'ai déjà décrite ne fait aucune distinction entre les deux. Si j'adopte la convention standard d'autoriser $\alpha = .05$ comme taux d'erreur de type I acceptable, ces deux résultats sont significatifs.

C'est là que la valeur p est utile. Pour comprendre son fonctionnement, supposons que nous ayons effectué de nombreux tests d'hypothèses sur le même ensemble de données, mais avec une valeur différente de α dans chaque cas. Lorsque nous faisons cela pour mes données d'origine sur la PES, nous obtenons quelque chose comme ceci

Valeur de α	0.05	0.04	0.03	0.02	0.01
Rejeter H0 ?	Oui	Oui	Oui	Non	Non

Quand on teste les données de la PES ($X = 62$ succès sur $N = 100$ observations), en utilisant les niveaux α de .03 et plus, on se retrouve toujours à rejeter l'hypothèse nulle. Pour les niveaux de α de .02 et moins, nous finissons toujours par conserver l'hypothèse nulle. Par conséquent, quelque part entre .02 et.03, il doit y avoir une valeur minimale de α qui nous permettrait de rejeter l'hypothèse nulle pour ces données. C'est la valeur p. Il s'avère que les données de la PES, on a $p = .021$. En bref,

p est défini comme étant le plus petit taux d'erreur de type I (α) que vous devez être prêt à tolérer si vous voulez rejeter l'hypothèse nulle.

S'il s'avère que p décrit un taux d'erreur que vous trouvez intolérable, alors vous devez conserver la valeur nulle. Si vous êtes à l'aise avec un taux d'erreur égal à p, vous pouvez rejeter l'hypothèse nulle en faveur de votre alternative préférée.

En effet, p est un résumé de tous les tests d'hypothèses possibles que vous auriez pu faire, à travers toutes les valeurs possibles de α Et par conséquent, cela a pour effet de « moduler » notre processus de décision. Pour les tests dans lesquels $p \leq \alpha$ vous auriez rejeté l'hypothèse nulle, alors que pour les tests dans lesquels $p > \alpha$ vous auriez retenu l'hypothèse nulle. Dans mon étude sur la PES j'ai obtenu $X = 62$ et par conséquent j'ai obtenu p =.021. Le taux d'erreur que je dois tolérer est donc de 2,1 p. 100. Par contre, supposons que mon expérience ait donné $X = 97$. Qu'advient-il de ma valeur p maintenant ?

Cette fois-ci, il est réduit à p = 1.36*10-25, qui est un taux d'erreur de type I[59] minuscule. Pour ce second cas, je serais en mesure de rejeter l'hypothèse nulle avec beaucoup plus de confiance, parce qu'il me suffit d'être « disposé » à tolérer un taux d'erreur de type I d'environ 1 sur 10 trillions de milliards pour justifier ma décision de rejeter.

La probabilité de données extrêmes

La deuxième définition de la *p-value* vient de Sir Ronald Fisher, et c'est en fait celle-ci que vous avez tendance à voir dans la plupart des manuels d'introduction aux statistiques. Remarquez comment, lorsque j'ai construit la région critique, elle correspondait aux *queues* (c'est-à-dire aux valeurs extrêmes) de la distribution d'échantillonnage ? Ce n'est pas une coïncidence, presque tous les « bons » tests ont cette caractéristique (bon dans le sens de minimiser notre taux d'erreur de type II, β). La raison en est qu'une bonne région critique correspond presque toujours aux valeurs de la statistique du test qui sont les moins susceptibles d'être observées si l'hypothèse nulle est vraie. Si cette règle est vraie, nous pouvons définir la *valeur p* comme la probabilité d'observer une statistique de test qui est au moins aussi extrême que celle que nous avons effectivement obtenue. En d'autres termes, si les données sont extrêmement peu plausibles selon l'hypothèse nulle, alors l'hypothèse nulle est probablement fausse.

Une erreur courante

Bien, vous pouvez donc voir qu'il y a deux façons plutôt différentes mais légitimes d'interpréter la valeur p, l'une fondée sur l'approche de Neyman pour vérifier les hypothèses et l'autre sur celle de Fisher. Malheureusement, il y a une troisième explication que les gens donnent parfois, surtout lorsqu'ils apprennent des statistiques pour la première fois, et celle-ci est *absolument et complètement fausse*. Cette approche erronée consiste à désigner la valeur p comme « la probabilité que l'hypothèse nulle soit vraie ». C'est une façon intuitive et attrayante de penser, mais elle est erronée à deux égards. Premièrement, le test d'hypothèse nulle est un outil fréquentiste et l'approche fréquentiste de la probabilité *ne* vous permet *pas* d'attribuer des probabilités à l'hypothèse nulle. Selon cette conception de la probabilité, l'hypothèse nulle est soit vraie, soit fausse, elle ne peut pas avoir « 5% de chance » d'être vraie. Deuxièmement, même dans l'approche bayésienne, qui permet d'assigner des probabilités aux hypothèses, la valeur p ne correspondrait pas à la probabilité que la valeur nulle soit vraie. Cette interprétation est tout à fait incompatible avec les mathématiques du calcul de la valeur *p.* En clair, malgré l'attrait intuitif d'une telle façon de penser, il n'y a *aucune* justification pour interpréter une valeur p de cette façon. Ne jamais le faire.

---

[59] C'est p " . 00000000000000000000000000000136 pour ceux qui n'aiment pas la notation scientifique !

Communication des résultats d'un test d'hypothèse

Lorsque vous rédigez les résultats d'un test d'hypothèse, il y a habituellement plusieurs éléments d'information que vous devez rapporter, mais cela varie beaucoup d'un test à l'autre. Tout au long du reste du livre, je vais passer un peu de temps à parler de la façon de rapporter les résultats des différents tests (voir la Section 10.1.9 pour un exemple particulièrement détaillé), afin que vous puissiez vous faire une idée de la façon dont cela se fait habituellement. Cependant, peu importe le test que vous faites, la seule chose que vous devez toujours faire est de dire quelque chose sur la valeur p et si le résultat est significatif ou non.

Le fait que vous deviez le faire n'est pas surprenant, c'est tout l'intérêt de faire le test. Ce qui peut surprendre, c'est le fait qu'il y a une certaine controverse sur la façon exacte de faire les choses. Si l'on laisse de côté les personnes qui ne sont pas du tout d'accord avec l'ensemble du cadre qui sous-tend les tests d'hypothèse nulle, il existe une certaine tension quant à savoir s'il faut déclarer ou non la valeur exacte de p que vous avez obtenue, ou si vous devez déclarer seulement ce $p < \alpha$ pour un niveau de signification que vous avez préalablement choisi (p. ex., p<.05).

La question

Pour comprendre pourquoi il s'agit d'un problème, le point clé à comprendre est que les valeurs p sont terriblement pratiques. En pratique, le fait que nous puissions calculer une valeur p signifie que nous n'avons pas besoin du tout de spécifier un niveau α pour exécuter le test. A la place, vous pouvez calculer votre valeur p et l'interpréter directement. Si vous obtenez p = 0,062, cela signifie que vous devez être prêt à tolérer un taux d'erreur de type I de 6,2 % pour justifier le rejet de la valeur nulle. Si vous trouvez personnellement 6,2 % intolérable, vous conservez l'hypothèse nulle. Par conséquent, l'argument est le suivant : pourquoi ne pas simplement déclarer la valeur réelle de p et laisser au lecteur le soin de se faire sa propre opinion sur ce qu'est un taux d'erreur de type I acceptable ? Cette approche a le grand avantage « d'adoucir » le processus de prise de décision. En fait, si vous acceptez la définition de Neyman de p, c'est le sens même de la valeur de p. Nous n'avons plus un niveau de signification fixe de $\alpha = .05$ comme la limite qui sépare les décisions « accepter » et « rejeter », ce qui élimine le problème plutôt pathologique de devoir traiter p=.051 d'une manière fondamentalement différente de p =.049.

Cette flexibilité est à la fois l'avantage et l'inconvénient de la valeur p. La raison pour laquelle beaucoup de gens n'aiment pas l'idée de rapporter une valeur p exacte est que cela donne un peu *trop* de liberté au chercheur. En particulier, il vous permet de changer d'avis sur la tolérance aux erreurs que vous êtes prêt à tolérer *après avoir* examiné les données. Prenons, par exemple, mon expérience sur la PES. Supposons que j'ai fait mon test et que j'ai obtenu une valeur p de 0,09. Dois-je accepter ou rejeter ? Maintenant, pour être honnête, je n'ai pas encore pris la peine de penser au niveau d'erreur de type I que je suis « vraiment » prêt à accepter. Je n'ai pas d'opinion à ce sujet. Mais j'ai une opinion sur l'existence ou non de la PES, et j'ai *certainement* une opinion sur le fait que mes recherches devraient être publiées dans une revue scientifique réputée. Et étonnamment, maintenant que j'ai examiné les données, je commence à penser qu'un taux d'erreur de 9 % n'est pas si

mauvais, surtout si on le compare au fait qu'il serait ennuyeux d'avoir à reconnaitre publiquement que mon expérience a échoué. Donc, pour ne pas avoir l'air d'avoir tout inventé après coup, je dis maintenant que mon α est .1, avec l'argument qu'un taux d'erreur de type I de 10% n'est pas trop mauvais et à ce niveau mon test est significatif ! J'ai gagné.

En d'autres termes, ce qui m'inquiète ici, c'est qu'ayant les meilleures intentions et étant la plus honnête des personnes, la tentation de cacher un peu les choses ici et là est vraiment, vraiment forte. Comme tous ceux qui ont déjà fait une expérience peuvent en témoigner, c'est un processus long et difficile et on s'attache souvent *beaucoup à* ses hypothèses. C'est difficile de laisser aller et d'admettre que l'expérience n'a pas permis de trouver ce que vous vouliez trouver. Et c'est là que réside le danger. Si nous utilisons la *valeur p* « brute », les gens commenceront à interpréter les données en fonction de ce qu'ils *veulent* croire, et non de ce que les données disent réellement et, si nous le permettons, pourquoi nous donnons-nous la peine de faire de la science ? Pourquoi ne pas laisser tout le monde croire ce qu'il veut de n'importe quoi, quels que soient les faits ? C'est un peu extrême, mais c'est de là que vient l'inquiétude. Selon ce point de vue, vous *devez* vraiment spécifier votre valeur α à l'avance et ensuite seulement indiquer si le test était significatif ou non. C'est le seul moyen de rester honnête.

Tableau 9-1 : Une convention communément adoptée pour la déclaration des valeurs p : dans de nombreuses publications, il est conventionnel de déclarer un de quatre résultats différents (par exemple, p < .05) comme indiqué ci-dessous. J'ai inclus la notation des « étoiles » (i.e., un * indique p < .05) parce que vous voyez parfois cette notation produite par un logiciel statistique. Il convient également de noter que certaines personnes écrivent ns. (non significatif) plutôt que p > .05.

Notation usuelle	Signif. étoiles	En langage courant	HO est….
$p > .05$	ns	Le test n'était pas significatif	Retenue
$p < .05$	*	Le test était significatif sur $\alpha = .05$ mais pas sur $\alpha = .01$ ou $\alpha = .001$.	Rejetée
$p < .01$	**	Le test était significatif sur $\alpha = .05$ et $\alpha = .01$ mais pas sur $\alpha = .001$.	Rejetée
$p < .001$	***	Le test était significatif à tous les niveaux	Rejetée

Deux solutions proposées

Dans la pratique, il est assez rare qu'un chercheur spécifie un seul niveau α à l'avance. La convention veut plutôt que les scientifiques se fondent sur trois niveaux de signification standard : .05,.01 et.001. Lorsque vous rapportez vos résultats, vous indiquez lesquels (le cas échéant) de ces niveaux de signification vous permettent de rejeter l'hypothèse nulle. Ce point est résumé dans le Tableau 9-1. Cela nous permet d'assouplir un peu la règle de décision, puisque $p < .01$ implique que les données satisfont à une norme de preuve plus

stricte que $p < .05$ le ferait. Néanmoins, comme ces niveaux sont fixés à l'avance par convention, cela empêche les gens de choisir leur niveau α après avoir examiné les données.

Néanmoins, un grand nombre de personnes préfèrent encore communiquer des valeurs *p* exactes. Pour beaucoup, l'avantage de permettre au lecteur de se faire sa propre opinion sur la façon d'interpréter $p = .06$ l'emporte sur les inconvénients éventuels. Dans la pratique, cependant, même parmi les chercheurs qui préfèrent des valeurs *p* exactes, il est assez courant d'écrire simplement $p < .001$ au lieu de rapporter une valeur exacte pour un petit *p*. C'est en partie parce que beaucoup de logiciels n'impriment pas réellement la valeur de *p* quand elle est si petite (par exemple, SPSS écrit simplement $p = .000$ chaque fois que $p < .001$), et en partie parce qu'une très petite valeur *p* peut être un peu trompeuse. L'esprit humain voit un nombre comme .00000000000001 et il est difficile de supprimer le sentiment instinctif que la preuve en faveur de l'hypothèse alternative est une quasi-certitude. Dans la pratique, cependant, ce n'est généralement pas le cas. La vie est une activité importante, désordonnée et compliquée, et tous les tests statistiques jamais inventés reposent sur des simplifications, des approximations et des hypothèses. Par conséquent, il n'est probablement pas raisonnable de s'éloigner d'*une* analyse statistique avec un sentiment de confiance plus fort que $p < .001$. En d'autres termes, p<.001 est vraiment un véritable indice de ce que sont "les preuves convaincantes pour le test concerné".

À la lumière de tout cela, vous vous demandez peut-être ce que vous devriez faire exactement. Il y a pas mal de conseils contradictoires sur le sujet, certaines personnes soutenant que vous devriez déclarer la valeur exacte de *p*, et d'autres que vous devriez utiliser l'approche par paliers illustrée dans le Tableau 9-1. Par conséquent, le meilleur conseil que je puisse donner est de vous suggérer de consulter les documents/rapports écrits dans votre domaine et de voir ce que semble être la convention. S'il ne semble pas y avoir de tendance constante, utilisez la méthode que vous préférez.

Exécuter le test d'hypothèse dans la pratique

Certains d'entre vous se demandent peut-être s'il s'agit d'un « vrai » test d'hypothèse, ou simplement d'un exemple de jeu que j'ai inventé. C'est réel. Dans la discussion précédente, j'ai construit le test à partir de principes de base, pensant que c'était le problème le plus simple que l'on puisse rencontrer dans la vie réelle. Cependant, ce test existe déjà. C'est ce qu'on appelle le *test binomial*, et il est implémenté dans Jamovi comme l'une des analyses statistiques disponibles lorsque vous cliquez sur le bouton « Frequencies ». Pour vérifier l'hypothèse nulle selon laquelle la probabilité de réponse est égale à la moitié de p = 0,5,[60] et en utilisant des données dans lesquelles x = 0,62 de n = 100 personnes ont donné la bonne réponse, disponible dans le fichier de données binomialtest.omv, nous obtenons les résultats présentés à la Figure 9-4.

---

[60] Notez que le p ici n'a rien à voir avec une valeur p. L'argument p du test binomial de Jamovi correspond à la probabilité d'obtenir une réponse correcte, selon l'hypothèse nulle. En d'autres termes, c'est la valeur θ.

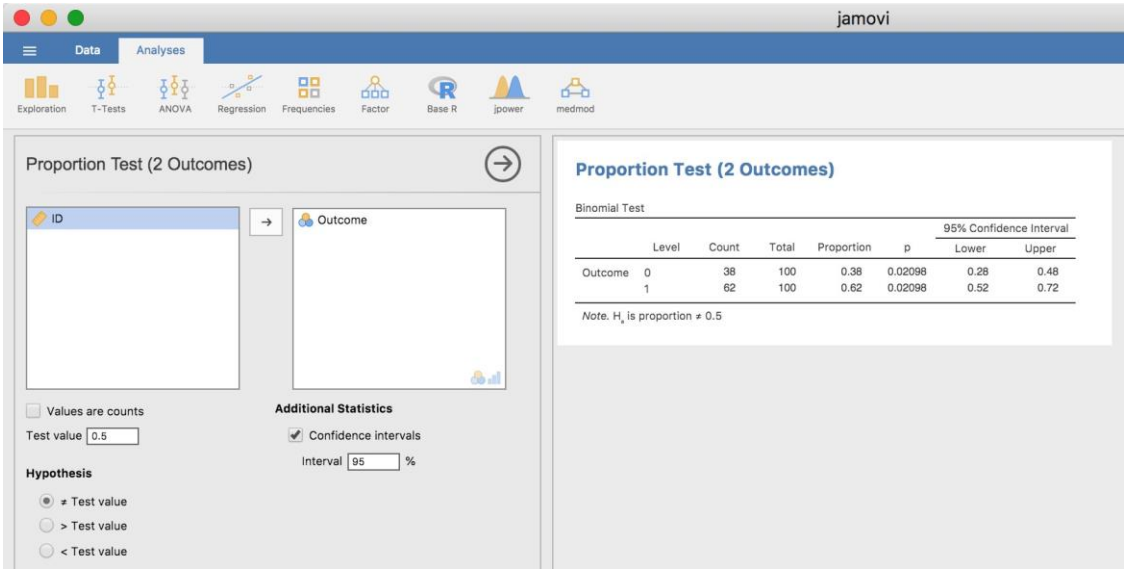


Figure 9-4 : Analyse du test binomial et résultats dans Jamovi

Pour l'instant, cette sortie ne vous semble pas très familière, mais vous pouvez voir qu'elle vous donne plus ou moins les bons résultats. Plus précisément, la *valeur p* de 0,02 est inférieure au choix habituel de $\alpha < .05$, vous pouvez donc rejeter l'hypothèse nulle. Nous parlerons beaucoup plus de la façon de lire ce genre de résultats au fur et à mesure, et après un certain temps, nous espérons que vous le trouverez assez facile à lire et à comprendre.

Taille de l'effet, taille de l'échantillon et puissance

Dans les sections précédentes, j'ai insisté sur le fait que le principal principe de conception qui sous-tend les tests d'hypothèses statistiques est que nous essayons de contrôler notre taux d'erreur de type I. Quand on fixe $\alpha = .05$, nous essayons de nous assurer que seulement 5 % des hypothèses nulles réelles sont rejetées à tort. Cependant, cela ne signifie pas que nous ne nous soucions pas des erreurs de type II. En fait, du point de vue du chercheur, l'erreur de ne pas rejeter l'hypothèse nulle alors qu'elle est en fait fausse est extrêmement ennuyeuse. En gardant cela à l'esprit, un objectif secondaire des tests d'hypothèse est d'essayer de minimiser β, le taux d'erreur de type II, bien que nous ne *parlions* généralement pas en termes de minimiser les erreurs de type II. Au lieu de cela, nous parlons de maximiser la *puissance* du test. Puisque la puissance est définie comme $1 - \beta$, c'est la même chose.

La fonction de puissance

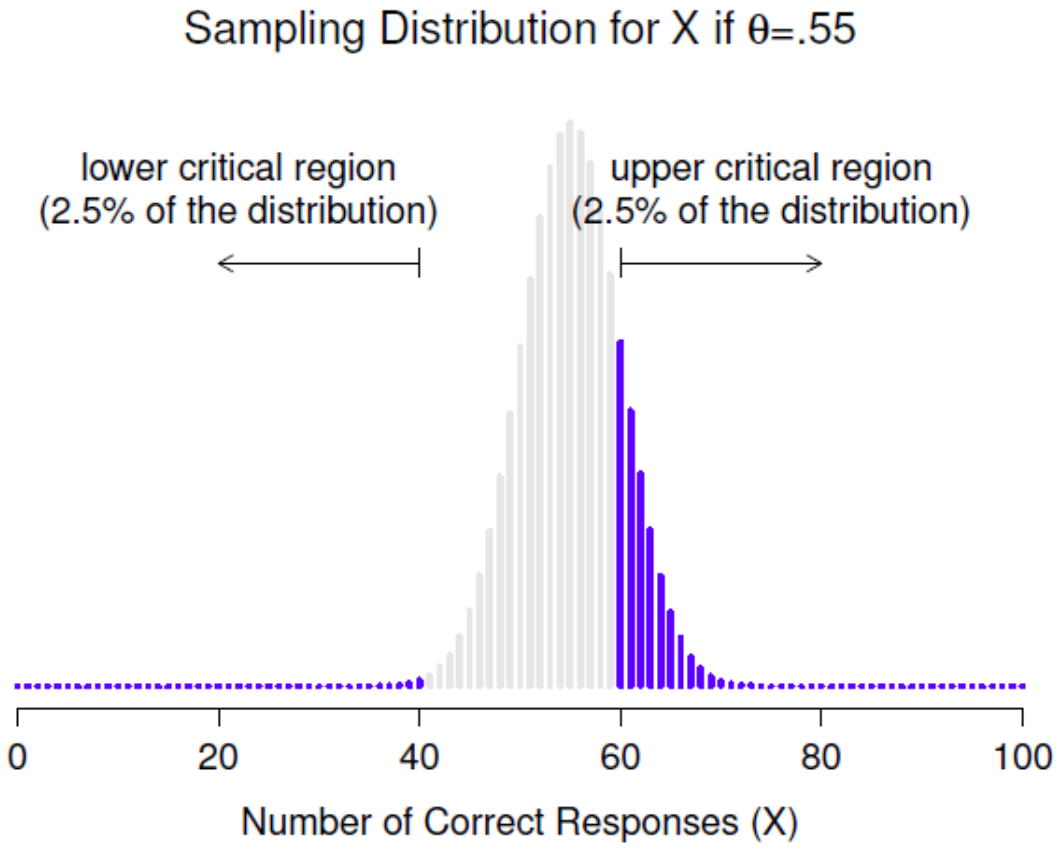


Figure 9-5 : Distribution de l'échantillonnage sous l'hypothèse alternative pour une valeur de paramètre de population de $\theta = 0.55$. Une proportion raisonnable de la distribution se trouve dans la région de rejet.

Prenons un moment pour réfléchir à ce qu'est une erreur de type II. Une erreur de type II se produit lorsque l'hypothèse alternative est vraie, mais que nous ne sommes pas en mesure de rejeter l'hypothèse nulle. Idéalement, nous pourrions calculer un nombre unique β qui nous indique le taux d'erreur de type II, de la même manière que nous pouvons régler $\alpha = .05$ pour le taux d'erreur de type I. Malheureusement, c'est beaucoup plus difficile à faire. Pour vous en rendre compte, notez que dans mon étude sur la PES l'hypothèse alternative correspond en fait à beaucoup de valeurs possibles de θ. En fait, l'hypothèse alternative correspond à toutes les valeurs de θ *à l'exception de* 0,5. Supposons que la probabilité réelle que quelqu'un choisisse la bonne réponse est de 55 % (c.-à-d. $\theta = 0.55$). Si c'est le cas, la distribution d'échantillonnage *réelle* pour *X* n'est pas la même que celle prévue par l'hypothèse nulle, car la valeur la plus probable pour *X* est maintenant de 55 sur 100. De plus, toute la distribution d'échantillonnage a maintenant changé, comme le montre la Figure 9-5. Les régions critiques, bien sûr, ne changent pas. Par définition, les régions critiques sont basées sur ce que l'hypothèse nulle prédit. Ce que nous constatons dans cette figure, c'est que lorsque l'hypothèse nulle est fausse, une proportion beaucoup plus importante de la distribution d'échantillonnage se situe dans la région critique. Et bien sûr,

c'est ce qui devrait arriver. La probabilité de rejeter l'hypothèse nulle est plus grande lorsque l'hypothèse nulle est fausse ! Cependant $\theta = 0.55$ n'est pas la seule possibilité compatible avec l'hypothèse alternative. Supposons plutôt que la valeur réelle de θ est en fait 0.70. Qu'arrive-t-il à la distribution d'échantillonnage lorsque cela se produit ? La réponse, illustrée à la Figure 9-6, est que la quasi-totalité de la distribution d'échantillonnage est maintenant passée dans la région critique. Par conséquent, si $\theta = 0.70$, la probabilité que nous rejetons correctement l'hypothèse nulle (c'est-à-dire la puissance du test) est beaucoup plus grande que si $\theta = 0.55$. Bref, alors que $\theta = 0.55$ et $\theta = 0.70$ font partie de l'hypothèse alternative, le taux d'erreur de type II est différent.

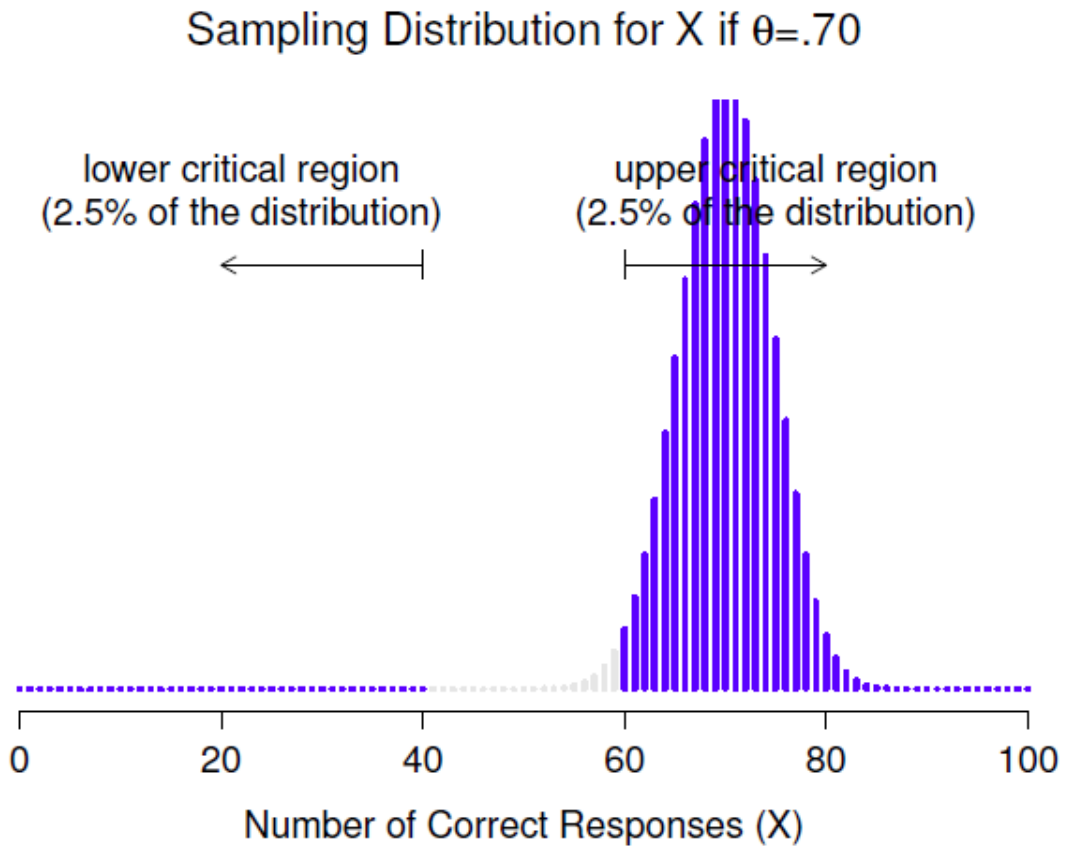


Figure 9-6 : Distribution de l'échantillonnage sous l'hypothèse alternative pour une valeur de paramètre de population de $\theta = .70$. Presque toute la distribution se trouve dans la région de rejet.

Tout cela signifie que la puissance d'un test (c.-à-d. $1 - \beta$) dépend de la valeur réelle de θ. Pour illustrer cela, j'ai calculé la probabilité prévue de rejeter l'hypothèse nulle pour toutes les valeurs de θ, et je l'ai représentée à la Figure 9-7. Ce graphique décrit ce que l'on appelle habituellement la **fonction de puissance** du test. C'est un bon résumé de la qualité du test, car il vous indique la puissance $(1 - \beta)$ pour toutes les valeurs possibles de θ. Comme vous pouvez le constater, lorsque la valeur réelle de θ est très élevée proche de 0,5, la puissance du test diminue très fortement, mais lorsqu'il est plus éloigné, la puissance est importante.

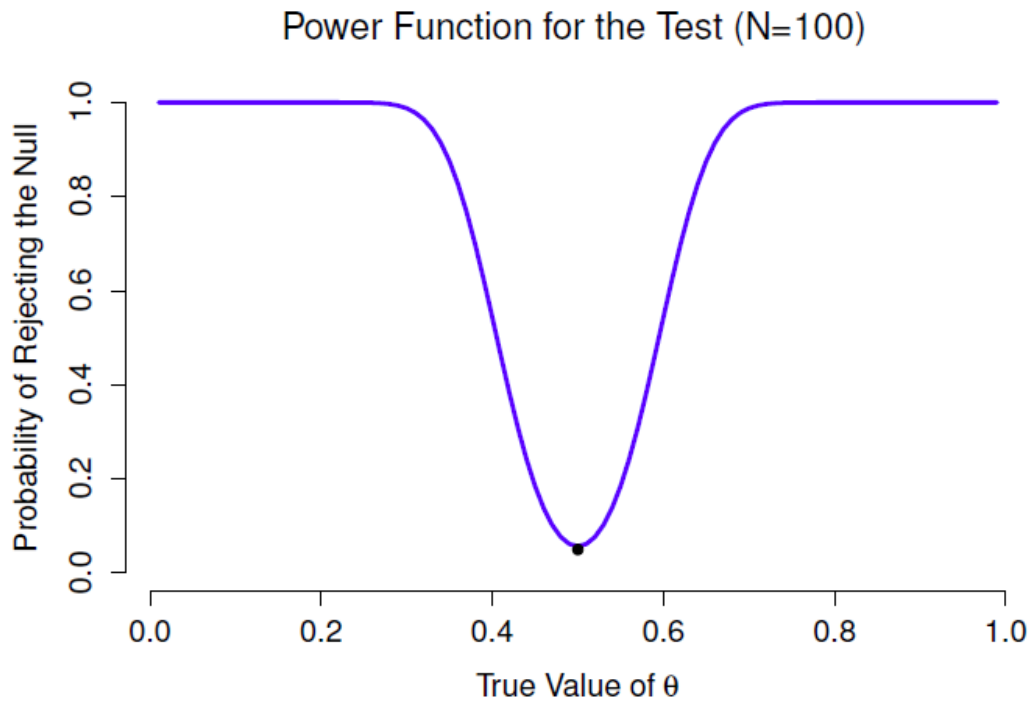


Figure 9-7 : La probabilité que nous rejetterons l'hypothèse nulle, tracée en fonction de la valeur réelle de θ. Évidemment, le test est plus puissant (plus grande chance de rejet correct) si la valeur réelle de θ est très différente de la valeur spécifiée dans l'hypothèse nulle (i.e. $\theta = .5$). Notez que lorsque θ est en fait égal à .5 (représenté par un point noir), l'hypothèse nulle est en fait vraie et rejeter l'hypothèse nulle dans ce cas serait une erreur de type I.

Taille de l'effet

Puisque tous les modèles sont faux, le scientifique doit être attentif à ce qui est faux. Il est inapproprié de s'inquiéter des souris quand il y a des tigres à la maison. - George Box (Box, George E. P. Box (1976), p. 792)

Le graphique de la Figure 9-7 illustre un point assez élémentaire de la vérification des hypothèses. Si l'état réel du monde est très différent de ce que l'hypothèse nulle prédit, alors votre pouvoir sera très élevé, mais si l'état réel du monde est similaire à l'état nul (mais pas identique), la puissance du test sera très faible. Il est donc utile d'avoir un moyen de quantifier à quel point l'état réel du monde est « semblable » à l'hypothèse nulle. Une statistique qui le fait s'appelle une mesure de **la taille de l'effet** (p. ex. Cohen, Cohen (1988) ; Ellis, Ellis (2010)). La taille de l'effet est définie légèrement différemment selon les contextes (c'est pourquoi la présente section ne fait que parler en termes généraux), mais l'idée qualitative qu'elle tente de saisir est toujours la même. Quelle est l'ampleur de la différence entre les paramètres *réels* de la population et les valeurs des paramètres qui sont assumés par l'hypothèse nulle ? Dans notre exemple de la PES, si nous laissons $\theta_0 = 0.5$ indiquer la valeur assumée par l'hypothèse nulle et laissons θ indiquer la vraie valeur, alors une simple mesure de l'ampleur de l'effet pourrait être quelque chose comme la différence entre la vraie valeur et l'hypothèse nulle (c'est-à-dire $\theta - \theta_0$), ou peut-être simplement l'ampleur de cette différence, $abs(\theta - \theta_0)$

Tableau 9-2 : Un guide sommaire pour comprendre la relation entre la signification statistique et la valeur de l'effet. Fondamentalement, si vous n'avez pas de résultat significatif, l'ampleur de l'effet n'a pas beaucoup de sens parce que vous n'avez aucune preuve qu'il est même réel. D'un autre côté, si vous avez un effet significatif mais que votre taille d'effet est petite, il y a de fortes chances que votre résultat (bien que réel) ne soit pas très intéressant. Cependant, ce guide est très grossier. Cela dépend beaucoup de ce que vous étudiez exactement. Les petits effets peuvent être d'une importance pratique considérable dans certaines situations. Alors ne prenez pas cette table trop au sérieux. C'est un guide approximatif, au mieux.

	taille du grand effet	faible ampleur de l'effet
résultat significatif	la différence est réelle, et d'importance pratique	la différence est réelle, mais pourrait ne pas être intéressant
résultat non significatif	aucun effet observé	aucun effet observé

Pourquoi calculer la taille de l'effet ? Supposons que vous avez mené votre expérience, recueilli les données et obtenu un effet significatif lorsque vous avez effectué votre test d'hypothèse. Ne suffit-il pas de dire que vous avez eu un effet significatif ? C'est le *but* des tests d'hypothèse, non ? Enfin, en quelque sorte. Oui, le but d'un test d'hypothèse est d'essayer de démontrer que l'hypothèse nulle est fausse, mais ce n'est pas la seule chose qui nous intéresse. Si l'hypothèse nulle affirmait que $\theta = .5$ et nous montrons qu'elle n'est pas correcte, mais nous n'avons vraiment raconté que la moitié des choses. Rejeter l'hypothèse nulle implique que nous croyons que $\theta \neq .5$, mais il y a une grande différence entre $\theta = .51$ et $\theta = .8$. Si nous trouvons que $\theta = .8$, alors non seulement nous avons trouvé que l'hypothèse nulle est fausse, mais elle semble être *très* fausse. Par contre, supposons que nous ayons réussi à rejeter l'hypothèse nulle, mais il semble que la vraie valeur de θ ne soit que de .51 (ce qui ne serait possible qu'avec une très grande étude). Bien sûr, l'hypothèse nulle est fausse, mais il n'est pas du tout sûr que nous ayons à nous en soucier parce que la taille de l'effet est si petite. Dans le contexte de mon étude sur la PES, nous pourrions encore nous en soucier puisque toute démonstration de vrais pouvoirs psychiques serait en fait plutôt cool[61], mais dans d'autres contextes, une différence de 1% n'est généralement pas très intéressante, même si c'est une vraie différence. Supposons, par exemple, que nous examinions les différences dans les résultats aux examens du secondaire entre les garçons et les filles et qu'il s'avère que les résultats des filles sont en moyenne 1 % plus élevés que ceux des garçons. Si j'ai des données provenant de milliers d'étudiants, cette différence sera presque certainement *statistiquement significative*, mais peu importe à quel point la valeur *p* est faible, ce n'est pas très intéressant. Vous ne voudriez pas affirmer que l'éducation des

[61] Bien qu'en pratique, une très petite taille de l'effet soit inquiétante parce que même des défauts méthodologiques mineurs peuvent être responsables de l'effet, et qu'en pratique aucune expérience n'est parfaite, il y a toujours des problèmes méthodologiques à prendre en compte.

garçons est en crise sur la base d'une si petite différence, n'est-ce pas ? C'est pour cette raison qu'il est de plus en plus courant (lentement, mais sûrement) de rapporter une mesure standard de la taille de l'effet avec les résultats du test d'hypothèse. Le test d'hypothèse lui-même vous dit si vous devez croire que l'effet que vous avez observé est réel (c.-à-d., pas seulement dû au hasard), alors que l'ampleur de l'effet vous dit si vous devez ou non vous y accorder de l'importance.

Augmenter la puissance de votre étude

Il n'est pas surprenant que les scientifiques soient assez obsédés par la maximisation de la puissance de leurs expériences. Nous voulons que nos expériences fonctionnent et nous voulons donc maximiser les chances de rejeter l'hypothèse nulle si elle est fausse (et bien sûr nous voulons généralement croire qu'elle est fausse !). Comme nous l'avons vu, l'un des facteurs qui influencent la puissance est la taille de l'effet. La première chose que vous pouvez faire pour augmenter votre puissance est donc d'augmenter la taille de l'effet. Dans la pratique, cela signifie que vous voulez concevoir votre étude de manière à ce que la taille de l'effet soit agrandie. Par exemple, dans mon étude sur la PES, je pourrais croire que les pouvoirs psychiques fonctionnent mieux dans une pièce calme et sombre avec moins de distractions pour perturber l'esprit. C'est pourquoi j'essaierais de mener mes expériences dans un tel environnement. Si je peux renforcer les capacités de PES des gens d'une manière ou d'une autre, alors la vraie valeur de θ augmentera[62] et donc ma taille d'effet sera plus grande. En bref, un design expérimental intelligent est un moyen d'augmenter la puissance, car il peut modifier la taille de l'effet.

Malheureusement, il arrive souvent que même avec les meilleures conceptions expérimentales, vous n'ayez qu'un effet minime. Peut-être, par exemple, l'ESP existe vraiment, mais même dans les meilleures conditions, elle est très très faible. Dans ces circonstances, le meilleur moyen d'augmenter la puissance est d'augmenter la taille de l'échantillon. En général, plus vous avez d'observations disponibles, plus il est probable que vous puissiez faire la distinction entre deux hypothèses. Si j'avais fait mon expérience sur la PES avec 10 participants et que 7 d'entre eux avaient deviné correctement la couleur de la carte cachée, vous ne seriez pas terriblement impressionné. Mais si je l'avais fait avec 10 000 participants, et que 7 000 d'entre eux avaient obtenu la bonne réponse, vous auriez beaucoup plus tendance à penser que j'avais découvert quelque chose. En d'autres termes, la puissance augmente avec la taille de l'échantillon. Ceci est illustré dans la Figure 9-8, qui montre la puissance du test pour un paramètre vrai de $\theta = 0.7$ pour toutes les tailles d'échantillon N de 1 à 100, où je suppose que l'hypothèse nulle prédit que $\theta_0 = 0.5$.

---

[62] Notez que le vrai paramètre de population θ ne correspond pas nécessairement à un fait immuable de la nature. Dans ce contexte, θ n'est que la véritable probabilité que les gens devinent correctement la couleur de la carte dans l'autre pièce. En tant que tel, le paramètre de population peut être influencé par toutes sortes de choses. Bien sûr, tout ceci est basé sur l'hypothèse que la PES existe réellement !

Parce que la puissance est importante, chaque fois que vous envisagez de faire une expérience, il serait très utile de savoir quel niveau de puissance vous êtes susceptible d'avoir. Il n'est jamais possible d'en être sûr car il est impossible de connaître la taille réelle de votre effet. Cependant, il est souvent (enfin, parfois) possible de deviner sa taille. Si oui, vous pouvez deviner la taille de l'échantillon dont vous avez besoin ! Cette idée s'appelle l'**analyse de puissance**, et si c'est faisable, alors c'est très utile. Il peut vous dire si vous avez assez de temps ou d'argent pour mener à bien l'expérience. Il est de plus en plus courant de voir des gens affirmer que l'analyse de puissance devrait faire partie intégrante de la conception expérimentale, alors cela vaut la peine d'en savoir plus. Je ne parle pas de l'analyse de puissance dans ce livre, cependant. C'est en partie pour une raison ennuyeuse et en partie pour une raison de fond. La raison ennuyeuse est que je n'ai pas encore eu le temps d'écrire sur l'analyse de puissance. Le plus important, c'est que je me méfie encore un peu de l'analyse de puissance. En tant que chercheur, je me suis très rarement trouvé en mesure de le faire. Soit (a) mon expérience est une expérience un peu non standard et je ne sais pas comment définir correctement la taille de l'effet, ou (b) j'ai littéralement si peu d'idée sur la taille de l'effet que je ne saurais pas comment interpréter les réponses.

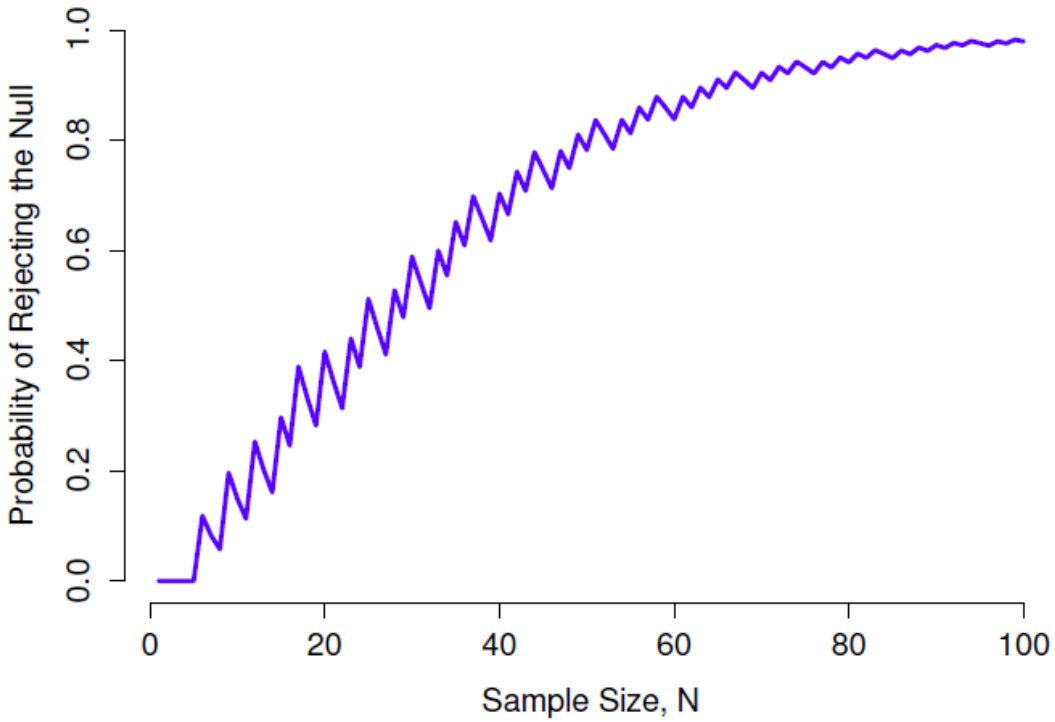


Figure 9-8 : La puissance de notre test tracée en fonction de la taille de l'échantillon N. Dans ce cas, la valeur réelle de θ est 0,7 mais l'hypothèse nulle est que $\theta = 0{,}5$. Dans l'ensemble, plus N est grand, plus la puissance est grande. (Les petits zig-zags dans cette fonction se produisent à cause de quelques interactions étranges entre $theta$, α et le fait que la distribution binomiale est discrète, cela ne pose pas de problème pour notre propos).

En plus de cela, après de longues conversations avec quelqu'un qui est consultant en statistiques pour gagner sa vie (ma femme, en l'occurrence), je ne peux m'empêcher de

remarquer qu'en pratique, le *seul* moment où quelqu'un lui demande de faire une analyse de puissance c'est quand elle aide quelqu'un à rédiger une demande de subvention. En d'autres termes, le seul moment où un scientifique semble vouloir une analyse de puissance dans la vraie vie, c'est lorsqu'il est forcé de le faire par un processus bureaucratique. Ça ne fait partie du travail quotidien de personne. Bref, j'ai toujours été d'avis que même si la puissance est un concept important, l'*analyse de puissance* n'est pas aussi utile que les gens le disent, sauf dans les rares cas où (a) quelqu'un a trouvé comment calculer la puissance pour votre plan expérimental réel et (b) vous avez une assez bonne idée de la taille de l'effet probable.[63] Peut-être que d'autres personnes ont eu de meilleures expériences que moi, mais personnellement je n'ai jamais été dans une situation où les deux (a) et (b) étaient vrais. Peut-être que je serai convaincu du contraire à l'avenir, et probablement qu'une prochaine version de ce livre inclurait une discussion plus détaillée de l'analyse de puissance, mais pour l'instant, c'est à peu près tout ce que je suis capable de dire sur le sujet.

Quelques points à prendre en considération

Ce que je vous ai décrit dans ce chapitre est le cadre orthodoxe du test de signification des hypothèses nulles (Null hypothesis significance testing : NHST). Comprendre le fonctionnement des NHST est une nécessité absolue parce qu'il s'agit de l'approche dominante en matière de statistiques inférentielles depuis qu'elles ont pris de l'importance au début du XXe siècle. C'est ce sur quoi la grande majorité des scientifiques se fient pour l'analyse de leurs données, donc même si vous détestez cela, vous devez le savoir. Cependant, l'approche n'est pas sans problèmes. Ce cadre a un certain nombre de bizarreries, des bizarreries historiques sur la façon dont il a vu le jour, des différends théoriques sur le bien-fondé du cadre et de nombreux pièges pratiques pour ceux qui ne sont pas prudents. Je ne vais pas entrer dans les détails à ce sujet, mais je pense qu'il vaut la peine d'aborder brièvement quelques-unes de ces questions.

Neyman contre Fisher

La première chose que vous devez savoir, c'est que la NHST orthodoxe est en fait une synthèse de deux approches assez différentes de la vérification des hypothèses, l'une proposée par Sir Ronald Fisher et l'autre par Jerzy Neyman (voir Lehmann (2011), pour un résumé historique). L'histoire est confuse parce que Fisher et Neyman étaient de vraies personnes dont les opinions ont changé au fil du temps, et à aucun moment ils n'ont offert « la publication définitive » sur la façon dont nous devrions interpréter leur travail plusieurs décennies plus tard. Cela dit, voici un bref résumé de ce que je pense de ces deux approches.

---

[63] Les chercheurs qui étudient l'efficacité d'un nouveau traitement médical et qui précisent à l'avance l'importance de l'effet à détecter, par exemple en plus de tout traitement existant, peuvent faire exception à cette règle. De cette façon, il est possible d'obtenir des informations sur la valeur potentielle d'un nouveau traitement.

Parlons d'abord de l'approche de Fisher. Pour autant que je sache, Fisher a supposé que vous n'aviez qu'une seule hypothèse (l'hypothèse nulle) et que ce que vous voulez faire est de découvrir si l'hypothèse nulle est incompatible avec les données. De son point de vue, ce qu'il faut faire, c'est vérifier si les données sont « suffisamment improbables » selon l'hypothèse nulle. En fait, si vous vous souvenez de ce que nous avons dit plus tôt, c'est ainsi que Fisher définit la *valeur p*. Selon Fisher, si l'hypothèse nulle fournissait un très mauvais résumé des données, vous pourriez la rejeter en toute sécurité. Mais, comme vous n'avez pas d'autres hypothèses pour la comparer, il n'y a aucun moyen « d'accepter l'alternative » parce que vous n'avez pas nécessairement une alternative explicitement énoncée. C'est plus ou moins tout ce qu'il y a à faire.

En revanche, Neyman pensait que la vérification des hypothèses servait de guide d'action et que son approche était un peu plus formelle que celle de Fisher. Selon lui, il y a plusieurs choses que vous pourriez *faire* (accepter l'hypothèse nulle ou accepter l'alternative) et le but du test était de vous dire laquelle est supportées par les données. De ce point de vue, il est essentiel de bien préciser votre hypothèse alternative. Si vous ne connaissez pas l'hypothèse alternative, alors vous ne savez pas quelle est la puissance du test, ni même quelle action a du sens. Son cadre exige véritablement une concurrence entre les différentes hypothèses. Pour Neyman, la valeur *p* ne mesurait pas directement la probabilité des données (ou des données plus extrêmes) sous l'hypothèse nulle, il s'agissait plutôt d'une description abstraite au sujet de laquelle les « tests possibles » vous disaient d'accepter l'hypothèse nulle ou d'accepter l'alternative.

Comme vous pouvez le constater, ce que nous avons aujourd'hui est un étrange méli-mélo des deux. Nous parlons d'avoir à la fois une hypothèse nulle et une hypothèse alternative (Neyman), mais nous définissons généralement[64] la valeur *p* en termes de données extrêmes (Fisher), alors nous avons toujours des valeurs α (Neyman). Certains des tests statistiques ont explicitement spécifié des alternatives (Neyman) mais d'autres sont assez vagues à ce sujet (Fisher). Et, selon certains au moins, nous n'avons pas le droit de parler d'accepter l'alternative (Fisher). C'est le bazar, mais j'espère au moins que ça explique pourquoi c'est le bazar.

Bayésiens contre fréquentistes

Plus tôt dans ce chapitre, j'ai insisté sur le fait qu'on *ne peut pas* interpréter la valeur *p* comme la probabilité que l'hypothèse nulle soit vraie. Le NHST est fondamentalement un outil fréquentiste (voir chapitre 7) et ne permet donc pas d'attribuer des probabilités aux hypothèses. L'hypothèse nulle est soit vraie, soit fausse. L'approche bayésienne des statistiques interprète la probabilité comme un degré de croyance, alors il est tout à fait normal de dire qu'il y a une probabilité de 10 % que l'hypothèse nulle soit vraie. Ce n'est qu'un reflet du degré de confiance que vous avez dans cette hypothèse. Vous n'êtes pas

---

[64] Bien que ce livre décrive à la fois la définition de Neyman et de Fisher de la valeur p, la plupart ne le font pas. La plupart des manuels d'introduction ne vous donneront que la version Fisher.

autorisé à le faire dans le cadre de l'approche fréquentiste. Rappelez-vous que si vous êtes un fréquentiste, une probabilité ne peut être définie qu'en fonction de ce qui se produit après un grand nombre de répétitions indépendantes (c.-à-d., une fréquence à long terme). Si c'est votre interprétation de la probabilité, parler de la « probabilité » que l'hypothèse nulle soit vraie est un charabia complet : une hypothèse nulle est soit vraie, soit fausse. Vous ne pouvez pas parler d'une fréquence à long terme pour cette affirmation. Parler de « la probabilité de l'hypothèse nulle » n'a pas plus de sens que de « la couleur de la liberté ».

Plus important encore, *il ne s'agit pas* d'une question purement idéologique. Si vous décidez que vous êtes Bayésien et que vous êtes d'accord pour faire des déclarations de probabilité au sujet des hypothèses, vous devez suivre les règles bayésiennes pour calculer ces probabilités. J'en parlerai plus en détail au chapitre 15, mais pour l'instant, ce que je tiens à vous faire remarquer, c'est que la valeur *p* est une *terrible* approximation de la probabilité que *H0* soit vrai. Si ce que vous voulez savoir est la probabilité de l'hypothèse nulle, alors la valeur *p* n'est pas celle que vous recherchez !

Pièges

Comme vous pouvez le constater, la théorie qui sous-tend les tests d'hypothèse est un fouillis, et même maintenant il y a des controverses dans les statistiques sur la façon dont cela devrait fonctionner. Cependant, les désaccords entre statisticiens ne sont pas notre véritable préoccupation ici. Notre véritable préoccupation est l'analyse pratique des données. Et bien que l'approche « orthodoxe » de la vérification de l'importance des hypothèses nulles présente de nombreux inconvénients, même un Bayésien non repentant comme moi sera d'accord pour dire qu'elles peuvent être utiles si elles sont utilisées de façon responsable. La plupart du temps, ils donnent des réponses sensées et vous pouvez les utiliser pour apprendre des choses intéressantes. Mis à part les différentes idéologies et les confusions historiques dont nous avons parlé, il n'en demeure pas moins que le plus grand danger dans toutes les statistiques est l'usage irréfléchi. Je ne parle pas de stupidité, je parle littéralement d'insouciance. La hâte d'interpréter un résultat sans prendre le temps de réfléchir à ce que chaque test dit réellement au sujet des données, et de vérifier si cela correspond à la façon dont vous l'avez interprété. C'est là que se trouve le plus grand piège.

Pour en donner un exemple, prenons la situation suivante (voir Gelman et Stern (2006)). Supposons que je mène mon étude sur la PES et que j'ai décidé d'analyser les données séparément pour les participants et les participantes. Parmi les participants, 33 sur 50 ont deviné la couleur de la carte correctement. C'est un effet significatif (p = .03). Parmi les participantes, 29 sur 50 ont deviné correctement. Ce n'est pas un effet significatif (p = .32). En observant cela, il est extrêmement tentant pour les gens de commencer à se demander pourquoi il y a une différence entre les hommes et les femmes du point de vue de leurs capacités psychiques. Cependant, c'est une erreur. Si vous y réfléchissez *bien*, nous n'avons pas fait un test qui compare explicitement les hommes et les femmes. Tout ce que nous avons fait est de comparer les hommes au hasard (le test binomial était significatif) et les femmes au hasard (le test binomial était non significatif). Si nous voulons argumenter qu'il y a une différence réelle entre les hommes et les femmes, nous devrions probablement tester l'hypothèse nulle qu'il n'y a pas de différence ! Nous pouvons le faire en utilisant un

test d'hypothèse différent,[65] mais lorsque nous le faisons, il s'avère que nous n'avons aucune preuve que les hommes et les femmes sont significativement différents (p = .54). Pensez-vous qu'il y a quelque chose de fondamentalement différent entre les deux groupes ? Bien sûr que non. Ce qui s'est passé ici, c'est que les données des deux groupes (hommes et femmes) sont assez limitées. Par pur hasard, l'un d'eux s'est retrouvé du côté magique du p = .05, et l'autre ne l'a pas fait. Cela ne veut pas dire que les hommes et les femmes sont différents. Cette erreur est si courante qu'il faut toujours s'en méfier. La différence entre significatif et non significatif *n'est pas* la preuve d'une différence réelle. Si vous voulez dire qu'il y a une différence entre deux groupes, alors vous devez tester cette différence !

L'exemple ci-dessus n'est qu'un exemple. Je l'ai choisi parce qu'il s'agit d'un problème courant, mais, dans l'ensemble, l'analyse des données peut être difficile à faire correctement. Pensez à ce que vous voulez tester, pourquoi vous voulez le tester, et si oui ou non les réponses que vous voulez tester pourraient avoir un sens dans le monde réel.

Résumé

La vérification des hypothèses nulles est l'un des éléments les plus omniprésents de la théorie statistique. La grande majorité des articles scientifiques rapportent les résultats d'un test d'hypothèse ou d'un autre. En conséquence, il est presque impossible de s'en sortir en science sans avoir au moins une compréhension superficielle de ce que signifie une *valeur p*, ce qui en fait l'un des chapitres les plus importants de l'ouvrage. Comme d'habitude, je terminerai le chapitre par un bref résumé des idées clés dont nous avons parlé :

- Hypothèses de recherche et hypothèses statistiques. Hypothèses nulles et alternatives. (Section 9.1).
- Erreurs de type 1 et de type 2 (section 9.2)
- Statistiques des tests et distributions d'échantillonnage (Section 9.3)
- Les tests d'hypothèse en tant que processus décisionnel (Section 9.4)
- *p* comme décisions « douces « (Section 9.5)
- Rédaction des résultats d'un test d'hypothèse (section 9.6)
- Exécution du test d'hypothèse dans la pratique (Section 9.7)
- Ampleur et puissance de l'effet (Section 9.8)
- Quelques questions à prendre en considération concernant la vérification des hypothèses (Section 9.9)

Plus loin dans le livre, au chapitre 16, je reviendrai sur la théorie des tests d'hypothèse nulle d'un point de vue bayésien et présenterai un certain nombre de nouveaux outils que vous pouvez utiliser si vous n'aimez pas particulièrement l'approche orthodoxe. Mais pour

[65] Dans ce cas, le test du chi carré de Pearson sur l'indépendance (chapitre 10)

l'instant, nous en avons terminé avec la théorie statistique abstraite, et nous pouvons commencer à discuter d'outils spécifiques d'analyse des données.

Analyse des données catégoriques

Maintenant que nous avons couvert la théorie de base qui sous-tend les tests d'hypothèse, il est temps de commencer à examiner des tests spécifiques qui sont couramment utilisés en psychologie. Par où devrions-nous commencer ? Tous les manuels ne s'entendent pas sur les points de départ, mais je vais commencer par « les test de χ^2 » (ce chapitre, prononcé « chi-carré »[66]) et « *les tests de t* » (chapitre 11. Ces deux outils sont très fréquemment utilisés dans la pratique scientifique, et bien qu'ils ne soient pas aussi puissants que la « régression » (chapitre 12) et « l'analyse de variance » (chapitre 13), ils sont beaucoup plus faciles à comprendre.

Le terme « données catégorielles » n'est qu'un autre nom pour « données d'échelle nominale ». Ce n'est rien dont nous n'avons pas déjà discuté, c'est juste que dans le contexte de l'analyse des données, les gens ont tendance à utiliser le terme « données catégorielles » plutôt que « données à échelle nominale ». Je ne sais pas pourquoi. Dans tous les cas, l'**analyse catégorielle des données** fait référence à un ensemble d'outils que vous pouvez utiliser lorsque vos données sont sur une échelle nominale. Cependant, il existe de nombreux outils différents qui peuvent être utilisés pour l'analyse des données catégorielles, et ce chapitre ne couvre que quelques-uns des plus courants.

Le test d'adéquation χ^2 (chi carré)

Le test d'adéquation χ^2 est l'un des plus anciens tests d'hypothèse. Il a été inventé par Karl Pearson au tournant du siècle (Pearson 1900), avec quelques corrections apportées plus tard par Sir Ronald Fisher (Fisher 1922). Il vérifie si une distribution de fréquence observée d'une variable nominale correspond à une distribution de fréquence attendue. Par exemple, supposons qu'un groupe de patients a subi un traitement expérimental et que leur état de santé a été évalué pour voir si leur état s'est amélioré, s'il est demeuré le même ou s'il s'est aggravé. Un test d'adéquation pourrait être utilisé pour déterminer si les chiffres dans chaque catégorie - améliorés, sans changement, aggravés - correspondent aux chiffres auxquels on pourrait s'attendre avec l'option de traitement standard. Réfléchissons encore un peu, avec un peu de psychologie.

Les données des cartes

Au fil des ans, de nombreuses études ont montré que les humains ont de la difficulté à simuler le hasard. Si nous essayons « d'agir » au hasard, nous *pensons* en termes de modèles et de structure et donc, quand on nous demande de « faire quelque chose au hasard », ce que les gens font en réalité est tout sauf aléatoire. En conséquence, l'étude de l'aléatoire humain (ou de la non aléatoire, selon le cas) soulève beaucoup de questions psychologiques

[66] Parfois aussi appelé "chi-carré".

profondes sur la façon dont nous pensons le monde. En gardant cela à l'esprit, considérons une étude très simple. Supposons que je demande aux gens d'imaginer un jeu de cartes mélangé, et de choisir mentalement une carte dans ce jeu imaginaire « au hasard ». Après avoir choisi une carte, je leur demande d'en choisir une deuxième mentalement. Pour les deux choix, ce que nous allons regarder, c'est la couleur (coeur, trèfle, pique ou carreau) que les gens ont choisi. Après avoir demandé, disons, à *N=200* personnes de le faire, j'aimerais regarder les données et déterminer si les cartes que les gens prétendaient choisir étaient vraiment aléatoires. Les données sont contenues dans le fichier randomness.csv dans lequel, lorsque vous l'ouvrez dans Jamovi et regardez la feuille de calcul, vous verrez trois variables. Il s'agit d'une variable id qui attribue un identifiant unique à chaque participant et des deux variables choice_1 et choice_2 qui indiquent les combinaisons de cartes choisies par les participants.

Pour l'instant, concentrons-nous sur le premier choix que les gens ont fait. Nous utiliserons l'option Tableaux de fréquence sous « Exploration » - « Descriptive » pour compter le nombre de fois où nous avons observé des gens choisir chaque combinaison. Voilà ce qu'on obtient :

Trèfle	Carreau	Coeur	Pique
35	51	64	50

Ce petit tableau de fréquence est très utile. En y regardant de plus près, il y a un indice que les sujets *sont* plus enclins à choisir des cœurs que des trèfles, mais ce n'est pas tout à fait évident que ce soit réellement vrai, ou si que ce soit le fruit du hasard. Nous devrons donc probablement faire une sorte d'analyse statistique pour le découvrir, ce dont je vais vous parler dans la prochaine section.

Très bien. À partir de maintenant, nous traiterons ce tableau comme les données que nous cherchons à analyser. Cependant, comme je vais devoir parler de ces données en termes mathématiques (désolé !), ce serait bien d'être clair sur la notation. En notation mathématique, on raccourcit le mot « observé » lisible par l'homme par la lettre *O*, et on utilise des indices pour indiquer la position de l'observation. Ainsi, la deuxième observation de notre tableau s'écrit O_2 en mathématiques. La relation entre les descriptions en langage nature et les symboles mathématiques est illustrée ci-dessous :

Intitulé	indice, *i*	symbole mathématique	la valeur
Trèfles	1	O_1	35
Carreaux	2	O_2	51
Cœurs	3	O_3	64
Piques	4	O_4	50

J'espère que c'est assez clair. Il est également intéressant de noter que les mathématiciens préfèrent parler de choses générales plutôt que spécifiques, de sorte que vous verrez aussi

la notation O_i, qui se réfère au nombre d'observations qui entrent dans la *i-ème* catégorie (où *i* pourrait être 1, 2, 3 ou 4). Enfin, si l'on veut se référer à l'ensemble des fréquences observées, les statisticiens regroupent toutes les valeurs observées dans un vecteur[67], que je vais appeler *O*.

$$O = (O_1, O_2, O_3, O_4)$$

Encore une fois, Il n'y a rien de nouveau ou de particulier. C'est juste une notation. Si je dis que *O=(35, 51 ,64,50)* tout ce que je fais est de décrire le tableau des fréquences observées (c'est-à-dire observées), mais je m'y réfère en utilisant la notation mathématique.

L'hypothèse nulle et l'hypothèse alternative

Comme l'indique la dernière section, notre hypothèse de recherche est que « les gens ne choisissent pas les cartes au hasard ». Ce que nous allons maintenant vouloir faire, c'est traduire cela en hypothèses statistiques, puis construire un test statistique de ces hypothèses. Le test que je vais vous décrire est le **test d'adéquation de Pearson** χ^2 (chi carré), et comme c'est souvent le cas, nous devons commencer par construire soigneusement notre hypothèse nulle. Dans ce cas, c'est assez facile. Tout d'abord, énonçons l'hypothèse nulle avec des mots :

H_0: Les quatre couleurs sont choisies avec une probabilité égale

Maintenant, comme il s'agit de statistiques, nous devons pouvoir dire la même chose d'une manière mathématique. Pour ce faire, utilisons la notation P_J pour faire référence à la véritable probabilité que la *j-ième* couleur soit choisie. Si l'hypothèse nulle est vraie, alors chacune des quatre couleurs a 25% de chance d'être sélectionnée. En d'autres termes, notre hypothèse nulle prétend que $P_1=.25$, $P_2=.25$, $P_3=.25$ et enfin que $P_4=.25$. Cependant, de la même manière que nous pouvons regrouper nos fréquences observées dans un vecteur *O* qui résume l'ensemble des données, nous pouvons utiliser *P* pour nous référer aux probabilités qui correspondent à notre hypothèse nulle. Donc si je laisse le vecteur $P=(P_1, P_2, P_3, P_4)$ se référer à l'ensemble des probabilités qui décrivent notre hypothèse nulle, alors nous avons :

$$H_0: P = (.25, .25, .25, .25)$$

Dans ce cas particulier, notre hypothèse nulle correspond à un vecteur de probabilités *P* dans lequel toutes les probabilités sont égales entre elles. Mais ce n'est pas forcément le cas. Par exemple, si la tâche expérimentale consistait pour les gens à imaginer qu'ils tiraient des cartes d'un jeu qui avait deux fois plus de trèfles que toute autre couleur, alors l'hypothèse nulle correspond à quelque chose comme *P= (.4, .2, .2, .2)*. Tant que les probabilités sont toutes des nombres positifs, et qu'elles totalisent toutes 1, alors c'est un choix parfaitement légitime pour l'hypothèse nulle. Toutefois, l'utilisation la plus courante du test de la qualité

[67] Un vecteur est une séquence d'éléments de données du même type basique

de l'ajustement consiste à vérifier une hypothèse nulle selon laquelle toutes les catégories sont également probables, et nous nous en tiendrons donc à cela pour notre exemple.

Et notre hypothèse alternative, H_1 ? Tout ce qui nous intéresse vraiment, c'est de démontrer que les probabilités en jeu ne sont pas toutes identiques (c'est-à-dire que les choix des gens n'étaient pas complètement aléatoires). En conséquence, les versions « humaine » de nos hypothèses ressemblent à ceci :

- H_0 : Les quatre couleurs sont choisies avec une probabilité égale
- H_1 : Au moins un des choix de la combinaison *n'a pas une probabilité de* 0,25

et la version « mathématique » est :

$$H_0: P = (.25, .25, .25, .25)$$
$$H_0: P \neq (.25, .25, .25, .25)$$

Le test statistique d'ajustement

A ce stade, nous avons nos fréquences observées *O* et un ensemble de probabilités *P* correspondant à l'hypothèse nulle que nous voulons tester. Ce que nous voulons maintenant faire, c'est construire un test de l'hypothèse nulle. Comme toujours, si nous voulons tester H_0 contre H_1, nous allons avoir besoin d'une statistique de test. L'astuce de base d'un test d'adéquation consiste à construire une statistique de test qui mesure à quel point les données sont « proches » de l'hypothèse nulle. Si les données ne ressemblent pas à ce à quoi vous « vous attendriez » si l'hypothèse nulle était vraie, alors ce n'est probablement pas vrai. Bien, si l'hypothèse nulle est vraie, qu'est-ce qu'on s'attendrait à voir ? Ou, pour employer la terminologie correcte, quelles sont les **fréquences attendues**. Il y a *N=200* observations, et (si l'hypothèse nulle est vraie) la probabilité que l'un d'eux choisisse un coeur est $P_3=.25$, donc on peut supposer qu'on attend *200 X 0,25=50 cœurs*. Ou, plus précisément, si nous laissons E_i référer au « nombre de réponses de catégorie *i* que nous attendons si l'hypothèse nulle est vraie », alors :

$$E_i = N \times P_i$$

S'il y a 200 observations qui peuvent être classées dans quatre catégories, et que nous pensons que les quatre catégories sont également probables, alors en moyenne nous nous attendrions à voir 50 observations dans chaque catégorie.

Maintenant, comment traduire cela en une statistique de test ? De toute évidence, ce que nous voulons faire, c'est comparer le nombre d'observations *attendues* dans chaque catégorie (E_i) avec le nombre d'observations *observées* dans cette catégorie (O_i). Puis sur la base de cette comparaison, nous devrions être en mesure d'établir une bonne statistique de test. Pour commencer, calculons la différence entre ce que nous attendions sous l'hypothèse nulle et ce que nous avons réellement trouvé. C'est-à-dire que nous calculons le score de différence « observé moins attendu », (O_i-E_i). Ceci est illustré dans le tableau suivant :

		Trèfle	Carreau	Coeur	Pique
fréquence prévue	E_i	50	50	50	50

fréquence observée	O_i	35	51	64	50
score de différence	O_i-E_i	15	-1	-14	0

D'après nos calculs, il est clair que les gens ont choisi plus de cœurs et moins de trèfles que ne le prévoyait l'hypothèse nulle. Cependant, un moment de réflexion suggère que ces différences brutes ne sont pas tout à fait ce que nous recherchons. Intuitivement, on a l'impression que c'est aussi mauvais quand l'hypothèse nulle prédit trop peu d'observations (ce qui est arrivé avec les cœurs) que quand elle prédit trop d'observations (ce qui est arrivé avec les trèfles). C'est donc un peu bizarre que nous ayons un nombre négatif pour les carreaux et un nombre positif pour les trèfles. Un moyen facile de résoudre ce problème est de tout élever au carré, de sorte que nous calculons maintenant les différences au carré, $(E_i - Oi)^2$. Comme auparavant, nous pouvons le faire à la main :

$(\text{observé} - \text{attendu})^2$

Trèfle	Carreau	Cœur	Pique
225	1	196	0

Maintenant, nous faisons des progrès. Ce que nous avons maintenant, c'est une collection de chiffres qui sont gros quand l'hypothèse nulle fait une mauvaise prédiction (trèfles et cœurs), mais qui sont petits quand elle fait une bonne prédiction (carreaux et piques). Ensuite, pour des raisons techniques que j'expliquerai dans un instant, divisons aussi tous ces nombres par la fréquence *Ei* attendue, donc nous calculons actuellement $\frac{(O_i - E_i)^2}{E_i}$. Depuis E_i*=50* pour toutes les catégories dans notre exemple, ce n'est pas un calcul très intéressant, mais faisons-le quand même :

$(\text{observé} - \text{attendu})^2/\text{attendu}$

Trèfle	Carreau	Cœur	Pique
4,5	0,02	3,92	0

En effet, nous avons ici quatre scores « d'écarts » différents, chacun d'eux nous indiquant l'ampleur de « l'erreur » que l'hypothèse nulle a faite lorsque nous avons essayé de l'utiliser pour prédire nos fréquences observées. Donc, afin de convertir ces données en une statistique de test utile, une chose que nous pourrions faire, c'est d'additionner ces chiffres. Le résultat s'appelle la statistique de **la qualité de l'ajustement**, conventionnellement appelée χ^2 (chi carré) ou Goodness of fit (GOF) en anglais. On peut le calculer comme suit :

$\text{sum}((\text{observed} - \text{expected})^2/\text{expected})$

Cela nous donne une valeur de 8,44.

Si nous laissons k référer au nombre total de catégories (i.e. $k = 4$ pour les données de nos cartes), alors la statistique χ^2 est donnée par :

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Intuitivement, il est clair que si χ^2 est petit, alors les données observées O_i sont très proches de ce que l'hypothèse nulle prédisait E_i, donc nous allons avoir besoin d'une grande statistique χ^2 afin de rejeter la nulle. Comme nous l'avons vu lors de nos calculs, dans notre jeu de données de cartes nous avons une valeur de $\chi^2 = 8.44$. La question est donc maintenant de savoir si c'est une valeur assez grande pour rejeter l'hypothèse nulle.

La distribution d'échantillonnage de la statistique d'ajustement

Pour déterminer si une valeur particulière de χ^2 est suffisamment grande pour justifier le rejet de l'hypothèse nulle, nous devrons déterminer quelle serait la distribution d'échantillonnage pour χ^2 si l'hypothèse nulle est vraie. C'est donc ce que je vais faire dans cette section. Je vais vous montrer en détail comment cette distribution d'échantillonnage est construite, puis, dans la section suivante, l'utiliser pour construire un test d'hypothèse. Si vous voulez aller droit au but et que vous êtes prêt à croire que la distribution d'échantillonnage est une **distribution** χ^2 (chi carré) avec *k-1* degrés de liberté, vous pouvez sauter le reste de cette section. Cependant, si vous voulez comprendre *pourquoi* le test d'ajustement fonctionne de cette façon, lisez ce qui suit.

Supposons que l'hypothèse nulle soit vraie. Si c'est le cas, la véritable probabilité qu'une observation tombe dans la *i-ème* catégorie est P_i. Après tout, c'est à peu près la définition de notre hypothèse nulle. Réfléchissons à ce que cela signifie vraiment. C'est un peu comme dire que c'est la « nature » qui décide si l'observation appartient ou non à la catégorie *i* en lançant une pièce pondérée (c'est-à-dire une pièce dont la probabilité de tomber sur face est P_j. Nous pouvons ainsi penser à notre fréquence O_j observée en imaginant que la nature a retourné N de ces pièces (une pour chaque observation dans l'ensemble de données), et exactement O_i d'entre elles sont tombées sur face. Évidemment, c'est une façon assez bizarre de penser à l'expérience. Mais le résultat (je l'espère), c'est vous rappeler que nous avons déjà vu ce scénario auparavant. C'est exactement la même configuration qui a donné lieu à la distribution binomiale de la Section 7.4. En d'autres termes, si l'hypothèse nulle est vraie, il s'ensuit que nos fréquences observées ont été générées par échantillonnage à partir d'une distribution binomiale :

$$O_i \sim Binomiale(P_i, N)$$

Maintenant, si vous vous souvenez de notre discussion sur le théorème de la limite centrale (Section 8.3.3), la distribution binomiale commence à ressembler à peu près à la distribution normale, surtout quand N est grand et quand P_i n'est pas *trop* proche de 0 ou 1. En d'autres termes, tant que $N \times P_i$ est assez grand. Ou, en d'autres termes, lorsque la O_i est normalement distribué alors c'est aussi le cas de $(O_i - E_i)/\sqrt{E_i}$. Puisque E_i est une valeur fixe, soustraire E_i et diviser par $\sqrt{E_i}$ change la moyenne et l'écart-type de la distribution normale, mais c'est tout. Voyons maintenant ce qu'est notre statistique d'ajustement. Ce que nous faisons, c'est prendre un ensemble de choses qui sont normalement distribuées, les mettre au carré et les additionner. Un instant ! On a déjà vu ça aussi ! Comme nous l'avons

vu à la section 7.6, lorsque vous prenez une série de choses qui ont une distribution normale standard (c.-à-d. la moyenne 0 et l'écart-type 1), que vous les élevez au carré puis que vous les additionnez, la quantité résultante a une distribution de chi carré. Nous savons maintenant que l'hypothèse nulle prédit que la distribution d'échantillonnage de la statistique d'ajustement est une distribution du chi carré. Cool.

Il y a un dernier détail dont il faut parler, à savoir les degrés de liberté. Si vous vous souvenez de la section 7.6, j'ai dit que si le nombre de valeurs que vous additionnez est *k*, alors les degrés de liberté pour la distribution du chi carré résultante sont *k*. Pourtant, ce que j'ai dit au début de cette section est que les degrés de liberté réels pour le test d'ajustement du chi carré sont *k - 1*. Pourquoi ? La réponse ici est que ce que nous sommes censés examiner, c'est le nombre de valeur vraiment *indépendantes* qui s'additionnent. Et, comme je vais continuer à en parler dans la section suivante, même s'il y a *k* valeurs que nous ajoutons seulement *k- 1* d'entre elles sont vraiment indépendantes, et donc les degrés de liberté sont en fait seulement *k-1*. C'est le sujet de la section suivante.[68]

Degrés de liberté

[68] Si vous réécrivez l'équation pour la statistique d'ajustement comme une somme sur *k -1* valeurs indépendantes, vous obtenez la distribution d'échantillonnage « correcte », qui est le chi carré avec *k - 1* degrés de liberté. C'est au-delà de la portée d'un livre d'introduction que de montrer les développements mathématiques avec autant de détails. Tout ce que je voulais faire, c'est vous donner une idée de la raison pour laquelle la statistique d'ajustement est associée à la distribution du chi carré.

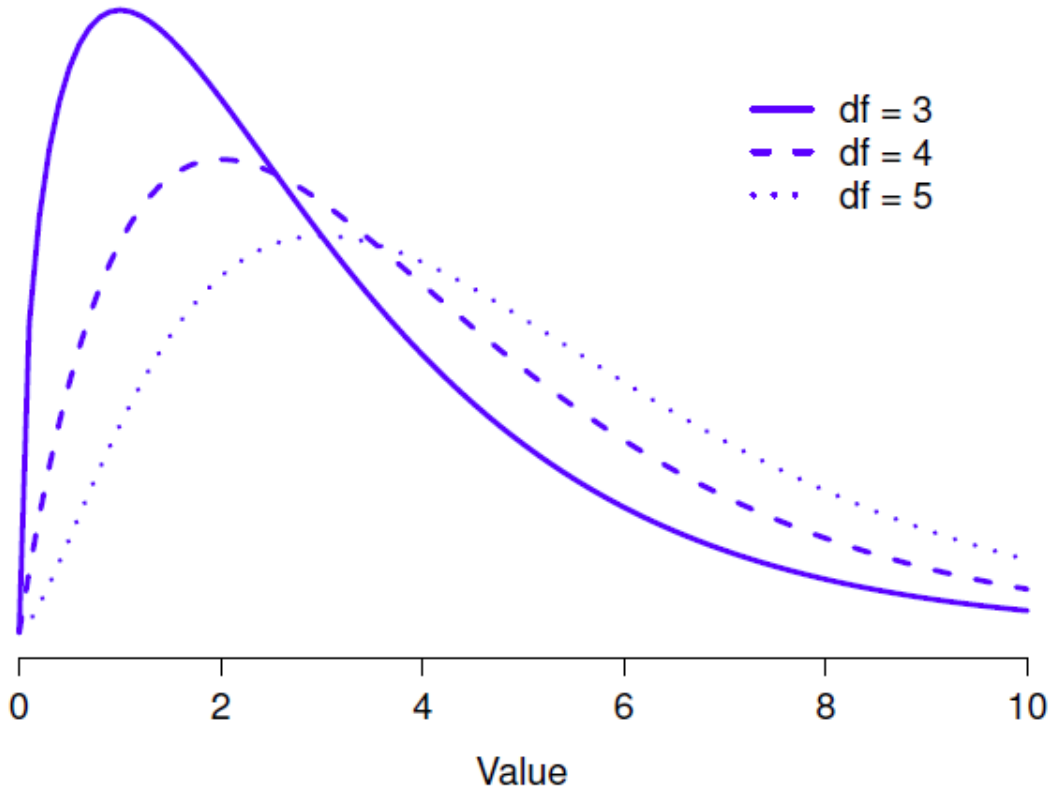


Figure 10-1 : distributions χ^2 (chi-carré) pour les différentes valeurs des « degrés de liberté ».

Lorsque j'ai introduit la distribution du khi-carré dans la section 7.6, j'étais un peu vague sur ce que *signifient* réellement les « **degrés de liberté** ». De toute évidence, c'est important. En regardant la Figure 10-1, vous pouvez voir que si nous changeons les degrés de liberté, alors la distribution du chi carré change de forme de façon assez substantielle. Mais qu'est-ce *que c'est* exactement ? Encore une fois, quand j'ai présenté la distribution et expliqué sa relation avec la distribution normale, j'ai offert une réponse : c'est le nombre de « valeurs normalement distribuées » que j'élève au carré et que j'additionne. Mais, pour la plupart des gens, c'est un peu abstrait et pas tout à fait utile. Ce que nous devons vraiment faire, c'est essayer de comprendre les degrés de liberté en termes de données. Alors c'est parti.

L'idée de base derrière les degrés de liberté est assez simple. Vous le calculez en comptant le nombre de « quantités » distinctes utilisées pour décrire vos données et en soustrayant ensuite toutes les « contraintes » que ces données doivent satisfaire.[69] C'est un peu vague,

---

[69] Je me sens obligé de souligner qu'il s'agit d'une simplification excessive. Cela fonctionne très bien dans de nombreuses situations, mais de temps en temps, nous rencontrons des valeurs de degrés de liberté qui ne sont pas des nombres entiers. Ne vous inquiétez pas trop ; quand vous rencontrez cela, rappelez-vous simplement que "degrés de liberté" est en fait

alors utilisons les données de nos cartes comme exemple concret. Nous décrivons nos données à l'aide de quatre chiffres, O_1, O_2, O_3 et O_4 correspondant aux fréquences observées dans les quatre catégories différentes (cœurs, trèfles, carreaux, piques). Ces quatre chiffres sont les *résultats aléatoires* de notre expérience. Mais mon expérience comporte en fait une contrainte fixe : la taille de l'échantillon N.[70] C'est-à-dire que si nous savons combien de personnes ont choisi des cœurs, combien ont choisi des carreaux et combien ont choisi des trèfles, alors nous serons en mesure de déterminer exactement combien ont choisi des piques. En d'autres termes, bien que nos données soient décrites à l'aide de quatre chiffres, elles ne correspondent en fait qu'à 4 - 1 = 3 degrés de liberté. Une façon légèrement différente d'y penser est de remarquer qu'il y a quatre *probabilités* qui nous intéressent (encore une fois, correspondant aux quatre catégories différentes), mais ces probabilités doivent s'additionner en une, ce qui impose une contrainte. Les degrés de liberté sont donc 4 - 1 = 3. Qu'on veuille y penser en termes de fréquences observées ou en termes de probabilités, la réponse est la même. En général, lorsque l'on effectue le test d'adéquation χ^2 (chi carré) pour une expérience impliquant k groupes, les degrés de liberté seront k- *1*.

Vérification de l'hypothèse nulle

L'étape finale dans le processus de construction de notre test d'hypothèse est de déterminer ce qu'est la région de rejet. C'est-à-dire, quelles valeurs de χ^2 nous amèneraient à rejeter l'hypothèse nulle. Comme nous l'avons vu précédemment, les grandes valeurs de χ^2 impliquent que l'hypothèse nulle a mal prédit les données de notre expérience, alors que les petites valeurs de χ^2 impliquent qu'elle est en fait assez bien faite. Par conséquent, une stratégie assez raisonnable serait de dire qu'il y a une valeur critique telle que si χ^2 est supérieur à la valeur critique, nous rejetons l'hypothèse nulle, mais si χ^2 est inférieur à cette valeur, nous conservons la valeur nulle. En d'autres termes, pour reprendre les termes que nous avons utilisés au chapitre 9, le test du chi carré est toujours **un test unilatéral**. Tout ce qu'on a à faire, c'est de trouver quelle est cette valeur critique. Et c'est assez simple. Si nous voulons que notre test ait un niveau de signification de $\alpha = .05$ (c'est-à-dire que nous sommes prêts à tolérer un taux d'erreur de type I de 5 %), alors nous devons choisir

---

un concept un peu foullis, et que la belle présentation simple que je vous fais ici n'est pas la présentation entière. Pour un cours d'introduction, il est généralement préférable de s'en tenir à la présentation simple, mais je pense qu'il est préférable de vous avertir que cette présentation simple va s'effondrer. Si je ne vous avais pas donné cet avertissement, vous pourriez commencer à vous tromper en voyant df = 3,4 ou quelque chose comme ça, pensant (à tort) que vous aviez mal compris ce que je vous ai appris plutôt que de réaliser (correctement) qu'il y a quelque chose que je ne vous ai jamais dit.

[70] Dans la pratique, la taille de l'échantillon n'est pas toujours fixe. Par exemple, nous pourrions mener l'expérience sur une période de temps fixe et le nombre de participants dépend du nombre de personnes qui se présentent. Cela n'a pas d'importance pour les objectifs actuels.

notre valeur critique de sorte qu'il n'y ait que 5 % de chances que χ^2 soit aussi grand si l'hypothèse nulle est vraie. C'est ce qu'illustre la Figure 10-2.

Vous vous demandez, comment puis-je trouver la valeur critique d'une distribution du chi carré avec *k-1* degrés de liberté ? Il y a de nombreuses années, lorsque j'ai suivi pour la première fois un cours de psychologie statistique, nous avions l'habitude de rechercher ces valeurs critiques dans un livre de tableaux des valeurs critiques, comme celui de la Figure 10-3. En regardant cette figure, nous pouvons voir que la valeur critique pour une distribution χ^2 avec 3 degrés de et *p=.05* est 7,815.

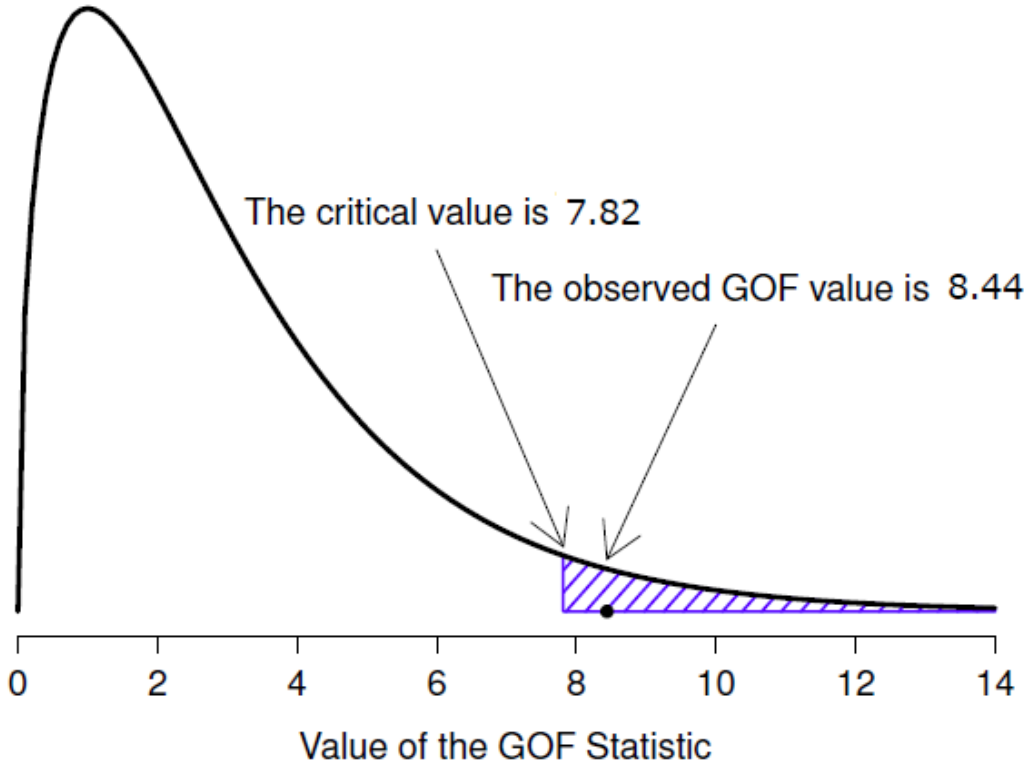


Figure 10-2 : Illustration du fonctionnement du test d'hypothèse pour le test d'ajustement (Goodness of fit)χ^2 (chi carré).

Ainsi, si notre statistique calculée χ^2 est supérieure à la valeur critique de 7,815, alors nous pouvons rejeter l'hypothèse nulle (souvenez-vous que l'hypothèse nulle, K_0, est que les quatre couleurs sont choisies avec une probabilité égale). Puisque nous avons déjà calculé qu'avant (i.e., $\chi^2 = 8.44$) nous pouvons rejeter l'hypothèse nulle. Et c'est tout, en gros. Vous connaissez maintenant le « test d'ajustement de Pearson χ^2 ». Vous avez de la chance.

Faire le test dans Jamovi

Comme il fallait s'y attendre, Jamovi fournit une analyse qui fera ces calculs pour vous. Dans la barre d'outils principale « Analyses », sélectionnez « Frequencies » - « One Sample Proportion Tests » - « N outcomes ». Ensuite, dans la fenêtre d'analyse qui apparaît,

déplacez la variable que vous voulez analyser (choix 1 dans la case « Variable »). Puis, cliquez sur la case à cocher « Expected counts » (effectifs attendus) pour qu'ils soient affichés dans le tableau des résultats. Lorsque vous avez fait tout cela, vous devriez voir les résultats de l'analyse dans Jamovi comme dans la Figure 10-4. Il n'est pas surprenant que Jamovi fournisse les mêmes chiffres et statistiques que ceux que nous avons calculés à la main ci-dessus, avec une valeur de 8,44 sur χ^2 avec 3 df[71] et *p=0,038*. Notez que nous n'avons plus besoin de chercher une valeur seuil critique de p-value, car Jamovi donne la *valeur p* réelle du χ^2 calculé pour 3 df.

Degrees of Freedom	**Probability**								
	0.95	**0.90**	**0.70**	**0.50**	**0.30**	**0.10**	**0.05**	**0.01**	**0.001**
1	0.004	0.016	0.148	0.455	1.074	2.706	3.841	6.635	10.828
2	0.103	0.211	0.713	1.386	2.408	4.605	5.991	9.210	13.816
3	0.352	0.584	1.424	2.366	3.665	6.251	7.815	11.345	16.266
4	0.711	1.064	2.195	3.357	4.878	7.779	9.488	13.277	18.467
5	1.145	1.610	3.000	4.351	6.064	9.236	11.070	15.086	20.515
6	1.635	2.204	3.828	5.348	7.231	10.645	12.592	16.812	22.458
7	2.167	2.833	4.671	6.346	8.383	12.017	14.067	18.475	24.322
8	2.733	3.490	5.527	7.344	9.524	13.362	15.507	20.090	26.124
9	3.325	4.168	6.393	8.343	10.656	14.684	16.919	21.666	27.877
10	3.940	4.865	7.267	9.342	11.781	15.987	18.307	23.209	29.588
	Non-significant						**Significant**		

Figure 10-3 : Tableau des valeurs critiques pour la distribution du chi carré

[71] NdT : l'abréviation française de degré de liberté est ddl, mais nous conservons l'abréviation anglaise df (degree of freedom) qui est utilisée dans les sorties Jamovi.

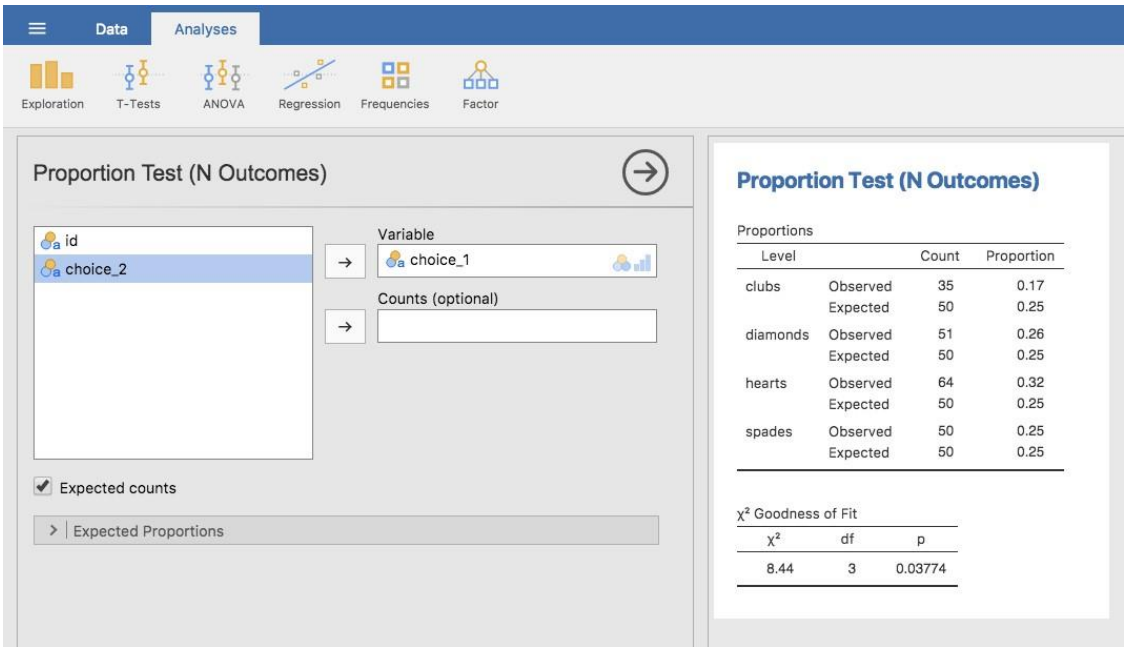


Figure 10-4 : Un test d'ajustement unilatéral χ^2 (One Sample Proportion Test) dans Jamovi, avec tableau montrant les fréquences et proportions observées et attendues.

Spécification d'une hypothèse nulle différente

À ce stade, vous vous demandez peut-être ce qu'il faut faire si vous voulez faire un test d'adéquation, mais votre hypothèse nulle n'est pas que toutes les catégories sont également probables. Supposons, par exemple, que quelqu'un ait prédit théoriquement que les gens devraient choisir des cartes rouges 60 % du temps, et des cartes noirs 40 % du temps (je ne sais pas pourquoi vous auriez prédit cela), mais qu'il n'avait aucune autre préférence. Si tel était le cas, l'hypothèse nulle serait de s'attendre à ce que 30% des choix soient des cœurs, 30% des carreaux, 20% des piques et 20% des trèfles. En d'autres termes, on s'attendrait à ce que les cœurs et les carreaux apparaissent 1,5 fois plus souvent que les piques et les trèfles (le ratio 30% :20% est le même que 1,5 :1). Cela me semble être une théorie idiote et il est assez facile de tester cette hypothèse nulle explicitement spécifiée avec les données de notre analyse Jamovi. Dans la fenêtre d'analyse (intitulée « Proportion Test (N Outcomes) » présentée dans la Figure 10-4, vous pouvez développer les options pour « Expected proportions ». Pour ce faire, vous avez la possibilité de saisir différentes valeurs de ratio pour la variable que vous avez sélectionnée, dans notre cas, il s'agit du choix 1. Modifiez le ratio pour tenir compte de la nouvelle hypothèse nulle, comme dans la Figure 10-5, et voyez comment les résultats changent et la statistique χ^2 est de 4,74, 3 df., $p = 0,182$.

Figure 10-5 : Changement des proportions attendues dans le test de la proportion d'un échantillon dans Jamovi (χ^2 One Sample Proportion)

Les nombres attendus sont maintenant :

		Trèfle	Carreau	Coeur	Pique
fréquence prévue	E_i	40	60	40	60

Maintenant, les résultats de nos hypothèses mises à jour et les fréquences attendues sont différentes de ce qu'elles étaient la dernière fois. En conséquence, notre statistique de test χ^2 est différente, et notre *valeur p* est également différente. Malheureusement, la *valeur p* est de 0,182, nous ne pouvons donc pas rejeter l'hypothèse nulle (reportez-vous à la section 9.5 pour vous rappeler pourquoi). Malheureusement, malgré le fait que l'hypothèse nulle corresponde à une théorie très stupide, ces données ne fournissent pas suffisamment de preuves contre elle.

Comment rapporter les résultats du test

Maintenant vous savez comment le test fonctionne, et vous savez comment faire le test à l'aide d'une merveilleuse baguette magique Jamovi. La prochaine chose que vous devez savoir, c'est comment rédiger les résultats. Après tout, il ne sert à rien de concevoir et de réaliser une expérience, puis d'analyser les données si vous n'en parlez à personne ! Parlons maintenant de ce que vous devez faire lorsque vous présentez votre analyse. Restons-en à l'exemple de nos combinaisons. Si je voulais écrire ce résultat pour un article ou quelque chose comme ça, alors la façon conventionnelle de le rapporter serait d'écrire quelque chose comme ceci :

Sur les 200 participants à l'expérience, 64 ont sélectionné des cœurs pour leur premier choix, 51 ont sélectionné des carreaux, 50 des piques et 35 des trèfles. Un test d'ajustement du chi carré a été effectué pour vérifier si les probabilités de choix étaient identiques pour

les quatre couleurs. Les résultats ont été significatifs ($\chi^2(3) = 8.44, p < .05$), ce qui suggère que les gens n'ont pas choisi les couleurs au hasard.

C'est assez simple et, espérons-le, cela semble assez banal. Cela dit, il y a certaines choses que vous devriez noter au sujet de cette description :

- *Le test statistique est précédé des statistiques descriptives*. C'est-à-dire, j'ai dit au lecteur à quoi ressemblaient les données avant de faire le test. En général, il s'agit d'une bonne pratique. Rappelez-vous toujours que votre lecteur ne connaît pas vos données aussi bien que vous. Donc, à moins que vous ne le leur décriviez correctement, les tests statistiques n'auront aucun sens pour eux et ils en seront frustrés et tristes.
- *La description vous indique quelle est l'hypothèse nulle testée*. Pour être honnête, les rédacteurs ne le font pas toujours, mais c'est souvent une bonne idée dans les situations où il existe une certaine ambiguïté ou lorsque vous ne pouvez pas compter sur le fait que vos lecteurs connaissent bien les outils statistiques que vous utilisez. Très souvent, le lecteur ne connaît pas (ou ne se souvient pas) de tous les détails du test que vous utilisez, c'est donc une sorte de politesse de leur « rappeler » ! En ce qui concerne le test d'adéquation, vous pouvez généralement compter sur un public scientifique qui sait comment il fonctionne (puisqu'il est couvert dans la plupart des cours d'introduction aux statistiques). Cependant, c'est quand même une bonne idée d'être explicite sur l'énoncé de l'hypothèse nulle (brièvement !) parce que l'hypothèse nulle peut être différente en fonction de l'objectif vous conduisant à utiliser le test. Dans l'exemple des cartes, mon hypothèse nulle était que les quatre probabilités des couleurs étaient identiques (c.-à-d. $P_1=P_2=P_4=P_4=0,25$), mais il n'y a rien de spécial dans cette hypothèse. J'aurais tout aussi bien pu tester l'hypothèse nulle que $P_1=0,7$ et $P_2=P_3=P_4= 0,1$ en utilisant un test d'adéquation. Il est donc utile pour le lecteur que vous lui expliquiez quelle était votre hypothèse nulle. Notez aussi que j'ai décrit l'hypothèse nulle avec des mots, pas avec des maths. C'est parfaitement acceptable. Vous pouvez la décrire en termes mathématiques si vous le souhaitez, mais comme la plupart des lecteurs trouvent les mots plus faciles à lire que les symboles, la plupart des auteurs ont tendance à décrire l'hypothèse nulle en utilisant des mots s'ils le peuvent.
- *Un « bloc statistique » est inclus*. En rapportant les résultats du test lui-même, je n'ai pas seulement dit que le résultat était significatif, j'ai inclus un « bloc statistique » (c'est-à-dire la partie mathématique dense entre parenthèses) qui rapporte toutes les informations statistiques « clés ». Pour le test d'ajustement du chi carré, l'information qui est rapportée est la statistique de test (que la statistique d'ajustement était de 8,44), l'information sur la distribution utilisée dans le test (χ^2 avec 3 degrés de liberté qui est habituellement raccourci à $\chi^2_{(3)}$, et ensuite la décision à savoir si le résultat est significatif (dans ce cas $p<.05$). L'information particulière qui doit entrer dans le bloc statistique est différente pour chaque test, et chaque fois que je présente un nouveau test, je vous montre à quoi doit ressembler le bloc statistique.[72] Cependant, le principe

[72] Les conventions relatives à la façon dont les statistiques doivent être présentées ont tendance à différer quelque peu d'une discipline à l'autre. J'ai tendance à m'en tenir à la

général est que vous devriez toujours fournir suffisamment d'informations pour que le lecteur puisse vérifier lui-même les résultats du test s'il le souhaite vraiment.

- *Les résultats sont interprétés*. En plus d'indiquer que le résultat était significatif, j'ai fourni une interprétation du résultat (c.-à-d. que les gens n'ont pas choisi au hasard). C'est aussi une politesse envers le lecteur, parce que cela lui dit quelque chose sur ce qu'il doit croire de ce qui se passe dans vos données. Si vous n'incluez pas quelque chose comme ça, c'est vraiment difficile pour votre lecteur de comprendre ce qui se passe.[73]

Comme pour tout le reste, votre souci premier devrait être d'*expliquer* les choses à votre lecteur. Rappelez-vous toujours que le but de rapporter vos résultats est de communiquer à un autre être humain. Je ne peux pas vous dire combien de fois j'ai vu la section des résultats d'un rapport, d'une thèse ou même d'un article scientifique qui n'est que du charabia, parce que l'auteur s'est concentré uniquement sur le fait d'avoir inclus tous les chiffres et oublié de communiquer avec le lecteur humain.

Un commentaire sur la notation statistique

Satan se plaît aussi bien dans les statistiques que dans la citation des Écritures. - H.G. Wells

Si vous avez lu très attentivement, et que vous êtes autant un pédant mathématique que moi, il y a une chose dans la façon dont j'ai écrit le test du chi carré dans la dernière section qui pourrait vous déranger un peu. vous vous dites peut-être : il y a quelque chose qui ne va pas du tout avec le fait d'écrire « $\chi^2(3) = 8.44$,. Après tout, c'est la statistique de la qualité de l'ajustement qui est égale à 8,44, alors n'aurais-je pas dû écrire $X^2 = 8.44$ ou peut-être GOF $= 8.44$? Cela semble confondre la *distribution d'échantillonnage* (c.-à-d. χ^2 avec *df* = 3) avec la *statistique du test* (c.-à-d. X^2). Il y a fort à parier que vous pensiez qu'il s'agissait d'une faute de frappe, puisque χ et X se ressemblent beaucoup. Curieusement, ça ne l'est pas. Ecrire $\chi^2(3) = 8.44$ est essentiellement une façon très condensée d'écrire «la

---

façon dont les choses se font en psychologie, puisque c'est ce que je fais. Mais le principe général de fournir suffisamment d'informations au lecteur pour lui permettre de vérifier vos résultats est assez universel, je pense.

[73] Pour certains, ce conseil peut sembler étrange, ou du moins en contradiction avec les conseils « habituels » sur la façon de rédiger un rapport technique. En règle générale, on dit aux élèves que la section « résultats » d'un rapport sert à décrire les données et à présenter une analyse statistique, tandis que la section « discussion » sert à fournir une interprétation. C'est vrai jusqu'à un certain point, mais je pense que les gens l'interprètent souvent beaucoup trop littéralement. La façon dont je l'aborde habituellement est de fournir une interprétation rapide et simple des données dans la section des résultats, afin que mon lecteur comprenne ce que les données nous disent. Puis, au cours de la discussion, j'essaie de tenir un discours plus générale sur la façon dont mes résultats s'intègrent au reste de la littérature scientifique. Bref, ne laissez pas le conseil « l'interprétation va dans la discussion » transformer votre section de résultats en un fouillis incompréhensible. Il est beaucoup plus important d'être compris par votre lecteur.

distribution d'échantillonnage de la statistique du test est $\chi^2(3)$, et la valeur de la statistique du test est 8,44».

Dans un sens, c'est un peu stupide. Il existe de *nombreuses* statistiques de test différentes qui s'avèrent avoir une distribution d'échantillonnage du chi carré. La statistique χ^2 que nous avons utilisée pour notre test d'ajustement n'est qu'une statistique parmi tant d'autres (bien qu'elle soit l'une des plus fréquemment rencontrées). Dans un monde sensible et parfaitement organisé, nous aurions *toujours* un nom distinct pour la statistique du test et la distribution d'échantillonnage. De cette façon, le bloc de statistiques lui-même vous dirait exactement ce que le chercheur a calculé. Parfois, ça arrive. Par exemple, la statistique de test utilisée dans le test d'ajustement de Pearson s'écrit χ^2, mais il existe un test étroitement lié connu sous le nom de *test G*[74] (Sokal et Rohlf 1994), dans lequel la statistique de test s'écrit *G*. En l'occurrence, le test d'ajustement de Pearson et le *test G* testent la même hypothèse nulle et la distribution de l'échantillonnage est exactement la même (c'est-à-dire chi carré avec *k -1 degrés de liberté*). Si j'avais fait un *test G* pour les données des cartes plutôt qu'un test d'ajustement, alors j'aurais obtenu une statistique de test de $G = 8.65$, qui est légèrement différente de la valeur $\chi^2 = 8{,}44$ que j'ai obtenue plus tôt et qui produit une *valeur p* légèrement inférieure de *p=.034*. Supposons que la convention consiste à déclarer la statistique du test, puis la distribution d'échantillonnage, et enfin la *valeur p*. Si c'était vrai, alors ces deux situations produiraient des blocs statistiques différents : mon résultat original serait écrit $X^2 = 8.44, \chi^2(3), p = .038$, alors que la nouvelle version utilisant le *test G serait* écrite $G = 8.65, \chi^2(3), p = .034$. Cependant, en utilisant la norme d'écriture condensée, le résultat original est écrit $\chi^2(3) = 8.44, p = .038$, et le nouveau est écrit $\chi^2(3) = 8.65, p = .034$, et savoir quel test j'ai fait n'est pas clair.

Alors pourquoi ne vivons-nous pas dans un monde dans lequel le contenu du bloc de statistiques spécifie de façon unique quels tests ont été effectués ? La raison profonde est que la vie est désordonnée. Nous (en tant qu'utilisateurs d'outils statistiques) voulons qu'il soit agréable, propre et organisé. Nous voulons qu'il soit *conçu* comme s'il s'agissait d'un produit, mais ce n'est pas ainsi que la vie fonctionne. La statistique est une discipline intellectuelle comme une autre, et en tant que telle, c'est un projet massivement distribué, partiellement collaboratif et partiellement compétitif que personne ne comprend vraiment complètement. Les choses que vous et moi utilisons comme outils d'analyse de données n'ont pas été créées par une loi des dieux de la statistique. Ils ont été inventés par beaucoup de personnes différentes, publiés sous forme d'articles dans des revues universitaires, mis en œuvre, corrigés et modifiés par beaucoup d'autres personnes, puis expliqués aux étudiants dans des manuels scolaires par quelqu'un d'autre. Par conséquent, il y a *beaucoup de* statistiques de test qui n'ont même pas de noms, et par conséquent, on leur donne juste le même nom que la distribution d'échantillonnage correspondante. Comme nous le verrons plus loin, toute statistique de test qui suit une distribution χ^2 est communément appelée « statistique du chi carré », tout ce qui suit une *distribution t* est appelé « *statistique t* », et

---

[74] Pour compliquer les choses, le test d'ajustement est un cas particulier de toute une classe de tests connus sous le nom de tests du rapport de vraisemblance (Likehood ratio test ; LRT). Je ne traite pas des LRT dans ce livre, mais ce sont des choses très pratiques à savoir.

ainsi de suite. Mais, comme l'illustre l'exemple de χ^2 par rapport à G, deux choses différentes avec la même distribution d'échantillonnage sont toujours … différentes.

Par conséquent, c'est parfois une bonne idée d'être clair sur le test que vous avez effectué, surtout si vous faites quelque chose d'inhabituel. Si vous dites simplement « test du chi carré », on ne sait pas vraiment de quel test vous parlez. Bien que, puisque les deux tests du chi carré les plus courants sont le test d'ajustement et le test d'indépendance (section 10.2), la plupart des lecteurs ayant une formation en statistiques peuvent probablement deviner. Néanmoins, c'est quelque chose dont il faut être conscient.

Le test d'indépendance (ou d'association) χ^2

GUARDBOT 1 : Halte ! GUARDBOT 2 : Etes-vous robot ou humain ? LEELA : Nous sommes robot. FRY : Euh, ouais ! Juste deux robots en train de le voler ! Hein ? GUARDBOT 1: Administrer le test. GUARDBOT 2: Lequel des énoncés suivants préféreriez-vous le plus ?

A : Un chiot, B : Une jolie fleur pour votre chéri, ou C : Un gros fichier de données correctement formaté ?

GUARDBOT 1 : Choisissez ! /- Futurama, «Fear of a Bot Planet»

L'autre jour, je regardais un documentaire animé examinant les coutumes pittoresques des indigènes de la planète *Chapek 9*. Apparemment, pour avoir accès à leur capitale, un visiteur doit prouver qu'il est un robot, pas un humain. Afin de déterminer si un visiteur est humain ou non, les autochtones lui demandent s'il préfère des chiots, des fleurs ou de gros fichiers de données correctement formatés. « Plutôt malin, me suis-je dit, et si les humains et les robots avaient les mêmes préférences ? Ce ne serait alors probablement pas un très bon test » Il se trouve que j'ai mis la main sur les données de test que les autorités civiles de *Chapek 9* utilisaient pour vérifier cela. Il s'avère que ce qu'ils ont fait était très simple. Ils ont trouvé un tas de robots et un tas d'humains et leur ont demandé ce qu'ils préféraient. J'ai sauvegardé leurs données dans un fichier appelé chapek9.omv, que nous pouvons maintenant charger dans Jamovi. En plus de la variable ID qui identifie les individus, il y a deux variables nominales, l'espèce et le choix. Au total, il y a 180 entrées dans l'ensemble de données, une pour chaque personne (en comptant à la fois les robots et les humains comme « personnes ») à qui on a demandé de faire un choix. Plus précisément, il y a 93 humains et 87 robots, et le fichier de données est de loin le meilleur choix. Vous pouvez le vérifier vous-même en demandant à Jamovi des tableaux de fréquences, sous le bouton « Exploration » - « Descriptives ». Cependant, ce résumé ne répond pas à la question qui nous intéresse. Pour ce faire, nous avons besoin d'une description plus détaillée des données. Ce que nous voulons faire, c'est examiner les choix ventilés *par* espèce. En d'autres termes, nous devons croiser les données (voir la section 6.1). Dans Jamovi, nous le faisons en utilisant l'analyse « Fréquencies » - « Contingency Tables » - « Independants Samples », et nous devrions obtenir un tableau comme celui-ci :

	Robot	Humain	Total
Chiot	13	15	28
Fleur	30	13	43

Données	44	65	109
Total	87	93	180

Il en ressort clairement que la grande majorité des humains ont choisi le fichier de données, alors que les robots avaient tendance à être beaucoup plus équilibrés dans leurs préférences. En laissant de côté la question de savoir *pourquoi* les humains seraient plus susceptibles de choisir le fichier de données pour le moment (ce qui semble assez étrange, il est vrai), notre première tâche consiste à déterminer si l'écart entre les choix humains et les choix robotiques dans l'ensemble de données est statistiquement significatif.

Construire notre test d'hypothèse

Comment analyser ces données ? Plus précisément, puisque mon hypothèse de *recherche* est que « les humains et les robots répondent à la question de façon différente », comment puis-je construire un test de l'hypothèse *nulle* selon laquelle « les humains et les robots répondent à la question de la même façon » ? Comme précédemment, nous commençons par établir une notation pour décrire les données :

	Robot	Humain	Total
Chiot	O_{11}	O_{12}	R_1
Fleur	O_{21}	O_{22}	R_2
Données	O_{31}	O_{32}	R_3
Total	C_1	C_2	N

Dans cette notation, nous disons que O_i est un comptage (fréquence observée) du nombre de répondants de l'espèce j (robots ou humains) qui ont donné la réponse i (chiot, fleur ou données) quand on leur a demandé de faire un choix. Le nombre total d'observations est écrit N, comme d'habitude. Enfin, j'ai utilisé R_i pour indiquer les totaux des lignes (par exemple, R_1 est le nombre total de personnes qui ont choisi la fleur), et C_j pour indiquer les totaux des colonnes (par exemple, C_1 est le nombre total de robots).[75]

---

[75] Note technique. La façon dont j'ai décrit le test prétend que les totaux des colonnes sont fixes (c.-à-d. que le chercheur avait l'intention de sonder 87 robots et 93 humains) et que les totaux des rangées sont aléatoires (c.-à-d. qu'il s'avère que 28 personnes ont choisi le chiot). Pour reprendre la terminologie de mon manuel de statistiques mathématiques (Hogg and Craig 2005), je devrais techniquement qualifier cette situation de test du Khi-deux d'homogénéité et réserver le terme test du Khi-carré d'indépendance à la situation où les totaux des lignes et des colonnes sont des résultats aléatoires de l'expérience. Dans les premières ébauches de ce livre, c'est exactement ce que j'ai fait. Cependant, il s'avère que ces deux tests sont identiques, et je les ai donc regroupés.

Réfléchissons maintenant à ce que dit l'hypothèse nulle. Si les robots et les humains répondent de la même manière à la question, cela signifie que la probabilité « qu'un robot choisisse chiot » est la même que la probabilité « qu'un humain choisisse chiot », et ainsi de suite pour les deux autres possibilités. Donc, si nous utilisons P_{ij} pour indiquer « la probabilité qu'un membre de l'espèce j donne la réponse i» alors notre hypothèse nulle est que : H_0: Toutes les affirmations suivantes sont vraies : $P_{11}=P_{12}$(même probabilité de dire «chiot») $P_{21}=P_{22}$ (même probabilité de dire «fleur») et $P_{31}=P_{32}$(même probabilité de dire «données»)

En fait, puisque l'hypothèse nulle prétend que les probabilités de choix réelles ne dépendent pas de l'espèce de la personne qui fait le choix, nous pouvons laisser P_i se référer à cette probabilité, par exemple, P_1 est la probabilité réelle de choisir le chiot.

Ensuite, de la même façon que nous l'avons fait avec le test de la qualité d'ajustement, nous devons calculer les fréquences attendues. En d'autres termes, pour chacun des effectifs O_{ij} observés, nous devons déterminer ce à quoi nous attendre avec l'hypothèse nulle. Notons cette fréquence attendue par E_{ij}. Cette fois, c'est un peu plus délicat. S'il y a un total de personnes C_j qui appartiennent à l'espèce j, et que la vraie probabilité pour quiconque (indépendamment de l'espèce) de choisir l'option i est P_i, alors la fréquence prévue est juste :

$$E_{ij} = C_j \times P_i$$

Tout cela est très bien, mais nous avons un problème. Contrairement à la situation que nous avons eue avec le test d'adéquation, l'hypothèse nulle ne spécifie pas réellement une valeur particulière pour P_i. C'est quelque chose que nous devons estimer (chapitre 8) à partir des données ! Heureusement, c'est assez facile à faire. Si 28 personnes sur 180 ont sélectionné les fleurs, alors une estimation naturelle de la probabilité de choisir des fleurs est 28/180, ce qui représente environ .16. Si nous formulons cela en termes mathématiques, ce que nous disons, c'est que notre estimation de la probabilité de choisir l'option i n'est que le total des lignes divisé par la taille totale de l'échantillon :

$$\hat{P}_i = \frac{R_i}{N}$$

Par conséquent, notre fréquence prévue peut s'écrire comme le produit (c.-à-d. la multiplication) du total des lignes et du total des colonnes, divisé par le nombre total d'observations :[76]

$$E_{ij} = \frac{R_i \times C_j}{N}$$

---

[76] Techniquement, E_{ij} ici est une estimation, donc je devrais probablement l'écrire $\hat{E}_{ij}$. Mais comme personne d'autre ne le fait, moi non plus.

Maintenant que nous avons trouvé comment calculer les fréquences attendues, il est simple de définir une statistique de test, suivant exactement la même stratégie que celle que nous avons utilisée dans le test d'ajustement. En fait, c'est à peu près la *même* statistique.

Pour un tableau de contingence avec *r* lignes et *c* colonnes, l'équation qui définit notre statistique χ^2 est la suivante

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(E_{ij} - O_{ij}\right)^2}{E_{ij}}$$

La seule différence est que je dois inclure deux signes de sommation (i.e., $\sum$) pour indiquer que nous faisons la somme sur les lignes et colonnes.

Comme précédemment, les grandes valeurs de χ^2 indiquent que l'hypothèse nulle fournit une mauvaise description des données, alors que les petites valeurs de χ^2 suggèrent qu'elle rend bien compte des données. Par conséquent, comme la dernière fois, nous cherchons à rejeter l'hypothèse nulle si χ^2 est trop grand.

Comme on pouvait s'y attendre, cette statistique est distribuée comme χ^2 Tout ce que nous avons à faire est de déterminer combien de degrés de liberté sont impliqués, ce qui n'est pas trop difficile. Comme je l'ai déjà mentionné, vous pouvez (habituellement) penser que les degrés de liberté sont égaux au nombre de données que vous analysez, moins le nombre de contraintes. Un tableau de contingence avec *r* lignes et *c* colonnes contient un total de $r \times c$ fréquences observées, donc c'est le nombre total d'observations. Qu'en est-il des contraintes ? Ici, c'est un peu plus délicat. La réponse est toujours la même mais l'explication de la raison pour laquelle les degrés de liberté prennent cette valeur est différente selon le plan expérimental.

$$df = (C - 1) \times (R - 1)$$

Supposons que nous avions honnêtement l'intention de sonder exactement 87 robots et 93 humains (totaux de colonnes fixés par l'expérimentateur), mais que nous avons laissé les totaux de lignes libres de varier (les totaux de lignes sont des variables aléatoires). Pensons aux contraintes qui s'appliquent ici. Eh bien, puisque nous avons délibérément fixé les totaux des colonnes par action de l'expérimentateur, nous avons *c* contraintes dans ce cas-là. Mais, en fait, il y a plus que ça. Rappelez-vous comment notre hypothèse nulle avait des paramètres libres (c.-à-d. que nous devions estimer les valeurs P_i) ? Celles-ci comptent aussi. Je n'expliquerai pas pourquoi dans ce livre, mais chaque paramètre libre dans l'hypothèse nulle est un peu comme une contrainte supplémentaire. Alors, combien sont-ils ? Eh bien, puisque ces probabilités doivent s'additionner à 1, il n'y en a que *r-1*. Donc notre degré total de liberté est :

$$\begin{aligned} df &= (\text{nombre d'observations} - \text{nombre de contraintes}) \\ &= (rc) - (c + r - 1)) \\ &= rc - c - r + 1 \\ &= (r - 1)(c - 1) \end{aligned}$$

Autrement dit, nous avons interrogé les 180 premières personnes que nous avons vues et il s'est avéré que 87 étaient des robots et 93 étaient des humains. Cette fois-ci, notre raisonnement serait légèrement différent, mais nous conduirait quand même à la même réponse. Notre hypothèse nulle a toujours r-1 paramètres libres correspondant aux probabilités de choix, mais elle a maintenant *c - 1* paramètres libres correspondant aux probabilités d'espèces, car il faudrait aussi estimer la probabilité qu'une personne échantillonnée au hasard se révèle être un robot.[77] Enfin, puisque nous avons effectivement fixé le nombre total d'observations *N*, c'est une contrainte de plus. Nous avons donc maintenant des observations *rc* et des contraintes *(c-1)+(r-1)+1*. Qu'est-ce que ça donne ?

$$\begin{aligned} df &= (\text{nombre d'observations} - \text{nombre de contraintes}) \\ &= rc - ((c-1) + (r-1) + 1 \\ &= rc - c - r + 1 \\ &= (r-1)(c-1) \end{aligned}$$

Incroyable !

Faire le test avec Jamovi

Maintenant qu'on sait comment fonctionne le test, voyons comment ça se passe en Jamovi. Aussi tentant que cela puisse paraître de vous guider à travers les calculs fastidieux pour que vous soyez forcé de l'apprendre sur le long terme, je me dis que cela n'a pas de sens. Je vous ai déjà montré comment le faire sur le long chemin pour le test d'adéquation dans la dernière section, et comme le test d'indépendance n'est pas différent sur le plan conceptuel, vous n'apprendrez rien de nouveau en le faisant à la main. Alors à la place, je vais vous montrer le chemin le plus facile. Après avoir effectué le test dans Jamovi (« Frequencies » - « Contingency Tables » -« Independent Samples »), tout ce que vous avez à faire est de regarder sous le tableau de contingence dans la fenêtre de résultats Jamovi, il y a la statistique χ^2 pour vous. Ceci montre une valeur statistique de χ^2 de 10.72, avec 2 df. et *p-value = 0,005*.

C'était facile, n'est-ce pas ? Vous pouvez également demander à Jamovi de vous montrer les effectifs attendus - il vous suffit de cliquer sur la case à cocher « Counts » - « Expected » dans les options « cells » et les effectifs attendus apparaîtront dans le tableau de contingence. Pendant que vous faites cela, une mesure de la taille de l'effet serait utile. Nous choisirons Cramer's V, et vous pouvez le spécifier à partir d'une case à cocher dans les options « Statistiques », et cela donne une valeur pour Cramer's V de 0,24. Nous en reparlerons dans un instant.

Ce résultat nous donne suffisamment d'informations pour rédiger le résultat :

Le χ^2 de Pearson révèle une association significative entre les espèces et le choix $\chi^2(2) = 10{,}7, p < .01$). Les robots semblaient plus portés à dire qu'ils préféraient les fleurs, alors que les humains étaient plus portés à dire qu'ils préféraient les données.

---

[77] Un problème dont beaucoup d'entre nous s'inquiètent dans la vraie vie.

Remarquez qu'une fois de plus, j'ai fourni un peu d'interprétation pour aider le lecteur humain à comprendre ce qui se passe avec les données. Plus tard dans ma section de discussion, je préciserai un peu plus le contexte. Pour illustrer la différence, voici ce que je dirai probablement plus tard :

Le fait que les humains semblaient avoir une préférence plus marquée pour les fichiers de données brutes que les robots est quelque peu contre-intuitif. Cependant, dans le contexte, cela a un certain sens, car l'autorité civile sur Chapek 9 a une tendance malheureuse à tuer et à disséquer les humains lorsqu'ils sont identifiés. Il semble donc très probable que les participants humains n'aient pas répondu honnêtement à la question, afin d'éviter des conséquences potentiellement indésirables. Il s'agit là d'une faiblesse méthodologique importante.

Je suppose que cela pourrait être considéré comme un exemple assez extrême d'effet de réactivité. De toute évidence, dans ce cas, le problème est suffisamment grave pour que l'étude soit plus ou moins inutile en tant qu'outil permettant de comprendre la différence entre les préférences des humains et des robots. Cependant, j'espère que cela illustre la différence entre obtenir un résultat statistiquement significatif (notre hypothèse nulle est rejetée en faveur de l'alternative), et trouver quelque chose de valeur scientifique (les données ne nous disent rien d'intéressant sur notre hypothèse de recherche en raison d'un grand défaut méthodologique).

Postscript

Plus tard, j'ai découvert que les données avaient été inventées et que j'avais regardé des dessins animés au lieu de travailler.

La correction de continuité

Bien, c'est l'heure d'une petite digression. Je vous ai un peu menti jusqu'ici. Il y a un petit changement que vous devez apporter à vos calculs lorsque vous n'avez qu'un seul degré de liberté. C'est ce qu'on appelle la « correction de continuité », ou parfois la **correction de Yates**. Souvenez-vous de ce j'ai souligné plus tôt : le test χ^2 est basé sur une approximation, en particulier sur l'hypothèse que la distribution binomiale commence à ressembler à une distribution normale lorsque N est grand. Le problème avec cela est qu'il ne fonctionne souvent pas bien, surtout quand vous avez seulement 1 degré de liberté (par exemple, lorsque vous faites un test d'indépendance sur une table de contingence 2 x 2). La raison principale est que la vraie distribution d'échantillonnage pour la statistique X^2 est en fait discrète (parce qu'il s'agit de données catégorielles !) mais la distribution χ^2 est continue. Cela peut introduire des problèmes systématiques. Plus précisément, lorsque N est petit et que $df = 1$, la statistique de la qualité de l'ajustement tend à être « trop grande », ce qui signifie que vous avez en fait une valeur α plus grande que vous ne le pensez (ou, de manière équivalente, les valeurs p sont un peu trop petites).

Yates (1934) a suggéré une solution simple, dans laquelle vous redéfinissez la statistique de la qualité de l'ajustement comme :

$$\chi^2 = \sum_i \frac{(|\ E_i - O_i\ | - 0{,}5)^2}{E_i}$$

En gros, il soustrait juste 0,5 partout.

D'après ce que j'en sais d'après l'article de Yates, la correction est essentiellement un bricolage. Elle ne découle d'aucune théorie fondée sur des principes. Elle repose plutôt sur un examen du comportement du test et sur l'observation que la version corrigée semble mieux fonctionner. Vous pouvez spécifier cette correction dans Jamovi à partir d'une case à cocher dans les options « Statistics », où elle est appelée « χ^2 continuity correction ».

Taille de l'effet

Comme nous l'avons mentionné précédemment (section 9.8), il est de plus en plus courant de demander aux chercheurs de déclarer une certaine mesure de la taille de l'effet. Supposons donc que vous ayez effectué votre test du chi carré, qui s'avère important. Vous savez donc maintenant qu'il y a une certaine association entre vos variables (test d'indépendance) ou un certain écart par rapport aux probabilités spécifiées (test d'ajustement). Vous voulez maintenant déclarer une mesure de l'ampleur de l'effet. Autrement dit, étant donné qu'il y a une association ou une déviation, quelle est sa force ?

Il existe différentes mesures que vous pouvez choisir de déclarer et plusieurs outils différents que vous pouvez utiliser pour les calculer. Je ne les aborderai pas tous, mais je me concentrerai plutôt sur les mesures de la taille de l'effet les plus couramment utilisés.

Par défaut, les deux mesures que les gens ont tendance à déclarer le plus souvent sont la statistique ϕ et la version supérieure, connue sous le nom de V.

Mathématiquement, ils sont très simples. Pour calculer la statistique ϕ, il suffit de diviser votre valeur χ^2 par la taille de l'échantillon et de prendre la racine carrée :

$$\phi = \sqrt{\frac{X^2}{N}}$$

L'idée ici est que la statistique de ϕ est supposée se situer entre 0 (aucune association du tout) et 1 (association parfaite), mais elle ne le fait pas toujours quand votre tableau de contingence est plus grand que 2x2, ce qui est une vraie difficulté. Pour des tables plus grandes, il est en fait possible d'obtenir $\phi < 1$, ce qui est plutôt insatisfaisant. Ainsi, pour corriger cela, les gens préfèrent généralement rapporter la statistique V proposée par Cramer (1999). C'est un ajustement assez simple à ϕ. Si vous avez un tableau de contingence avec r lignes et c colonnes, alors définissez $k = min(r, c)$ comme étant la plus petite des deux valeurs. Si c'est le cas, la statistique **V de Cramer** est la suivante

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

Et c'est fini. Cela semble être une mesure assez populaire, probablement parce qu'elle est facile à calculer et qu'elle donne des réponses qui ne sont pas complètement idiotes. Avec le *V* de Crame, vous savez que la valeur varie vraiment de 0 (aucune association du tout) à 1 (association parfaite).

Hypothèses relatives au(x) test(s)

Tous les tests statistiques font des hypothèses, et c'est généralement une bonne idée de vérifier que ces hypothèses sont respectées. Pour les tests du chi carré dont il a été question jusqu'à présent dans ce chapitre, les hypothèses sont les suivantes :

- *Les fréquences attendues sont suffisamment élevées*. Rappelez-vous comment dans la section précédente nous avons vu que la distribution d'échantillonnage de χ^2 émerge parce que la distribution binomiale est assez similaire à une distribution normale ? Eh bien, comme nous l'avons vu au chapitre 7, cela n'est vrai que lorsque le nombre d'observations est suffisamment important. En pratique, cela signifie que toutes les fréquences attendues doivent être raisonnablement grandes. Quelle est la taille de raisonnablement grand ? Les opinions divergent, mais l'hypothèse par défaut semble être que vous aimeriez généralement voir toutes vos fréquences attendues supérieures à environ 5, bien que pour des tables plus grandes, vous seriez probablement d'accord si au moins 80% des fréquences attendues sont supérieures à 5 et aucune d'entre elles n'est inférieure à 1. Cependant, de ce que j'ai pu découvrir (par exemple,Cochran (1952)) cela semble avoir été proposé comme des lignes directrices générales, non des règles strictes, et elles semblent être assez prudentes (Larntz 1978).

- *Les données sont indépendantes les unes des autres*. Une hypothèse quelque peu cachée du test du chi carré est qu'il faut vraiment croire que les observations sont indépendantes. Précisons ce que je veux dire. Supposons que je m'intéresse à la proportion de bébés nés dans un hôpital donné qui sont des garçons. Je me promène dans les maternités et j'observe 20 filles et seulement 10 garçons. C'est une différence assez convaincante, non ? Mais plus tard, il s'est avéré que j'étais entré 10 fois dans la même salle et que je n'avais vu que 2 filles et 1 garçon. Ce n'est pas aussi convaincant, n'est-ce pas ? Mes 30 observations initiales étaient massivement non indépendantes, et n'équivalaient en fait qu'à 3 observations indépendantes. C'est évidemment un exemple extrême (et extrêmement stupide), mais il illustre la question fondamentale. La non-indépendance « empoisonne les données ». Parfois, cela vous pousse à rejeter faussement l'hypothèse nulle, comme l'illustre l'exemple stupide de l'hôpital, mais cela peut aussi aller dans l'autre sens. Pour donner un exemple un peu moins stupide, considérons ce qui se passerait si j'avais fait l'expérience des cartes un peu différemment au lieu de demander à 200 personnes d'essayer d'imaginer d'échantillonner une carte au hasard, supposons que je demande à 50 personnes de choisir 4 cartes. Une possibilité serait que chacun choisisse un coeur, un trèfle, un carreau et un pique (en accord avec l'heuristique de la représentativité ;Tversky & Kahneman (1973). C'est un comportement particulièrement non aléatoire de la part des gens, mais dans ce cas, j'obtiendrais une fréquence observée de 50 pour les quatre couleurs. Pour cet exemple, le fait que les observations ne soient pas indépendantes

(parce que les quatre cartes que vous choisirez seront liées les unes aux autres) conduit en fait à l'effet inverse, conserver faussement l'hypothèse nulle.

Si vous vous trouvez dans une situation où l'indépendance est violée, il est possible d'utiliser le test de McNemar (dont nous parlerons) ou le test de Cochran (dont nous ne parlerons pas). De même, si le nombre de cellules attendu est trop faible, vérifiez le test exact de Fisher. C'est à ces sujets que nous nous intéressons maintenant.

Le test exact de Fisher

Que faire si votre nombre de cellules est trop faible, mais que vous souhaitez quand même tester l'hypothèse nulle que les deux variables sont indépendantes ? L'une des réponses serait de « recueillir plus de données », mais c'est beaucoup trop désinvolte. Il y a beaucoup de situations dans lesquelles il serait impossible ou contraire à l'éthique de le faire. Si c'est le cas, les statisticiens ont une sorte d'obligation morale de fournir aux scientifiques de meilleurs tests. En l'occurrence, Fisher (1922) a aimablement fourni la bonne réponse à la question. Pour illustrer l'idée de base, supposons que nous analysons les données d'une expérience sur le terrain portant sur l'état émotionnel de personnes qui ont été accusées de sorcellerie, dont certaines sont actuellement brûlées sur le bûcher.[Cet exemple est basé sur une plaisanterie publiée dans un article du Journal of Irreproducible Results.] Malheureusement pour le scientifique (mais heureusement pour la population en général), il est en fait très difficile de trouver des personnes en train d'être brûlées, de sorte que le nombre de d'observations est très faible dans certains cas. Un tableau de contingence des données salem.csv illustre ce point :

	Brûlée	
Heureuse	FAUX	VRAI
FAUX	3	3
VRAI	10	0

En regardant ces données, il serait difficile de ne pas soupçonner que les gens qui ne sont pas brulés sont plus susceptibles d'être heureux que ceux qui sont brûlés. Cependant, le test du khi-carré le rend très difficile à tester en raison de la petite taille de l'échantillon. Donc, en tant que personne qui ne veut pas être brûlée, j'aimerais vraiment obtenir une meilleure réponse que celle-ci. C'est là que le **test exact de Fisher** (Fisher 1922) est très utile.

Le test exact de Fisher fonctionne un peu différemment du test du chi carré (ou en fait de tout autre test d'hypothèse dont je parle dans ce livre) dans la mesure où il n'a pas de statistique de test, mais il calcule « directement » la *valeur p*. J'expliquerai les bases du fonctionnement du test pour un tableau de contingence *2X2*. Comme avant, voyons la notation:

	Heureuse	Triste	Total

Brûlée	O_{11}	O_{12}	R_1
Non brûlée	O_{21}	O_{22}	R_2
Total	C_1	C_1	N

Pour construire le test, Fisher traite les totaux des lignes et des colonnes (R_1, R_2, C_1 et C_2) comme des quantités fixes connues, puis calcule la probabilité que nous aurions obtenu les fréquences observées (O_{11}, O_{12}, O_{21} et O_{22}) avec ces totaux. Dans la notation que nous avons développée au chapitre 7, ceci est écrit :

$$P(O_{11}, O_{12}, O_{21}, O_{22} \mid R_1, R_2, C_1, C_2)$$

et comme vous pouvez l'imaginer, c'est un exercice un peu délicat de savoir quelle est cette probabilité. Mais il s'avère que cette probabilité est décrite par une distribution connue sous le nom de **distribution hypergéométrique**. Ce que nous devons faire pour calculer notre *valeur p* est de calculer la probabilité d'observer ce tableau particulier *ou un tableau qui est « plus extrême »*[78]. Dans les années 1920, le calcul de cette somme était décourageant même dans les situations les plus simples, mais de nos jours, c'est assez facile tant que les tables ne sont pas trop grandes et que l'échantillon n'est pas trop grand. La question conceptuellement délicate est de savoir ce que cela signifie de dire qu'une table de contingence est plus « extrême » qu'une autre. La solution la plus simple est de dire que la table avec la probabilité la plus faible est la plus extrême. Ceci nous donne alors la *valeur p*.

Vous pouvez spécifier ce test dans Jamovi à partir d'une case à cocher dans les options « Statistics » de l'analyse « Contingency Tables ». Lorsque vous le faites avec les données du fichier salem.csv, la statistique exacte du test de Fisher est affichée dans les résultats. La principale chose qui nous intéresse ici est la valeur de p (*p-value*), qui dans ce cas est suffisamment petite ($p<.036$) pour justifier le rejet de l'hypothèse nulle que les personnes brûlées sont aussi heureuses que celles qui ne le sont pas. Voir la Figure 11-6.

[78] Il n'est pas surprenant que le test exact de Fisher soit motivé par l'interprétation de Fisher d'une valeur p, et non par celle de Neyman ! Voir la section 9.5.

Contingency Tables

	on.fire		
happy	FALSE	TRUE	Total
FALSE	3	3	6
TRUE	10	0	10
Total	13	3	16

χ^2 Tests

	Value	df	p
χ^2 continuity correction	3.31	1	0.06888
Fisher's exact test	0.00		0.03571
N	16		

Figure 10-6 :Sortie du test exact de Fisher dans Jamovi. Ignorez la « Valeur » et référez-vous simplement à la valeur p.

Le test McNemar

Supposons que vous ayez été embauché pour travailler pour le *Parti politique générique australien* (AGPP) et qu'une partie de votre travail consiste à déterminer l'efficacité des publicités politiques de l'AGPP. Vous décidez donc de constituer un échantillon de N = 100 personnes et de leur demander de regarder les publicités de l'AGPP. Avant qu'ils répondent, vous leur demandez s'il a l'intention de voter pour l'AGPP, puis après avoir montré les annonces, vous leur demandez à nouveau s'il a changé d'avis. Évidemment, si vous êtes bon dans votre travail, vous feriez aussi beaucoup d'autres choses, mais considérons juste cette simple expérience. Une façon de décrire vos données est d'utiliser le tableau de contingence suivant :

	Avant	Après	Total
Oui	30	10	40
Non	70	90	160
Total	100	100	200

Au premier abord, vous pourriez penser que cette situation se prête au test d'indépendance de Pearson χ^2 (voir la section 10.2). Cependant, un peu de réflexion révèle que nous avons un problème. Nous avons 100 participants mais 200 observations. C'est parce que chaque personne nous a fourni une réponse à la fois dans la colonne avant et dans la colonne après. Cela signifie que les 200 observations ne sont pas indépendantes les unes des autres. Si l'électeur A dit « oui » la première fois et que l'électeur B dit « non », on s'attendrait à ce que

l'électeur A dise « oui » la deuxième fois plutôt que l'électeur B ! La conséquence en est que le test habituel de χ^2 ne donnera pas de réponses fiables en raison de la violation de l'hypothèse d'indépendance. Si c'était vraiment une situation rare, je ne perdrais pas mon temps à vous en parler. Mais ce n'est pas rare du tout. Il s'agit d'un plan *standard* à mesures répétées, et aucun des tests que nous avons considérés jusqu'à présent ne peut y faire face.

La solution au problème a été publiée par McNemar (1947). L'astuce consiste à commencer par organiser vos données d'une manière légèrement différente :

	Avant : Oui	Avant : Non	Total
Après : Oui	5	5	10
Après : Non	25	65	90
Total	30	70	100

Ce sont exactement les mêmes données, mais elles ont été réécrites de sorte que chacun de nos 100 participants apparaissent dans une seule cellule. Grâce à cette réécriture, l'hypothèse d'indépendance est maintenant satisfaite, et il s'agit d'un tableau de contingence que nous pouvons utiliser pour construire une statistique d'ajustement χ^2 Cependant, comme nous le verrons, nous devons le faire d'une manière légèrement non standard. Pour voir ce qui se passe, il est utile d'étiqueter les entrées de notre tableau un peu différemment :

	Avant : Oui	Avant : Non	Total
Après : Oui	a	b	$a + b$
Après : Non	c	d	$c + d$
Total	$a + c$	$b + d$	n

Pensons ensuite à notre hypothèse nulle : elle pose que le test « avant » et le test « après » ont la même proportion de personnes qui disent « Oui, je voterai pour AGPP ». En raison de la façon dont nous avons réécrit les données, cela signifie que nous vérifions maintenant l'hypothèse que les *totaux des lignes* et des *colonnes* proviennent de la même distribution. Ainsi, l'hypothèse nulle du test de McNemar est que nous avons une « homogénéité marginale ». En d'autres termes, les totaux de ligne et les totaux de colonne ont la même répartition : Pa + Pb = Pa + Pc, de même que Pc + Pd = Pb + Pd. Notez que cela signifie que l'hypothèse nulle se simplifie en fait à P_b=P_c. En d'autres termes, en ce qui concerne le test McNemar, seules les entrées hors diagonale de ce tableau (c.-à-d. b et c) comptent ! Après l'avoir remarqué, le **test de McNemar de l'homogénéité marginale** n'est pas différent d'un test χ^2 habituel. Après avoir appliqué la correction de Yates, notre statistique de test devient :

$$\chi^2 = \frac{(|\, b - c \,| - 0{,}5)^2}{b + c}$$

ou, pour revenir à la notation que nous avons utilisée plus tôt dans ce chapitre :

$$\chi^2 = \frac{(| O_{12} - O_{21} | - 0{,}5)^2}{O_{12} + O_{21}}$$

et cette statistique a une distribution χ^2 (approximativement) avec *df = 1*. Cependant, rappelez-vous que, comme les autres tests χ^2, il ne s'agit que d'une approximation, vous devez donc avoir un nombre de d'observations raisonnablement élevé pour que cela fonctionne.

Faire le test McNemar dans Jamovi

Maintenant que vous savez ce qu'est le test McNemar, faisons-en un. Le fichier agpp.csv contient les données brutes dont j'ai parlé précédemment. L'ensemble de données de l'agpp contient trois variables, une variable id qui identifie chaque participant dans l'ensemble de données (nous verrons pourquoi c'est utile dans un instant), une variable response_before qui enregistre la réponse de la personne quand on lui a posé la question la première fois, et une variable response_after qui montre la réponse qu'elle a donnée quand on a posé la même question une deuxième fois. Notez que chaque participant n'apparaît qu'une seule fois dans cet ensemble de données. Allez dans l'analyse « Analyses » - « Frequencies » - « Contingency Tables » -Paired Samples' dans Jamovi, et déplacez response_before dans la case « Rows », et response_after dans la case « Column ». Vous obtiendrez alors un tableau de contingence dans la fenêtre des résultats, avec la statistique pour le test McNemar juste en dessous, voir Figure 10-7.

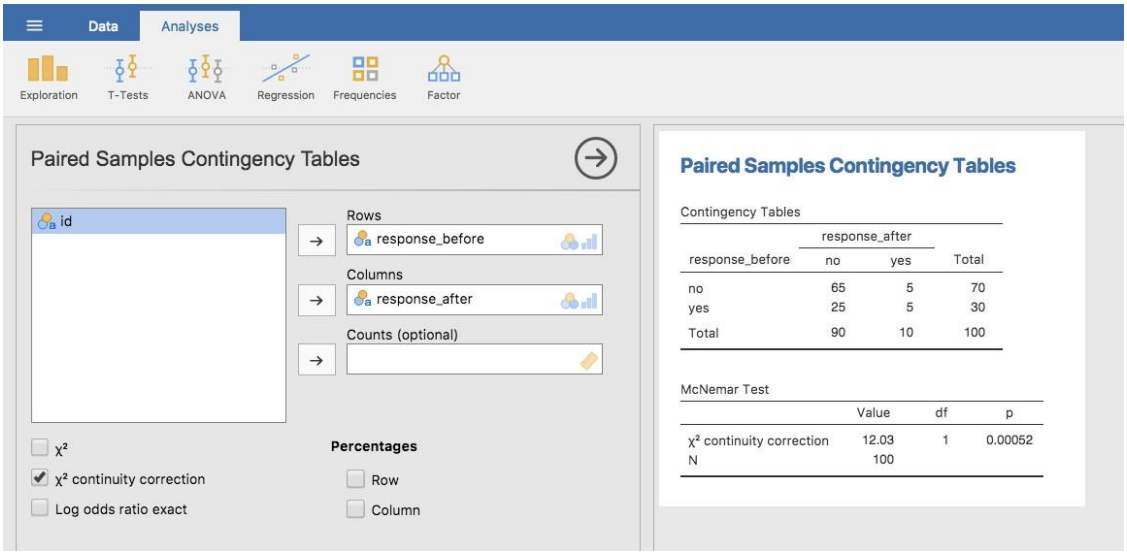


Figure 10-7 : McNemar test de sortie dans Jamovi

Et c'est fini ! Nous venons de faire un test de McNemar pour déterminer si les gens étaient tout aussi susceptibles de voter pour l'AGPP après les annonces qu'ils ne l'étaient avant. Le test était significatif ($\chi^2(1) = 12{,}03, p < .001$), suggérant qu'ils ne l'étaient pas. Et, en fait, il

semble que les publicités aient eu un effet négatif : les gens étaient moins enclins à voter pour AGPP après avoir vu les publicités. Ce qui est tout à fait logique si l'on considère la qualité d'une publicité politique typique.

Quelle est la différence entre McNemar et l'indépendance ?

Revenons au début du chapitre et regardons à nouveau l'ensemble des données des cartes. Si vous vous souvenez bien, le plan expérimental que j'ai décrit impliquait que les gens faisaient deux choix. Comme nous disposons d'informations sur le premier et le second choix que chacun a fait, nous pouvons construire le tableau de contingence suivant qui croise le premier choix et le second choix.

	choix_2			
choix_1	Trèfles	Carreaux	Cœurs	Piques
Trèfles	10	9	10	6
Carreaux	20	4	13	14
Coeur	20	18	3	23
Piques	18	13	15	4

Supposons que je voulais savoir si le choix que vous faites la deuxième fois dépend du choix que vous avez fait la première fois. C'est là qu'un test d'indépendance est utile, et ce que nous essayons de faire est de voir s'il y a une relation entre les lignes et les colonnes de ce tableau.

Contingency Tables

Contingency Tables

	choice_2				
choice_1	clubs	diamonds	hearts	spades	Total
clubs	10	9	10	6	35
diamonds	20	4	13	14	51
hearts	20	18	3	23	64
spades	18	13	15	4	50
Total	68	44	41	47	200

χ^2 Tests

	Value	df	p
χ^2	29.24	9	0.00059
N	200		

Paired Samples Contingency Tables

Contingency Tables

	choice_2				
choice_1	clubs	diamonds	hearts	spades	Total
clubs	10	9	10	6	35
diamonds	20	4	13	14	51
hearts	20	18	3	23	64
spades	18	13	15	4	50
Total	68	44	41	47	200

McNemar Test

	Value	df	p
χ^2	16.03	6	0.01358
N	200		

Figure 10-8 : Sortie d'un test d'indépendance ou apparié (McNemar) dans Jamovi

Sinon, supposons que je veuille savoir si, *en moyenne*, les fréquences des choix de cartes sont différentes la deuxième fois que la première fois. Dans cette situation, ce que j'essaie vraiment de voir, c'est si les totaux des lignes sont différents des totaux des colonnes. Pour cela, vous utilisez le test McNemar.

Les différentes statistiques produites par ces différentes analyses sont présentées à la Figure 10-8. Notez que les résultats sont différents ! Ce n'est pas le même test.

Résumé

Les idées clés discutées dans ce chapitre sont :

- Le test d'adéquation χ^2 (chi carré) (section 10.1) est utilisé lorsque vous avez un tableau des fréquences observées de différentes catégories, et l'hypothèse nulle vous donne un ensemble de probabilités « connues » pour les comparer.
- Le test d'indépendance χ^2 (chi carré) (section 10.2) est utilisé lorsque vous disposez d'un tableau de contingence (tableau croisé) de deux variables catégorielles. L'hypothèse nulle est qu'il n'y a aucune relation ou association entre les variables.
- La taille de l'effet d'un tableau de contingence peut être mesurée de plusieurs façons (section 10.4). En particulier, nous avons vu la statistique *V de Cramer*.
- Les deux versions du test de Pearson reposent sur deux hypothèses : que les fréquences attendues sont suffisamment élevées et que les observations sont indépendantes (section 10.5). Le test exact de Fisher (section 10.6) peut être utilisé lorsque les fréquences attendues sont faibles. Le test de McNemar (section 10.7) peut être utilisé pour certains types de violations de l'indépendance.

Si vous souhaitez en savoir plus sur l'analyse des données catégorielles, Agresti (1996), qui, comme son titre l'indique, fournit une *introduction à l'analyse des données catégorielles*, est un bon premier choix. Si le livre d'introduction ne vous suffit pas (ou ne peut pas résoudre le problème sur lequel vous travaillez), vous pouvez consulter Agresti (2013), *Categorical Data Analysis*. Ce dernier est un texte plus avancé, il n'est donc probablement pas sage de sauter directement de ce livre à celui-là.

Comparer deux moyennes

Au chapitre 10, nous avons examiné la situation où votre variable résultat a une échelle nominale et votre variable prédictive est également l'échelle nominale. Beaucoup de situations réelles sont ainsi, et vous constaterez que les tests du chi carré en particulier sont très largement utilisés. Cependant, vous êtes beaucoup plus susceptible de vous retrouver dans une situation où votre variable résultat est une échelle d'intervalle ou plus, et ce qui vous intéresse est de savoir si la valeur moyenne de la variable de résultat est plus élevée dans un groupe ou un autre. Par exemple, un psychologue pourrait vouloir savoir si les

niveaux d'anxiété sont plus élevés chez les parents que chez les non-parents, ou si la capacité de mémoire de travail est réduite en écoutant de la musique (par rapport à ne pas écouter de musique). Dans un contexte médical, nous pourrions vouloir savoir si un nouveau médicament augmente ou diminue la tension artérielle. Un agronome pourrait vouloir savoir si l'ajout de phosphore aux plantes indigènes australiennes les tuera.[79] Dans toutes ces situations, notre variable de résultat est une variable d'échelle d'intervalle ou de rapport continue, et notre prédicteur est une variable binaire de « regroupement ». En d'autres termes, nous voulons comparer les moyennes des deux groupes.

La réponse standard au problème de la comparaison des moyens est d'utiliser un *test t*, dont il existe plusieurs variantes en fonction de la question à laquelle vous voulez répondre. Par conséquent, la majeure partie de ce chapitre se concentre sur différents types de *tests t* : un *test t sur un échantillon* est discuté à la section 11.2, les *tests t pour* échantillons indépendants sont discutés aux sections 11.3 et 11.4, et les *tests t pour échantillons appariés* sont discutés à la section 11.5. Nous parlerons ensuite des tests unilatéraux (Section 11.6) et, après cela, nous parlerons un peu du *d de Cohen*, qui est la mesure standard de l'ampleur de l'effet pour un test t (Section 11.7). Les sections suivantes du chapitre se concentrent sur les hypothèses des *tests t* et sur les recours possibles en cas de violation de ces tests. Cependant, avant de discuter de ces choses utiles, nous allons commencer par une discussion sur les *tests z*.

Le *z-test pour* un échantillon

Dans cette section, je décrirai l'un des tests les plus inutiles de toutes les statistiques : le *test z*. Sérieusement, ce test n'est presque jamais utilisé dans la vie réelle. Son seul but réel est que, dans l'enseignement des statistiques, c'est un tremplin très pratique vers le *test t*, qui est probablement l'outil le plus (trop) utilisé dans toutes les statistiques.

Le problème d'inférence auquel le test s'adresse

Pour introduire l'idée derrière le *test z*, prenons un exemple simple. Un de mes amis, le Dr Zeppo, note son cours d'introduction à la statistique sur une courbe. Supposons que la note moyenne dans sa classe est de 67,5 et que l'écart-type est de 9,5. Parmi ses centaines d'étudiants, il s'avère que 20 d'entre eux suivent aussi des cours de psychologie. Par curiosité, je me demande si les étudiants en psychologie ont tendance à obtenir les mêmes notes que tout le monde (c.-à-d. 67,5 en moyenne) ou s'ils ont tendance à obtenir des notes plus ou moins élevées ? Il m'envoie par courriel le fichier zeppo.csv, que j'utilise pour examiner les notes de ces étudiants, dans le tableur Jamovi,

---

[79] Des expériences informelles dans mon jardin suggèrent que oui, c'est vrai. Les plantes indigènes australiennes sont adaptées à de faibles niveaux de phosphore par rapport à n'importe où ailleurs sur terre, donc si vous avez acheté une maison avec un tas de plantes exotiques et que vous voulez planter des plantes indigènes, gardez-les séparés ; les nutriments des plantes européennes sont un poison pour les Australiennes.

50 60 60 64 66 66 67 69 70 74 76 76 77 79 79 79 81 82 82 89

puis calculez la moyenne dans « Exploration » -« Descriptifs » [Pour ce faire, j'ai dû changer le niveau de mesure de X en « Continu », car lors de l'ouverture / importation du fichier csv, Jamovi en a fait une variable de niveau nominal, ce qui n'est pas correct pour mon analyse.]. La valeur moyenne est de 72,3.

Bien. Il se peut que les étudiants en psychologie obtiennent des résultats légèrement supérieurs à la normale. Cette moyenne d'échantillon de $\overline{X} = 72{,}3$ est un peu plus élevée que la moyenne de population supposée de $\mu = 67.5$ par ailleurs, une taille d'échantillon de N=20 n'est pas si grande. C'est peut-être un hasard.

Pour répondre à la question, il est utile de pouvoir écrire ce que je crois savoir. Premièrement, je sais que la moyenne de l'échantillon est $\overline{X} = 72{,}3$. Si je suis prêt à supposer que les étudiants en psychologie ont le même écart-type que le reste de la classe, je peux dire que l'écart-type de la population est $\sigma = 9{,}5$. Je supposerai aussi que puisque le Dr Zeppo évalue les étudiants dont les notes suivent une courbe qui est normalement distribuée.

Ensuite, il est utile d'être clair sur ce que je veux apprendre des données. Dans ce cas, mon hypothèse de recherche porte sur la moyenne de la *population* μ des notes des étudiants en psychologie, laquelle est inconnu. Plus précisément, je veux savoir si $\mu = 67{,}5$ ou non. Étant donné que c'est ce que je sais, pouvons-nous concevoir un test d'hypothèse pour résoudre notre problème ? Les données, ainsi que la distribution hypothétique dont on pense qu'elles proviennent, sont présentées à la Figure 11-1. Ce qui constitue la bonne réponse n'est pas du tout évident, n'est-ce pas ? Pour cela, nous allons avoir besoin de statistiques.

Construire le test d'hypothèse

La première étape de la construction d'un test d'hypothèse consiste à déterminer clairement ce que sont les hypothèses nulles et alternatives. Ce n'est pas trop difficile à faire. Notre hypothèse nulle, H_0, est que la moyenne réelle de la population μ pour les notes des étudiants en psychologie est de 67,5, et notre hypothèse alternative est que la moyenne de la population *n'est pas de* 67,5. Si nous écrivons ceci en notation mathématique, ces hypothèses deviennent :

$$\text{H}_{0}: \mu = 67,5\\ \text{H}_{1}: \mu\neq 67,5$$

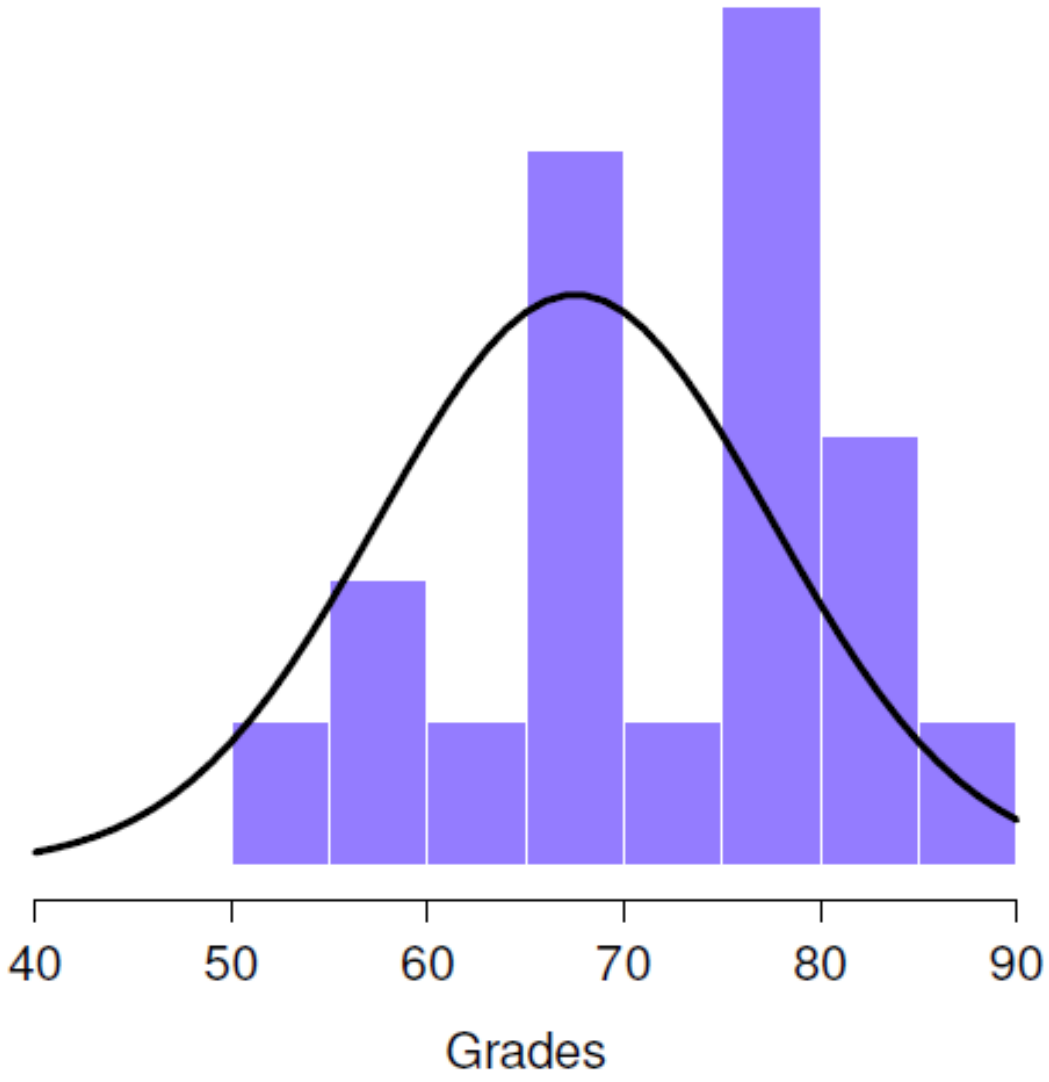


Figure 11-1 : La distribution théorique (ligne pleine) à partir de laquelle les notes des étudiants en psychologie (barres) sont censées avoir été générées.

bien que pour être honnête, cette notation n'ajoute pas grand-chose à notre compréhension du problème, c'est juste une façon compacte d'écrire ce que nous essayons d'apprendre à partir des données. Les hypothèses nulles H_0 et l'hypothèse alternative H_1 de notre test sont toutes deux illustrées à la Figure 11-2. En plus de nous fournir ces hypothèses, le scénario décrit ci-dessus nous fournit une bonne quantité de connaissances de base qui pourraient être utiles. Plus précisément, il y a deux éléments d'information particuliers que nous pouvons ajouter :

1. Les notes en psychologie sont normalement distribuées.
2. L'écart-type réel de ces scores σ est connu pour être 9,5.

Pour l'instant, nous allons agir comme si ce sont des faits absolument dignes de confiance. Dans la vraie vie, ce genre de connaissances de base absolument dignes de confiance n'existe pas, et donc si nous voulons nous fier à ces faits, nous devrons simplement *supposer que* ces choses sont vraies. Toutefois, comme ces hypothèses peuvent être justifiées ou non, nous devrons peut-être les vérifier. Mais pour l'instant, nous nous allons faire simple.

L'étape suivante consiste à déterminer ce qui considérerions comme un bon choix pour une statistique de test, quelque chose qui nous aiderait à faire la distinction entre H_0 et H_1. Étant donné que les hypothèses se réfèrent toutes à la moyenne de la population μ, vous seriez assez confiant que la moyenne de l'échantillon $\overline{X}$ constituerait un point de départ très utile. Ce que nous pourrions faire, c'est examiner la différence entre la moyenne de l'échantillon $\overline{X}$ et la valeur moyenne que l'hypothèse nulle prédit pour la population.

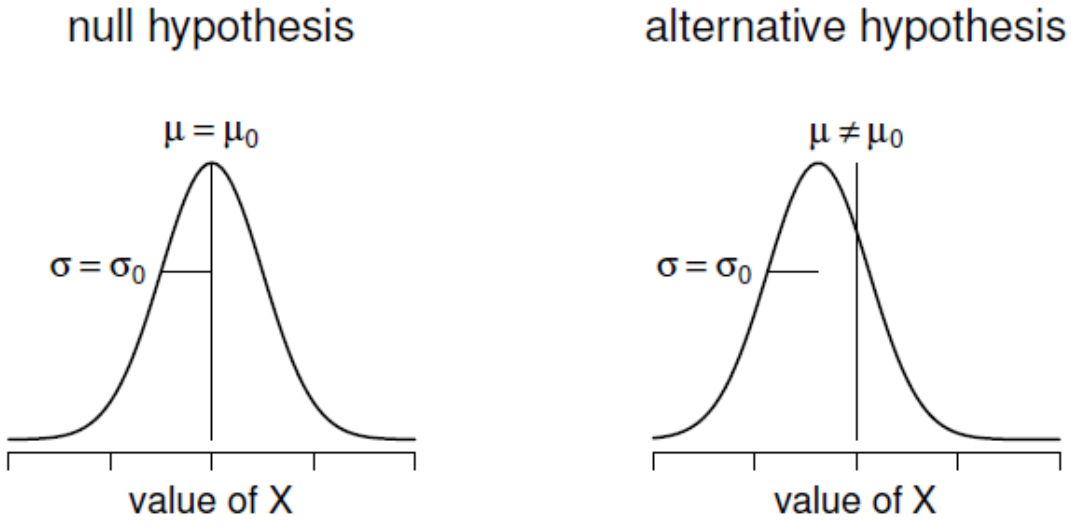


Figure 11-2 : Illustration graphique de l'hypothèse nulle et de l'hypothèse alternative supposée par le z-test à un échantillon. L'hypothèse nulle et l'hypothèse alternative supposent toutes deux que la distribution de la population est normale et supposent en outre que l'écart-type de la population est connu (fixé à une certaine valeur σ_0). L'hypothèse nulle (à gauche) est que la moyenne de population μ est égale à une valeur donnée μ_0. L'hypothèse alternative est que la moyenne de population diffère de cette valeur, $\mu - \mu_0$.

Dans notre exemple, cela signifie que nous calculons $\overline{X} - 67{,}5$. Plus généralement, si nous posons que μ_0 correspond à la valeur de la moyenne de notre population selon l'hypothèse nulle, alors nous devrions calculer

$$\overline{X} - \mu_0$$

Si cette quantité est égale ou très proche de 0, les choses s'annoncent bien pour l'hypothèse nulle. Si cette quantité est très éloignée de 0, il est moins probable que l'hypothèse nulle

vaille la peine d'être retenue. Mais à quelle distance de zéro doit-on s'éloigner pour rejeterH_0?

Pour comprendre cela, nous devons être un peu plus rusés, et nous devons faire le lien avec deux connaissances de base que j'ai présentées précédemment, à savoir que les données brutes sont normalement distribuées et que nous connaissons la valeur de l'écart-type de la population σ. Si l'hypothèse nulle est vraie et que la vraie moyenne est μ_0, alors ces faits réunis nous indiquent la distribution de la population : une distribution normale ayant pour moyenne μ_0 et pour écart-type σ. En adoptant la notation de la section 7.5, un statisticien pourrait écrire ceci sous la forme :

$$X \sim Normale(\mu_0, \sigma^2)$$

Bien, si c'est vrai, alors que dire de la distribution de $\overline{X}$? Comme nous l'avons mentionné précédemment (voir la section 8.3.3), la distribution d'échantillonnage de la moyenne $\overline{X}$ est également normale et a une moyenne μ. Mais l'écart-type de cette distribution d'échantillonnage $SE(\overline{X})$, qu'on appelle l'*erreur type de la moyenne*, est

$$SE(\overline{X}) = \frac{\sigma}{\sqrt{N}}$$

En d'autres termes, si l'hypothèse nulle est vraie, la distribution d'échantillonnage de la moyenne peut s'écrire comme suit :

$$\overline{X} \sim Normale(\mu_0, SE(\overline{X}))$$

Maintenant, le tour est joué. Ce que nous pouvons faire, c'est convertir la moyenne de l'échantillon $\overline{X}$ en un score standard (Section 4.5). Cela s'écrit conventionnellement comme z, mais pour l'instant je vais le noter $z_{\overline{X}}$. (La raison de l'utilisation de cette notation étendue est de vous aider à vous rappeler que nous calculons une version standardisée d'une moyenne d'échantillon, *et non* une version standardisée d'une seule observation, ce à quoi un *z-score* fait habituellement référence). Lorsque nous le faisons, le *score z de* la moyenne de notre échantillon est le suivant

$$z_{\overline{X}} = \frac{\overline{X} - \mu_0}{SE(\overline{X})}$$

ou, de manière équivalente

$$z_{\overline{X}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{N}}$$

Ce *z-score* est notre statistique de test. Ce qu'il y a d'intéressant à utiliser cette statistique comme statistique de test, c'est que, comme tous les *z-scores*, elle a une distribution normale standard :

$$z \sim Normale(1,0)$$

(encore une fois, voir la Section 4.5 si vous avez oublié pourquoi c'est vrai). En d'autres termes, quelle que soit l'échelle sur laquelle se trouvent les données d'origine, la *statistique z* elle-même a toujours la même interprétation : elle est égale au nombre d'erreurs-types qui séparent la moyenne observée de l'échantillon $\overline{X}$ de la moyenne de la population μ_0 prévue par l'hypothèse nulle. Mieux encore, quels que soient les paramètres de population pour les scores bruts, les 5 % de régions critiques pour le *test z* sont toujours les mêmes, comme l'illustre la Figure 11-3. Et ce que cela signifiait, à l'époque où les gens faisaient toutes leurs statistiques à la main, c'est que quelqu'un pouvait publier un tableau comme celui-ci :

	valeur critique z	
niveau de significativité α	test bilatéral	test unilatéral
.1	1.644854	1.281552
.05	1.959964	1.644854
.01	2.575829	2.326348
.001	3.290527	3.090232

Cela permettait aux chercheurs de calculer leur *statistique z* à la main et ensuite rechercher la valeur critique dans un manuel scolaire.

Un exemple travaillé à la main

Comme je l'ai mentionné plus tôt, le *z-test* n'est presque jamais utilisé dans la pratique. Il est si rarement utilisé dans la vie réelle que l'installation de base de Jamovi n'a pas de fonction intégrée pour cela. Cependant, le test est si simple qu'il est très facile de le faire manuellement. Revenons aux données de la classe du Dr Zeppo. Après avoir chargé les données des notes, la première chose que je dois faire est de calculer la moyenne de l'échantillon, ce que j'ai déjà fait (72,3). Nous avons déjà la norme de population connue ($\sigma = 9{,}5$), et la valeur de la population signifie que l'hypothèse nulle spécifie ($\mu_0 = 67{,}5$), et on connaît la taille de l'échantillon (N=20).

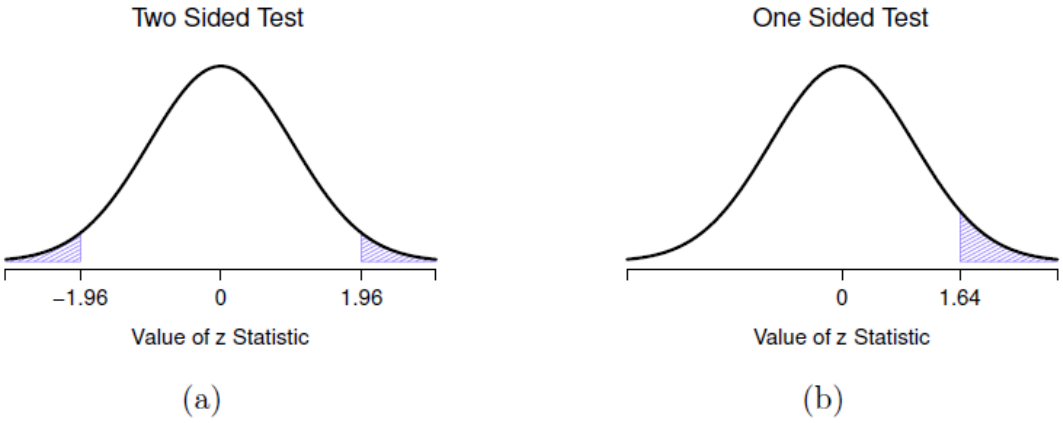


Figure 11-3 : Régions de rejet pour le test z bilatéral (panneau a) et le test z unilatéral (panneau b).

Ensuite, calculons l'erreur type (vraie) de la moyenne (facile à faire avec une calculatrice) :

$$\begin{aligned} \text{sem.true} &= \text{sd.true/sqrt} \\ &= 9{,}5/\text{sqrt} \\ &= 2{,}124265 \end{aligned}$$

Et enfin, nous calculons notre *z-score* :

$$\begin{aligned} \text{z.score} &= (\text{sample.mean} - \text{mu.null})/\text{sem.true} \\ &= (72{,}3 - 67{,}5)/2{,}124265 \\ &= 2{,}259606 \end{aligned}$$

À ce stade, nous chercherions traditionnellement la valeur 2,26 dans notre tableau des valeurs critiques. Notre hypothèse initiale était bilatérale (nous n'avions pas vraiment de théorie expliquant pourquoi les étudiants en psychologie seraient meilleurs ou pires en statistiques que les autres étudiants), alors notre test d'hypothèse est bilatéral. En regardant le petit tableau que j'ai présenté plus tôt, on constate que 2,26 est supérieur à la valeur critique de 1,96 qui devrait être significative à $\alpha = .05$, mais inférieure à la valeur de 2,58 qui devrait être significative à un niveau de $\alpha = .01$. Par conséquent, nous pouvons conclure que nous avons un effet significatif, que nous pourrions écrire en disant quelque chose comme ceci :

Avec une note moyenne de 73,2 dans l'échantillon d'étudiants en psychologie, et en supposant un écart type de population réel de 9,5, nous pouvons conclure que les étudiants en psychologie ont des scores statistiques significativement différents de la moyenne de la classe ($z = 2{,}26, N = 20, p < .05$).

Hypothèses du *z-test*

Comme je l'ai déjà dit, tous les tests statistiques font des hypothèses. Certains tests font des hypothèses raisonnables, alors que d'autres ne le font pas. Le test que je viens de décrire, le *test z-test* à un échantillon, fait trois hypothèses de base. Celles-ci le sont :

- *Normalité*. Comme on le décrit habituellement, le *test z* suppose que la distribution réelle de la population est normale.[80] C'est souvent une hypothèse assez raisonnable, et c'est aussi une hypothèse que nous pouvons vérifier si nous nous en inquiétons (voir Section 11.8).

- *Indépendance*. La deuxième hypothèse du test est que les observations de votre ensemble de données ne sont pas corrélées entre elles, ou reliées entre elles. Ce n'est pas aussi facile à vérifier statistiquement, cela repose un peu sur un bon plan expérimental. Un exemple évident (et stupide) d'une situation qui viole cette hypothèse est un ensemble de données où vous « copiez » la même observation encore et encore dans votre fichier de données de sorte que vous obtenez une « taille d'échantillon » massive, qui consiste en une seule observation réelle. De façon plus réaliste, vous devez vous demander s'il est vraiment plausible d'imaginer que chaque observation est un échantillon complètement aléatoire de la population qui vous intéresse. Dans la pratique, cette hypothèse n'est jamais respectée, mais nous faisons de notre mieux pour concevoir des études qui minimisent les problèmes de corrélation des données.

- *Écart-type connu*. La troisième hypothèse du *test z* est que le chercheur connaît l'écart-type réel de la population. C'est juste stupide. Dans aucun problème réel d'analyse de données, vous connaissez l'écart-type σ de certaines populations, tout en ignorant complètement la moyenne μ. En d'autres termes, cette hypothèse est *toujours* fausse.

Compte tenu de la stupidité de supposer que σ est connu, voyons si nous pouvons nous en passer. Cela nous emmène hors du domaine morne du *test z*, et dans le royaume magique du *test t*, avec des licornes, des fées et des lutins !

Le *test t sur* un échantillon

Après mûre réflexion, j'ai décidé qu'il ne serait peut-être pas prudent de supposer que les notes des étudiants en psychologie ont nécessairement le même écart-type que les autres étudiants de la classe du Dr Zeppo. Après tout, si je fais l'hypothèse qu'ils n'ont pas la même moyenne, alors pourquoi devrais-je croire qu'ils ont absolument le même écart-type ? Compte tenu de cela, je devrais vraiment arrêter de supposer que je connais la vraie valeur de σ. Cela viole les suppositions de mon *test z*, donc dans un sens, je suis de retour à la case départ. Cependant, ce n'est pas comme si j'étais complètement à court d'options. Après tout,

---

[80] En fait, c'est trop fort. Strictement parlant, le test z exige seulement que la distribution d'échantillonnage de la moyenne soit normalement distribuée. Si la population est normale, il s'ensuit nécessairement que la distribution d'échantillonnage de la moyenne est également normale. Cependant, comme nous l'avons vu en parlant du théorème de la limite centrale, il est tout à fait possible (voire banal) que la distribution d'échantillonnage soit normale même si la distribution de la population elle-même n'est pas normale. Cependant, à la lumière du ridicule de l'hypothèse selon laquelle l'écart-type véritable est connu, il n'y a vraiment pas beaucoup de raison d'entrer dans les détails à ce sujet !

j'ai toujours mes données brutes, et ces données brutes me donnent une *estimation de l'*écart-type de la population, qui est de 9,52. En d'autres termes, si je ne peux pas dire que je sais que $\sigma = 9{,}5$, je *peux* dire que $\hat{\sigma} = 9{,}52$.

Bien ! Une solution évidente à laquelle vous pourriez penser est d'effectuer un *test z*, mais en utilisant l'écart type estimé de 9,52 au lieu de vous fier à mon hypothèse de l'écart type réel de 9,5. Et vous ne seriez probablement pas surpris d'apprendre que cela nous donnerait quand même un résultat significatif. Cette solution est proche, mais elle n'est pas *tout à fait* correcte. Comme nous nous fions maintenant à une *estimation de* l'écart-type de la population, nous devons faire un certain ajustement pour tenir compte du fait que nous avons une certaine incertitude quant à la véritable valeur de l'écart-type de la population. Peut-être que nos données ne sont qu'un coup de chance… peut-être que l'écart-type réel de la population est de 11, par exemple. Mais si c'était vrai, et nous ayons réalisé le *test* z en supposant que $\sigma = 11$, alors le résultat serait *non significatif*. C'est un problème, et c'est un problème que nous allons devoir régler.

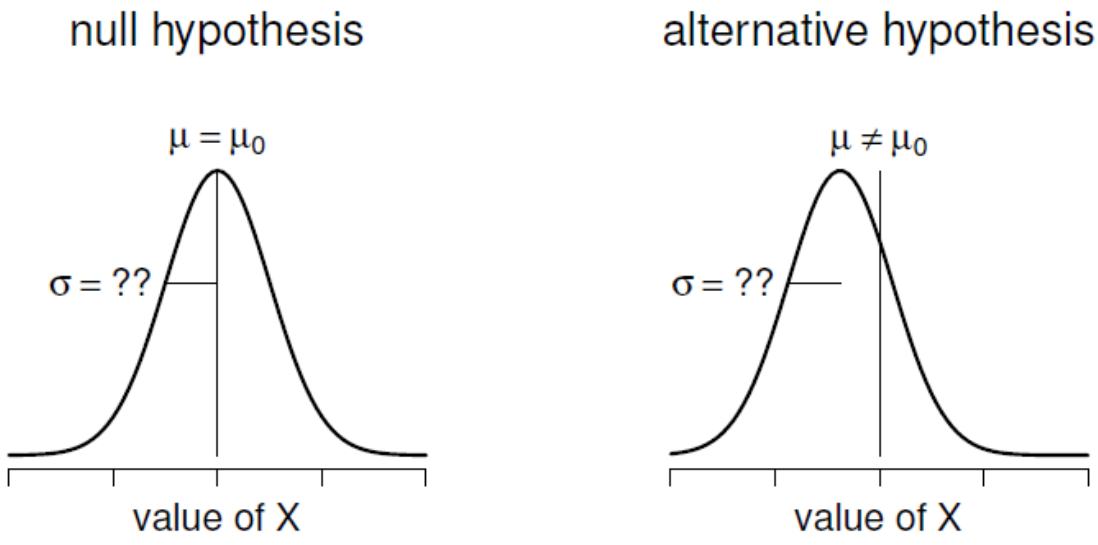


Figure 11-4 : Illustration graphique de l'hypothèse nulle et de l'hypothèse alternative supposée par le test t pour un échantillon (bilatéral). Notez la similitude avec le test z (Figure 11-2). L'hypothèse nulle est que la moyenne de population μ est égale à une valeur spécifiée μ_0, et l'hypothèse alternative est que ce n'est pas le cas. Comme pour le test z, nous supposons que les données sont normalement distribuées, mais nous ne supposons pas que l'écart type de population σ est connu à l'avance.

Présentation du *test t-test*

Cette ambiguïté est agaçante, et elle a été résolue en 1908 par un type appelé William Sealy Gosset (Student 1908), qui travaillait alors comme chimiste à la brasserie Guinness (voir J. F. Box (1987)). Parce que Guinness avait une mauvaise opinion de ses employés qui publiaient des analyses statistiques (apparemment ils pensaient que c'était un secret

commercial), il a publié l'ouvrage sous le pseudonyme « A Student » et, à ce jour, le nom complet du *t-test* est en fait le **test t de Student**.

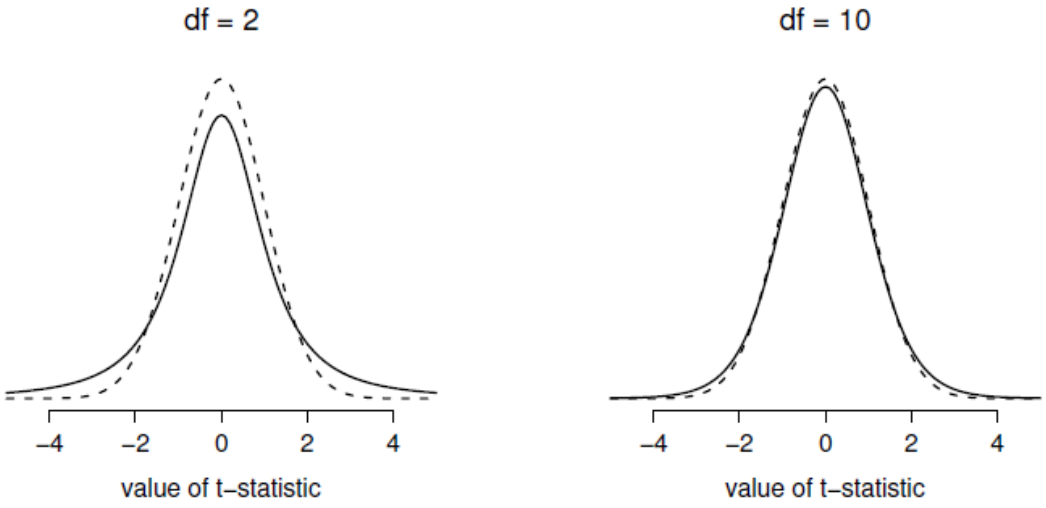


Figure 11-5 : La distribution t avec 2 degrés de liberté (à gauche) et 10 degrés de liberté (à droite), avec une distribution normale standard (c.-à-d. moyenne 0 et écart-type 1) représentée par des lignes pointillées à des fins de comparaison. Notez que la distribution t a des queues plus importantes (leptocurtique : kurtosis plus élevé) que la distribution normale ; cet effet est assez exagéré lorsque les degrés de liberté sont très faibles, mais négligeable pour des valeurs plus grandes. En d'autres termes, pour un grand df, la distribution t est pratiquement identique à une distribution normale.

Ce que Gosset a compris, c'est qu'il faut tenir compte du fait que nous ne savons pas exactement quel est l'écart-type réel.[81] La réponse est qu'il modifie subtilement la distribution d'échantillonnage. Dans le *test t,* notre statistique de test, maintenant appelée *statistique t*, est calculée exactement de la même manière que je l'ai exposé ci-dessus. Si notre hypothèse nulle est que la vraie moyenne est μ, mais que notre échantillon a une moyenne $\overline{X}$ et que notre estimation de l'écart-type de la population est $\hat{\sigma}$, alors notre statistique *t* est :

$$t = \frac{\overline{X} - \mu}{\hat{\sigma} / \sqrt{N}}$$

La seule chose qui a changé dans l'équation est qu'au lieu d'utiliser la valeur réelle connue σ, nous utilisons l'estimation $\hat{\sigma}$. Et si cette estimation a été construite à partir de N

[81] Si j'ai bien compris l'histoire, Gosset n'a fourni qu'une solution partielle ; la solution générale au problème a été fournie par Sir Ronald Fisher.

observations, alors la distribution d'échantillonnage se transforme en une *distribution t* avec *N - 1* degrés de liberté* (df). La distribution *t* est très semblable à la distribution normale, mais a des queues plus « importantes », comme on l'a vu plus haut à la section 7.6 et illustré à la Figure 11-5. Notez, cependant, qu'à mesure que df s'agrandit, la *distribution t* tend à être identique à la distribution normale standard. C'est comme il se doit : si vous avez un échantillon de *N* = 70 000 000 000, alors votre « estimation » de l'écart-type devrait être à peu près parfaite. Il faut donc s'attendre à ce que pour un grand *N*, le *t-test* se comporte exactement de la même manière qu'un *test z*. Et c'est exactement ce qui se passe !

Faire le test avec Jamovi

Comme on peut s'y attendre, la mécanique du *test t* est presque identique à celle du *test z*. Il n'y a donc pas beaucoup d'intérêt à faire l'exercice fastidieux de vous montrer comment faire les calculs à l'aide de commandes de bas niveau. C'est à peu près identique aux calculs que nous avons faits plus tôt, sauf que nous utilisons l'écart-type estimé et que nous testons ensuite notre hypothèse en utilisant la distribution *t* plutôt que la distribution normale. Ainsi, au lieu de passer en revue les calculs en détail fastidieux pour une deuxième fois, je vais vous montrer comment les *tests t* sont réellement effectués. Jamovi est livré avec une analyse dédiée pour les *tests t* qui est très flexible (il peut exécuter de nombreux types de *tests t* différents). C'est assez simple à utiliser ; tout ce que vous avez à faire est de sélectionner « Analyses » - « T-Tests » - « One Sample T-Test », déplacer la variable qui vous intéresse (X) dans la case « Variables », et taper la valeur moyenne de l'hypothèse nulle « 67.5 » dans la case « Hypothesis » -« Test value ». C'est assez facile. Voir Figure 11-6, qui, entre autres choses que nous allons aborder dans un instant, vous donne une statistique *t-test = 2,25*, avec 19 degrés de liberté et une *valeur p* associée de 0,036.

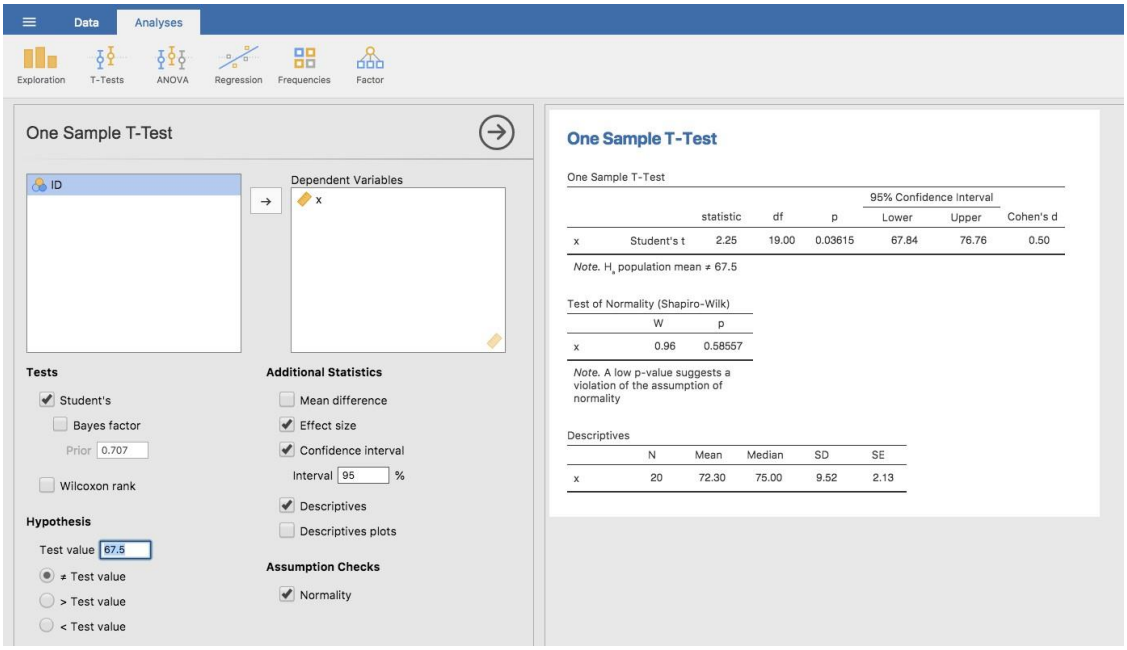


Figure 11-6 : Jamovi effectue le test t sur un échantillon.

Il y a aussi deux autres choses qui pourraient vous intéresser : l'intervalle de confiance à 95 % et une mesure de la taille de l'effet (on en reparlera plus loin). Cela semble donc assez simple. Maintenant, que *faisons-nous* de cette sortie ? Eh bien, nous sommes ravis de découvrir que le résultat est statistiquement significatif (c'est-à-dire que la *valeur p* est inférieure à 0,05). Nous pourrions rapporter le résultat en disant quelque chose comme ceci :

Avec une note moyenne de 72,3, les étudiants en psychologie ont obtenu une note légèrement supérieure à la note moyenne de 67,5 (*t(19)=2,25, p<.05)*; l'intervalle de confiance à 95 % est de 67,8 à 76.8.

où *t(19)* est une notation condensée pour une *statistique t* qui a 19 degrés de liberté. Cela dit, il arrive souvent que les gens ne signalent pas l'intervalle de confiance, ou qu'ils le fassent en utilisant une forme beaucoup plus comprimée que ce que j'ai fait ici. Par exemple, il n'est pas rare de voir l'intervalle de confiance inclus dans le bloc statistique, comme ceci :

$$t(19) = 2{,}25, p < .05, C_{I95} = [67{,}8; 76{,}8]$$

Avec tout ce jargon entassé dans une demi-ligne, vous savez qu'il doit être très intelligent.[82]

Hypothèses d'un *test t sur* un échantillon

Bien, quelles sont les hypothèses du *test t sur* un échantillon ? Eh bien, puisque le *test t* est fondamentalement un *test z* sans l'hypothèse d'écart type connu, vous ne devriez pas être surpris de voir qu'il fait les mêmes hypothèses que le *test z*, moins celle de l'écart type connu. C'est-à-dire

- *Normalité*. Nous supposons toujours que la distribution de la population est normale[83] et, comme nous l'avons déjà mentionné, il existe des outils standards que vous pouvez

---

[82] Plus sérieusement, j'ai tendance à penser que l'inverse est vrai. Je me méfie beaucoup des rapports techniques qui remplissent leurs sections de résultats avec rien d'autre que des chiffres. C'est peut-être juste que je suis un crétin arrogant, mais j'ai souvent l'impression d'être un auteur qui ne fait aucun effort pour expliquer et interpréter son analyse au lecteur, ou qui ne la comprend pas lui-même, ou qui est un peu paresseux. Vos lecteurs sont intelligents, mais pas infiniment patients. Ne les ennuyez pas si vous pouvez l'éviter.

[83] Un commentaire technique. De la même façon que nous pouvons affaiblir les hypothèses du *test z pour* ne parler que de la distribution d'échantillonnage, nous *pouvons* affaiblir les hypothèses du *test t pour* ne pas avoir à supposer la normalité de la population. Cependant, pour le *t-test,* c'est plus difficile à faire. Comme précédemment, nous pouvons remplacer l'hypothèse de normalité de la population par une hypothèse selon laquelle la distribution d'échantillonnage de $\overline{X}$ est normale. Cependant, n'oubliez pas que nous nous basons également sur une estimation d'échantillon de l'écart-type, et nous exigeons donc aussi que la distribution d'échantillonnage de $\hat{\sigma}$ soit le chi carré. Cela rend les choses plus désagréables, et cette version est rarement utilisée dans la pratique. Heureusement, si la répartition de la population est normale, ces deux hypothèses sont satisfaites.

utiliser pour vérifier si cette hypothèse est remplie (Section 11.8, et d'autres tests que vous pouvez faire à sa place si cette hypothèse est violée (Section 11.9).

- *Indépendance*. Encore une fois, nous devons supposer que les observations de notre échantillon sont générées indépendamment les unes des autres. Voir la discussion précédente sur le *test z* pour les spécificités (Section 11.1.4).

Dans l'ensemble, ces deux hypothèses ne sont pas très déraisonnables et, par conséquent, le *test t pour un échantillon* est assez largement utilisé dans la pratique pour comparer la moyenne d'un échantillon à une moyenne de population supposée.

Les tests de student pour échantillons indépendants

Bien que le *test t-test pour un échantillon* ait ses utilisations, ce n'est pas l'exemple le plus typique d'un *test t*[84]. Une situation beaucoup plus courante survient lorsque vous avez deux groupes d'observations différents. En psychologie, cela tend à correspondre à deux groupes différents de participants, où chaque groupe correspond à une condition différente dans votre étude. Pour chaque personne participant à l'étude, vous mesurez une variable d'intérêt, et la question de recherche que vous posez est de savoir si les deux groupes ont ou non la même moyenne de population. C'est la situation pour laquelle le *t-test des* échantillons indépendants est conçu.

Les données

Supposons que 33 étudiants suivent les cours du Dr Harpo sur les statistiques, et que le Dr Harpo n'est pas noté selon une courbe. En fait, la notation du Dr Harpo est un peu mystérieuse, nous ne savons donc pas vraiment quelle est la note moyenne pour l'ensemble de la classe. Il y a deux tuteurs pour la classe, Anastasia et Bernadette. Il y a $N_1=15$ élèves dans les travaux dirigés d'Anastasia, et $N_2=18$ dans ceux de Bernadette. La question de recherche qui m'intéresse est de savoir si Anastasia ou Bernadette est une meilleure tutrice, ou si cela ne fait pas une grande différence. Le Dr Harpo m'a envoyé les notes de cours par e-mail, dans le fichier harpo.csv. Comme d'habitude, je vais charger le fichier dans Jamovi et regarder quelles variables il contient - il y a trois variables, ID, grade et tuteur. La variable de note contient la note de chaque élève, mais elle n'est pas importée dans Jamovi avec l'attribut de niveau de mesure correct, donc je dois la modifier pour qu'elle soit considérée comme une variable continue (voir section 3.6). La variable du tuteur est un facteur qui indique qui était le tuteur de chaque élève - soit Anastasia ou Bernadette.

Nous pouvons calculer les moyennes et les écarts-types à l'aide de l'analyse « Exploration » -« Descriptives », et voici un joli petit tableau récapitulatif :

	mean	std dev	N
Les élèves d'Anastasia	74,53	9,00	15

[84] Bien que ce soit le plus simple, c'est pourquoi j'ai commencé par lui.

Les élèves de Bernadette	69,06	5,77	18

Pour vous donner une idée plus détaillée de ce qui se passe ici, j'ai tracé des histogrammes (non pas en Jamovi, mais en utilisant R) montrant la distribution des notes pour les deux tuteurs (Figure 11-7), ainsi qu'un graphique plus simple montrant les moyennes et les intervalles de confiance correspondants des deux groupes d'élèves (Figure 11-8).

Présentation du test

Le **t-test d'échantillons indépendants** se présente sous deux formes différentes, celle de Student et celle de Welch. Le *t-test* original de Student, qui est celui que je vais décrire dans cette section, est le plus simple des deux mais repose sur des hypothèses beaucoup plus restrictives que le *test t de Welch*.

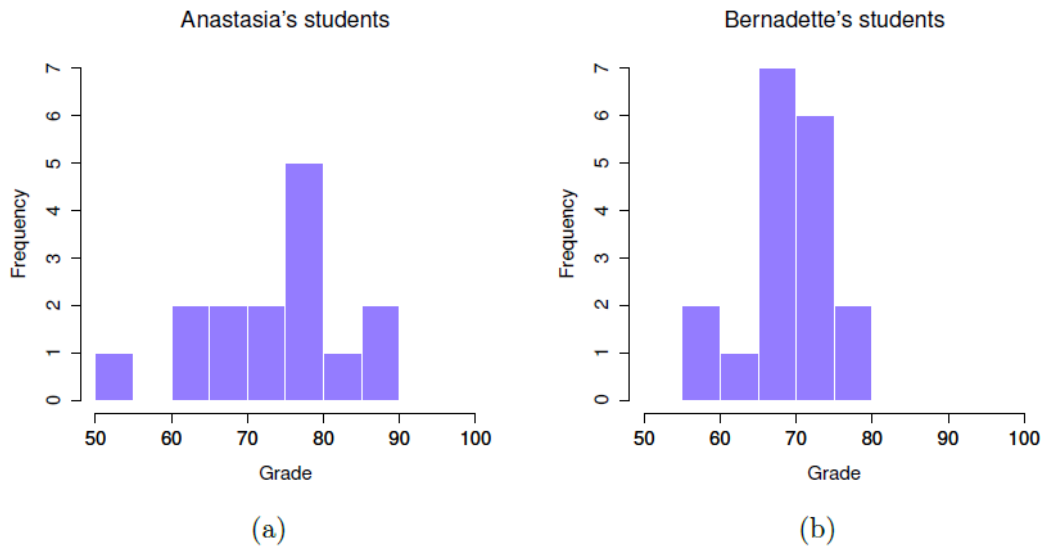


Figure 11-7 : Histogrammes montrant la distribution des notes des élèves des classes d'Anastasia (figure a) et de Bernadette (figure b). Visuellement, cela suggère que les élèves de la classe d'Anastasia peuvent obtenir de meilleures notes en moyenne, bien qu'elles semblent aussi un peu plus dispersées.

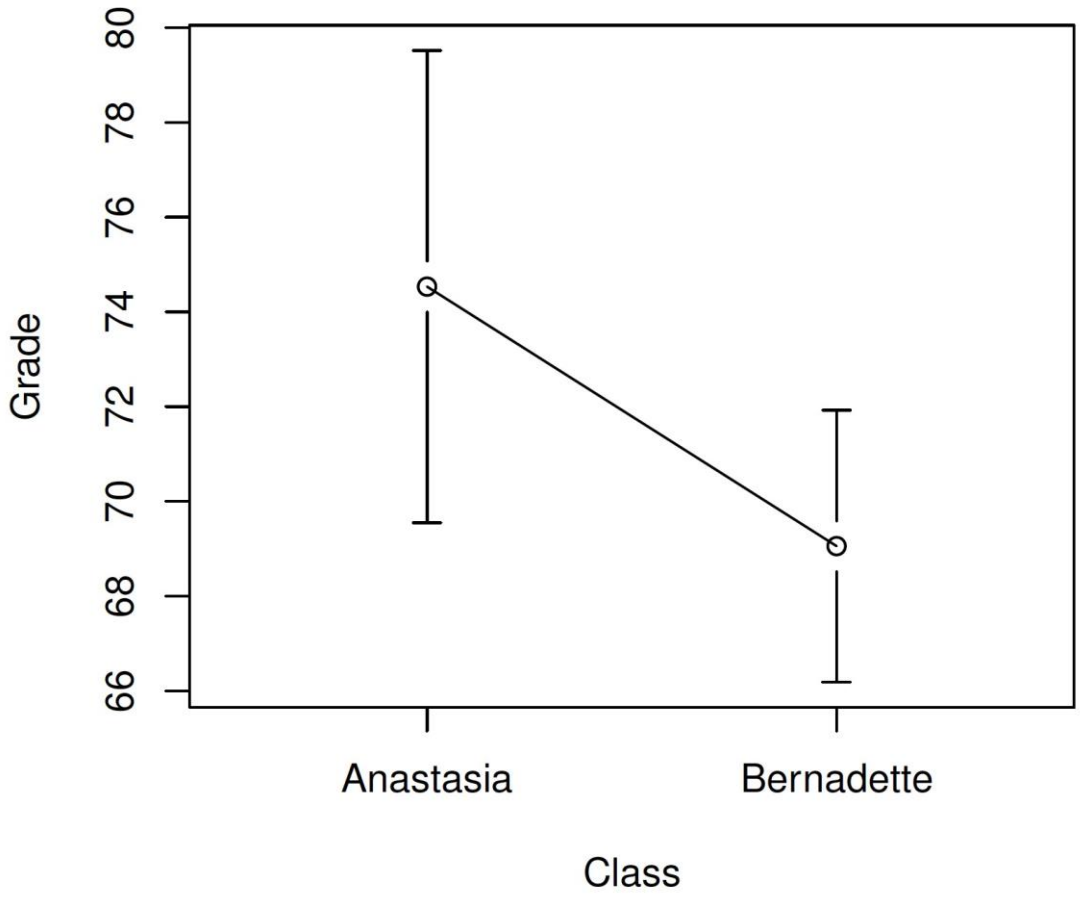


Figure 11-8 : Le graphique illustre la note moyenne des élèves des travaux dirigées d'Anastasia et de Bernadette. Les barres d'erreur représentent des intervalles de confiance à 95 % autour de la moyenne. Visuellement, on dirait qu'il y a une vraie différence entre les groupes, mais c'est difficile à dire avec certitude.

En supposant pour l'instant que l'on veuille effectuer un test bilatéral, le but est de déterminer si deux « échantillons indépendants » de données sont tirés de populations ayant la même moyenne (l'hypothèse nulle) ou des moyennes différentes (l'hypothèse alternative). Lorsque nous parlons d'échantillons « indépendants », ce que nous voulons dire ici, c'est qu'il n'y a pas de relation particulière entre les observations des deux échantillons. Cela n'a probablement pas beaucoup de sens pour l'instant, mais ce sera plus clair lorsque nous parlerons des tests t pour échantillons appariés plus tard. Pour l'instant, signalons simplement que si nous avons un plan expérimental où les participants sont répartis au hasard dans l'un des deux groupes et que nous voulons comparer la performance moyenne des deux groupes en fonction d'une mesure des résultats, alors c'est un *test t* pour échantillons indépendants (plutôt qu'un *test t apparié*) qui nous intéresse.

Ok, donc posons μ_1 pour indiquer la vraie moyenne de population pour le groupe 1 (par exemple, les étudiants d'Anastasia), et μ_2 sera la vraie moyenne de population pour le

groupe 2 (par exemple, les étudiants de Bernadette),[85] et comme d'habitude nous laissons $\overline{X_1}$ et $\overline{X_2}$ indiquer les moyennes observées pour ces deux groupes. Notre hypothèse nulle indique que les deux moyennes de population sont identiques ($\mu_1 = \mu_2$) et l'alternative est qu'elles ne le sont pas ($\mu_1 \neq \mu_2$). Voici cela écrit en termes mathématique :

$$\text{H}_{0}=\mu_{1} = \mu_{2})\\ \text{H}_{1}=\mu_{1} \neq \mu_{2})$$

Pour construire un test d'hypothèse qui traite ce scénario, nous commençons par noter que si l'hypothèse nulle est vraie, alors la différence entre les moyennes de population est *exactement* nulle, $\mu_1 - \mu_2 = 0$. Par conséquent, une statistique de test diagnostique sera basée sur la différence entre les deux moyennes de l'échantillon. Parce que si l'hypothèse nulle est vraie, on s'attendrait à ce que $\overline{X_1} - \overline{X_2}$ soit *assez proche* de zéro. Cependant, tout comme nous l'avons vu avec nos tests sur un échantillon (c.-à-d. le *test z sur un échantillon* et le *test t sur un échantillon*), nous devons être précis quant à savoir exactement dans quelle mesure cette différence devrait être *proche* de zéro. Et la solution au problème est plus ou moins la même. Nous calculons une estimation de l'erreur-type (SE pour standard error), comme la dernière fois, puis nous divisons la différence entre les moyennes par cette estimation. Notre *statistique T* sera donc de la forme :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\text{SE}}$$

Nous avons juste besoin de savoir ce qu'est réellement cette estimation de l'erreur type. C'est un peu plus délicat que pour l'un ou l'autre des deux tests que nous avons examinés jusqu'à présent, et nous devons donc l'examiner beaucoup plus attentivement pour comprendre comment il fonctionne.

Une « estimation globale » de l'écart-type

Dans le « test t de Student *t-test* » original, nous partons du principe que les deux groupes ont le même l'écart-type de la population. C'est-à-dire que, peu importe si les moyennes de

---

[85] Une drôle de question surgit presque toujours à ce moment-là : à quoi diable se réfère-t-on dans ce cas-ci ? Est-ce l'ensemble des étudiants qui suivent les cours du Dr Harpo (tous les 33) ? L'ensemble des personnes qui pourraient suivre le cours (un nombre inconnu d'entre elles) ? Ou autre chose ? Est-ce que ça a de l'importance de savoir ce que nous choisissons ? C'est la tradition dans un cours d'introduction aux statistiques comportementales de marmonner beaucoup à ce stade, mais comme mes élèves me posent cette question chaque année, je vais donner une brève réponse. Techniquement oui, c'est important. Si vous changez votre définition de ce qu'est réellement la population dans le "monde réel", alors la distribution d'échantillonnage de votre moyenne observée change aussi. Le test t repose sur l'hypothèse que les observations sont échantillonnées au hasard dans une population infiniment grande et, dans la mesure où la vie réelle n'est pas comme ça, le test t peut être erroné. Dans la pratique, cependant, ce n'est généralement pas grand-chose. Même si l'hypothèse est presque toujours fausse, elle ne provoque pas beaucoup de biais du test, alors nous avons tendance à l'ignorer.

population sont les mêmes, nous supposons que les écarts-types de population sont identiques, $\sigma_1 = \sigma_2$, l'écart-type de la population.

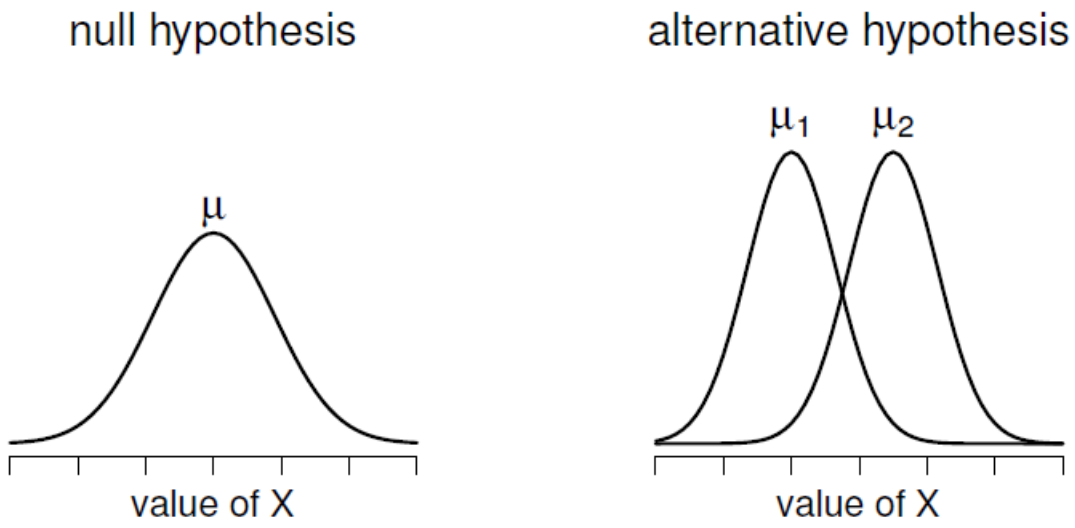


Figure 11-9 : Illustration graphique des hypothèses nulle et alternative supposées par le test t de Student. L'hypothèse nulle suppose que les deux groupes ont la même moyenne μ, alors que l'hypothèse alternative suppose qu'ils ont des moyennes différentes μ_1 et μ_2. Notez qu'on suppose que les distributions de la population sont normales et que, bien que l'hypothèse alternative permette au groupe d'avoir des moyennes différentes, elle suppose qu'ils ont le même écart-type.

C'est-à-dire que, peu importe si les moyennes de population sont les mêmes, nous supposons que les écarts-types de population sont identiques, $\sigma_1 = \sigma_2$. Puisque nous supposons que les deux écarts-types sont les mêmes, nous supprimons les indices et les appelons tous les deux σ. Comment devrions-nous les estimer ? Comment construire une estimation unique d'un écart-type lorsque nous avons deux échantillons ? La réponse est qu'en gros, nous en faisons la moyenne. Enfin, en quelque sorte. En fait, nous prenons une moyenne *pondérée des* estimations de la *variance*, que nous utilisons comme **estimation globale de la variance**. Le poids attribué à chaque échantillon est égal au nombre d'observations dans cet échantillon, moins 1.

Mathématiquement, nous pouvons écrire ceci comme suit

$$w_{1}=N_{1}-1\\ w_{2}=N_{2}-1$$

Maintenant que nous avons attribué des pondérations à chaque échantillon, nous calculons l'estimation globale de la variance en prenant la moyenne pondérée des deux estimations de variance, $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$

$$\hat{\sigma}_p^2 = \frac{w_1 \hat{\sigma}_1^2 + w_2 \hat{\sigma}_2^2}{w_1 + w_2}$$

Enfin, nous convertissons l'estimation de la variance globale en une estimation de l'écart-type regroupée, en prenant la racine carrée.

$$\hat{\sigma}_p^2 = \sqrt{\frac{w_1 \hat{\sigma}_1^2 + w_2 \hat{\sigma}_2^2}{w_1 + w_2}}$$

Et si vous remplacez mentalement $w_1 = N_1 - 1$ et $w_2 = N_2 - 1$ dans cette équation, vous obtenez une formule très moche. Une formule très moche qui semble en fait être la façon « standard » de décrire l'estimation de l'écart-type globale. Ce n'est toutefois pas ma façon préférée de voir les écarts-types globaux. Je préfère voir les choses comme ça. Notre ensemble de données correspond en fait à un ensemble de *N* observations qui sont classées en deux groupes. Utilisons donc la notation X_{ik} pour faire référence à la note reçue par le *i-ème* élève dans le *k-ème* groupe de tutorat. C'est-à-dire, X_{11} est la note reçue par le premier élève de la classe d'Anastasia, X_{21} est son deuxième élève, et ainsi de suite. Et nous avons deux moyennes de groupe séparées $\overline{X}_1$ et $\overline{X}_2$, que nous pourrions « génériquement » désigner en utilisant la notation $\overline{X}_k$, c'est-à-dire la note moyenne pour le *k-ème* groupe de tutorat. Pour l'instant, tout va bien. Maintenant, puisque chaque élève tombe dans l'un des deux travaux dirigés, nous pouvons décrire son écart par rapport à la moyenne du groupe comme étant la différence suivante

$$X_{ik} - \overline{X}_k$$

Alors pourquoi ne pas simplement utiliser ces écarts (c.-à-d. la mesure dans laquelle la note de chaque élève diffère de la note moyenne dans son groupe de travaux dirigés) ? N'oubliez pas qu'une variance n'est que la moyenne d'un ensemble d'écarts quadratiques, alors faisons-le. Mathématiquement, nous pourrions l'écrire comme ceci

$$\frac{\sum_{ik}\left(X_{ik} - \overline{X_k}\right)^2}{N}$$

où la notation $\sum_{ik}$ est une façon paresseuse de dire « calculer une somme en regardant tous les élèves dans tous les travaux dirigés», puisque chaque « *ik* » correspond à un élève[86]. Mais, comme nous l'avons vu au chapitre 8, calculer la variance en divisant par *N* produit une estimation biaisée de la variance de la population. Et auparavant, nous devions diviser par *N - 1* pour résoudre ce problème. Toutefois, comme je l'ai mentionné à l'époque, la raison pour laquelle ce biais existe est que l'estimation de la variance repose sur la moyenne de l'échantillon, et dans la mesure où la moyenne de l'échantillon n'est pas égale à celle de la population, elle peut systématiquement biaiser notre estimation de la variance. Mais cette fois, nous comptons sur deux moyennes d'échantillons ! Est-ce que cela signifie

[86] Une notation plus correcte sera introduite au chapitre 13.

que nous avons plus de biais ? Oui, c'est vrai. Et cela signifie-t-il que nous devons maintenant diviser par *N - 2* au lieu de *N - 1*, afin de calculer notre estimation de la variance globale ? Oui, bien sûr.

$$\hat{\sigma}_p^2 = \frac{\sum_{ik}\left(X_{ik} - \overline{X}_k\right)^2}{N-2}$$

Si vous prenez la racine carrée de ceci alors vous obtenez $\hat{\sigma}_p$, l'estimation de l'écart-type global. En somme, le calcul de l'écart-type global n'a rien de particulier. Ce n'est pas très différent du calcul normal de l'écart-type.

Terminer le test

Quelle que soit la façon dont vous voulez y penser, nous avons maintenant notre estimation globale de l'écart-type. A partir de maintenant, je vais laisser tomber le stupide indice *p*, et me référer simplement à cette estimation en tant que $\hat{\sigma}$. Super ! Revenons maintenant au test de l'hypothèse de départ, d'accord ? La raison pour laquelle nous avons calculé cette estimation est que nous savions qu'elle serait utile pour calculer notre estimation de l'erreur-type. Mais l'erreur type de *quoi* ? Dans le *test t sur un échantillon* c'était l'erreur-type de la moyenne de l'échantillon, $SE(\overline{X})$, et alors $SE(\overline{X}) = \sigma/\sqrt{N}$ c'est ce à quoi ressemblait le dénominateur de notre *statistique t.* Cette fois-ci, cependant, nous avons *deux* moyennes d'échantillon. Et ce qui nous intéresse, en particulier, c'est la différence entre les deux $\overline{X}_1 - \overline{X}_2$ Par conséquent, l'erreur-type par laquelle nous devons diviser est en fait l'**erreur-type de la différence** entre les moyennes.

Tant que les deux variables ont réellement le même écart-type, notre estimation de l'erreur type est la suivante :

$$\mathrm{SE}\left(\overline{X}_1 - \overline{X}_2\right) = \hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

et notre *statistique t* est donc

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\mathrm{SE}\left(\overline{X}_1 - \overline{X}_2\right)}$$

Comme nous l'avons vu avec notre test sur un échantillon, la distribution d'échantillonnage de cette *statistique t* est une *distribution t* (surprenant, non ?) tant que l'hypothèse nulle est vraie et que tous les présupposés du test sont remplis. Les degrés de liberté, cependant, sont légèrement différents. Comme d'habitude, on peut penser aux degrés de liberté pour être égal au nombre de points de données moins le nombre de contraintes. Dans ce cas, nous avons *N* observations (N_1 dans l'échantillon 1, et N_2 dans l'échantillon 2), et 2 contraintes (les moyennes des échantillons). Les degrés de liberté totaux pour ce test sont donc *N-2*.

Faire le test avec Jamovi

Vous ne serez pas surpris d'apprendre que vous pouvez faire un *t-test d'*échantillons indépendants facilement dans Jamovi. La variable résultat de notre test est la note de l'étudiant, et les groupes sont définis en fonction du tuteur pour chaque classe. Ainsi, vous ne serez probablement pas trop surpris que tout ce que vous avez à faire dans Jamovi est d'aller à l'analyse pertinente (« Analyses » - « T-Tests » -« Independent Samples T-Test ») et de déplacer la variable de note vers la case « Dependent Variables », et la variable tuteur vers la case « Grouping Variable », comme indiqué dans la Figure 11-10.

Le résultat a une forme très familière. Tout d'abord, il vous indique quel test a été exécuté et le nom de la variable dépendante que vous avez utilisée. Il communique ensuite les résultats des tests. Comme la dernière fois, les résultats du test sont constitués d'une *statistique t*, des degrés de liberté et de la *valeur p*. La dernière section rapporte deux choses : elle vous donne un intervalle de confiance et une taille d'effet. Je parlerai de la taille de l'effet plus tard. L'intervalle de confiance, cependant, je devrais en parler maintenant.

Il est très important d'être clair sur ce à quoi se réfère cet intervalle de confiance. C'est un intervalle de confiance pour la *différence* entre les moyennes des groupes. Dans notre exemple, les élèves d'Anastasia avaient une note moyenne de 74,53 et les élèves de Bernadette avaient une note moyenne de 69,06, de sorte que la différence entre les deux moyennes de l'échantillon est de 5,48.

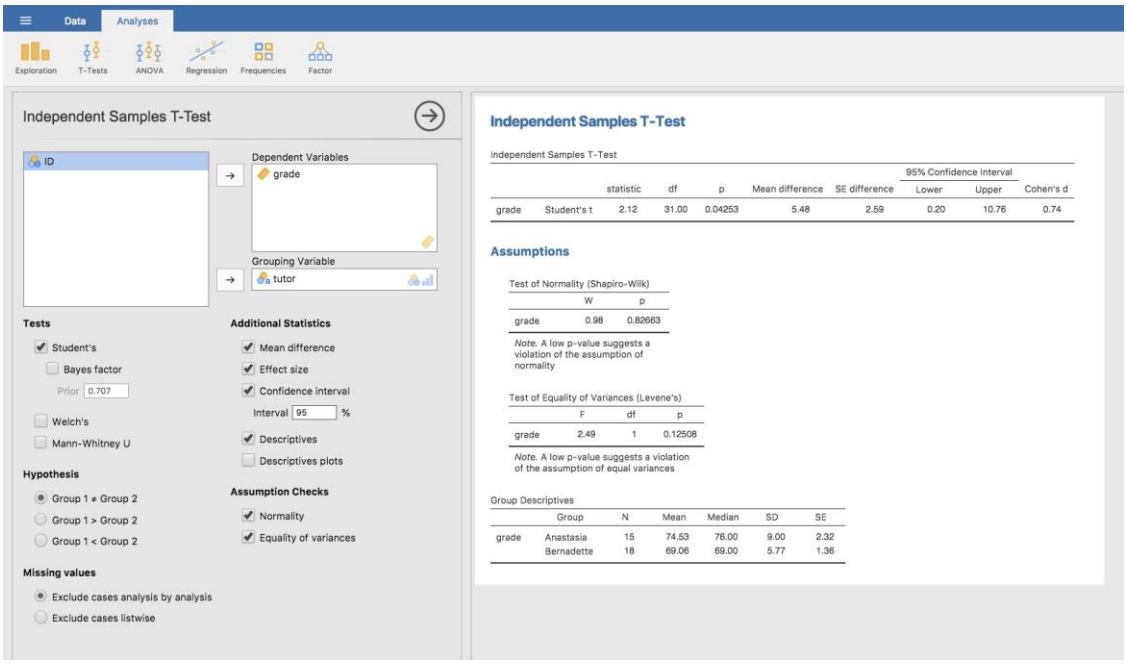


Figure 11-10 : Test t indépendant dans Jamovi, avec les options utiles aux résultats cochées.

Bien sûr, la différence entre les moyennes de population peut être plus grande ou plus petite que cela. L'intervalle de confiance indiqué à la Figure 11-10 vous indique que si nous répétons cette étude encore et encore, dans 95 % des cas, la vraie différence dans les

moyennes se situe entre 0,20 et 10,76. Reportez-vous à la section 8.5 pour un rappel sur la signification des intervalles de confiance.

Dans tous les cas, la différence entre les deux groupes est significative (à peine), nous pourrions donc écrire le résultat en utilisant un texte comme celui-ci :

La note moyenne dans la classe d'Anastasia était de 74,5 (écart-type = 9,0), tandis que la moyenne dans la classe de Bernadette était de 69,1 (écart-type = 5,8). Le *test t de Student sur des échantillons indépendants* a montré que cette différence de 5,4% était significative ($t(31) = 2{,}1, p < .05, CI_{95} = [0{,}2; 10{,}8], d = 0{,}74$), ce qui suggère qu'une différence réelle dans les résultats d'apprentissage a eu lieu.

Notez que j'ai inclus l'intervalle de confiance et la taille de l'effet dans le bloc de statistiques. Les auteurs ne font pas toujours ça. Au minimum, on s'attendrait à voir la *statistique t*, les degrés de liberté et la valeur *p.* Vous devriez donc inclure quelque chose comme ceci au minimum : $t(31) = 2{,}1, p < .05$. Si les statisticiens avaient leur mot à dire, tout le monde déclarerait aussi l'intervalle de confiance et probablement aussi la mesure de l'ampleur de l'effet, parce que ce sont des choses utiles à savoir. Mais la vraie vie ne fonctionne pas toujours de la façon dont les statisticiens le veulent, alors vous devrez faire un choix selon que vous pensez que cela aidera ou non vos lecteurs et, si vous rédigez un article scientifique, selon la norme éditoriale de la revue en question. Certains journaux s'attendent à ce que vous déclariez la taille de l'effet, d'autres non. Dans certaines communautés scientifiques, il est d'usage de signaler les intervalles de confiance, dans d'autres, ce n'est pas le cas. Vous devrez déterminer ce à quoi votre public s'attend. Mais, par souci de clarté, si vous prenez mon cours, ma position par défaut est qu'il vaut habituellement la peine d'inclure à la fois la taille de l'effet et l'intervalle de confiance.

Valeurs *t* positives et négatives

Avant de passer aux hypothèses du *test t,* j'aimerais faire une remarque supplémentaire sur l'utilisation des *tests t* dans la pratique. La première concerne le signe de la *statistique t* (c'est-à-dire s'il s'agit d'un nombre positif ou négatif). L'une des préoccupations les plus fréquentes des étudiants lorsqu'ils commencent à exécuter leur premier *test t* est qu'ils se retrouvent souvent avec des valeurs négatives pour la *statistique t* et qu'ils ne savent pas comment l'interpréter. En fait, il n'est pas rare que deux personnes travaillant indépendamment obtiennent des résultats presque identiques, sauf qu'une personne a une valeur *t* négative et l'autre une valeur *t* positive. En supposant que vous effectuez un test bilatéral, les *valeurs p* seront identiques. En y regardant de plus près, les étudiants remarqueront que les intervalles de confiance ont aussi les signes opposés. C'est tout à fait normal. Lorsque cela se produit, vous constaterez que les deux versions des résultats proviennent de deux façons légèrement différentes d'exécuter le *t-test*. Ce qui se passe ici est très simple. La *statistique t* que nous calculons ici est toujours de la forme

$$t = \frac{(\text{moyenne1} - \text{moyenne2})}{\text{SE}}$$

Si « moyenne 1 » est supérieur à « moyenne 2 », la statistique *t* sera positive, alors que si « moyenne 2 » est supérieure, la statistique *t* sera négative. De même, l'intervalle de

confiance que rapporte Jamovi est l'intervalle de confiance pour la différence « (moyenne 1) - (moyenne 2) », qui sera l'inverse de ce que vous obtiendriez si vous calculiez l'intervalle de confiance pour la différence « (moyenne 2) - (moyenne 1) ».

Bien, c'est assez simple quand on y pense, mais maintenant considérons notre *t-test* comparant la classe d'Anastasia à celle de Bernadette. Lequel devons-nous appeler « moyenne 1 » et lequel devons-nous appeler « moyenne 2 ». C'est arbitraire. Cependant, vous devez vraiment désigner l'un d'eux comme « moyenne 1 » et l'autre comme « moyenne 2 ». Il n'est pas surprenant que la façon dont Jamovi gère cela soit aussi assez arbitraire. Dans les versions antérieures du livre, j'essayais de l'expliquer, mais au bout d'un moment, j'ai abandonné, parce que ce n'est pas vraiment important et pour être honnête, je ne m'en souviens jamais moi-même. Chaque fois que j'obtiens un résultat significatif à un *test t*, et que je veux savoir quelle moyenne est la plus grande, je n'essaie pas de le déterminer en regardant la *statistique t*. Pourquoi je ferais ça ? C'est stupide. Il est plus facile de regarder simplement les moyennes réelles des groupes puisque la sortie Jamovi les montre !

Voilà le plus important. Parce que peu importe ce que Jamovi vous montre, j'essaie généralement de *rapporter* les *statistiques t de manière à* ce que les chiffres correspondent au texte. Supposons que ce que je veux écrire dans mon rapport est « La classe d'Anastasia avait des notes plus élevées que celle de Bernadette ». La formulation ici implique que le groupe d'Anastasia vient en premier, il est donc logique de rapporter la *statistique t* comme si la classe d'Anastasia correspondait au groupe 1. Si oui, j'écrirais que la classe d'Anastasia avait des notes *supérieures* à celle de Bernadette $t(31) = 2{,}1, p = .05$.

(Je ne soulignerais pas le mot « supérieur » dans la vie réelle, je le fais juste pour souligner le fait que « supérieur » correspond à des valeurs *t* positives). D'un autre côté, supposons que le commentaire que je voulais utiliser soit celui de la classe de Bernadette en premier. Si c'est le cas, il est plus logique de traiter sa classe comme un groupe 1, et si c'est le cas, l'écriture ressemble à ceci

La classe de Bernadette avait des notes *plus faibles* que celle d'Anastasia ($t(31) = -2{,}1, p < .05$).

Parce que je parle d'un groupe ayant des scores « inférieurs » cette fois-ci, il est plus raisonnable d'utiliser la forme négative de la *statistique t*. C'est juste une écriture plus propre.

Une dernière chose : veuillez noter que vous *ne pouvez pas* faire cela pour d'autres types de statistiques de test. Il fonctionne pour les *tests t*, mais il ne serait pas possible pour les tests de chi-carré, les *tests F* ou même pour la plupart des tests dont je parle dans ce livre. Ne généralisez donc pas trop ce conseil ! Je ne parle que de *tests t* et de rien d'autre !

Hypothèses du test

Comme toujours, notre test d'hypothèse repose sur certaines hypothèses. Alors qu'elles sont-elles ? Pour le test t de Student, il y a trois hypothèses, dont certaines que nous avons vues précédemment dans le contexte du *test t sur* un échantillon (voir section 11.2.3) :

- *Normalité*. Comme pour le *test t sur un échantillon*, on suppose que les données sont normalement distribuées. Plus précisément, nous supposons que les deux groupes sont normalement distribués. Dans la section 11.8, nous discuterons de la façon de tester la normalité, et dans la section 11.9, nous discuterons des solutions possibles.
- *Indépendance*. Encore une fois, on suppose que les observations sont échantillonnées de façon indépendante. Dans le contexte du test de Student, cela comporte deux aspects. Premièrement, nous supposons que les observations à l'intérieur de chaque échantillon sont indépendantes les unes des autres (exactement de la même façon que pour le test sur un échantillon). Toutefois, nous supposons également qu'il n'y a pas de dépendances inter échantillons. Si, par exemple, il s'avère que vous avez inclus certains participants dans les deux conditions expérimentales de votre étude (par exemple, en permettant accidentellement à la même personne de s'inscrire à des conditions différentes), alors il y a des dépendances croisées d'échantillon que vous devrez prendre en compte.
- *Homogénéité de la variance* (aussi appelée « homoscédasticité »). La troisième hypothèse est que l'écart-type de la population est le même dans les deux groupes. Vous pouvez tester cette hypothèse en utilisant le test de Levene, dont je parlerai plus loin dans le livre (Section 13.6.1). Cependant, il y a un remède très simple à cette hypothèse si vous êtes inquiet, dont je parlerai dans la prochaine section.

Le *test t des* échantillons indépendants (test de Welch)

Le plus gros problème que pose l'utilisation du test de Student dans la pratique est la troisième hypothèse énoncée dans la section précédente. Il suppose que les deux groupes ont le même écart-type. C'est rarement vrai dans la vraie vie.

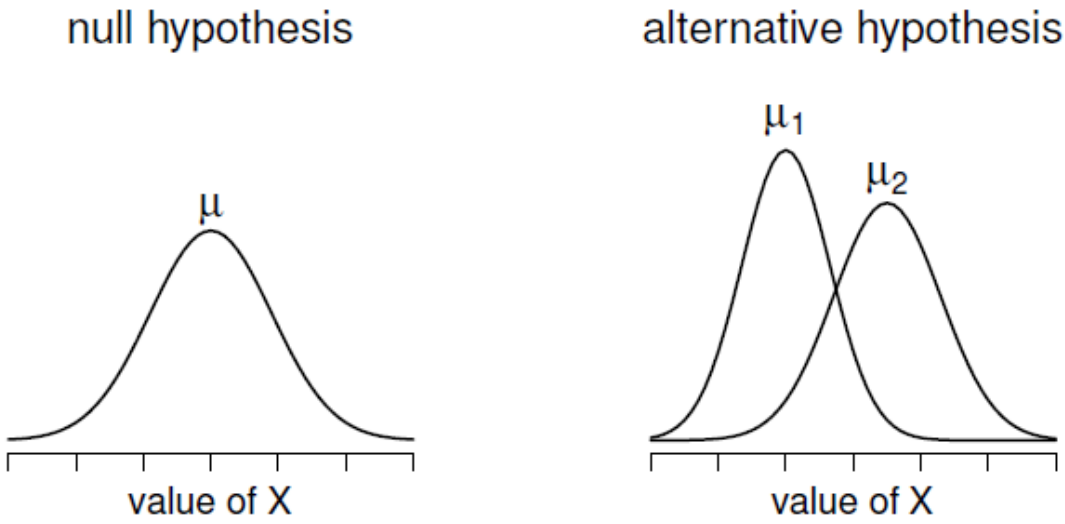


Figure 11-11 : Illustration graphique des hypothèses nulles et alternatives supposées par le test t de Welch. Comme pour le test de Student (Figure 11-9), nous supposons que les deux échantillons sont tirés d'une population normale, mais l'hypothèse alternative n'exige plus que les deux populations aient la même variance.

Si deux échantillons n'ont pas les mêmes moyennes, pourquoi devrions-nous nous attendre à ce qu'ils aient le même écart-type ? Il n'y a vraiment aucune raison de s'attendre à ce que cette hypothèse soit vraie. Nous parlerons un peu de la façon dont vous pourrez vérifier cette hypothèse plus tard parce qu'elle apparaît à différents endroits, et pas seulement au *test t*. Mais pour l'instant, je vais parler d'une autre forme de *test t* (Welch 1947) qui ne repose pas sur cette hypothèse. Une illustration graphique de ce que le **test t de Welch** suppose au sujet des données est présentée à la Figure 11-11, afin de fournir un contraste avec la version du test de Student à la Figure 11-9. J'admets qu'il est un peu étrange de parler du remède avant de parler du diagnostic, mais il se trouve que le test de Welch peut être spécifié comme l'une des options du « Independent Samples T-Test » dans Jamovi, c'est donc probablement le meilleur endroit pour en parler.

Le test de Welch est très semblable au test de Student. Par exemple, la *statistique t* que nous utilisons dans le test de Welch est calculée de la même façon que pour le test de Student. C'est-à-dire que nous prenons la différence entre les moyennes de l'échantillon et la divisons ensuite par une estimation de l'erreur type de cette différence.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\mathrm{SE}(\overline{X}_1 - \overline{X}_2)}$$

La principale différence est que les calculs de l'erreur type sont différents. Si les deux populations ont des écarts-types différents, il est complètement absurde d'essayer de calculer une estimation de l'écart type global, parce ce serait faire la moyenne de pommes et d'oranges.[Je suppose que vous pouvez faire la moyenne de pommes et d'oranges, et ce que vous obtenez est un délicieux smoothie aux fruits. Mais personne ne pense vraiment qu'un smoothie aux fruits est une très bonne façon de décrire les fruits originaux, non ?]

Mais vous pouvez toujours estimer l'erreur-type de la différence entre les moyennes de l'échantillon, elle finit par sembler différente.

$$\mathrm{SE}(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_1^2}{N_2}}$$

La raison pour laquelle il est calculé de cette façon est au-delà de la portée de ce livre. Ce qui compte pour nous, c'est que la *statistique t* qui sort du *test t de Welch* soit en fait quelque peu différente de celle qui sort du *test t de Student*.

La deuxième différence entre Welch et Student est que les degrés de liberté sont calculés d'une manière très différente. Dans le test de Welch, les « degrés de liberté » n'ont plus besoin d'être un nombre entier, et ils ne correspondent plus du tout à règle du « nombre de points de données moins le nombre de contraintes » que j'ai utilisée jusqu'à présent.

Les degrés de liberté sont en fait

$$ddl = \frac{\left(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2\right)^2}{\left(\hat{\sigma}_1^2/N_1\right)^2/(N_1-1) + \left(\hat{\sigma}_1^2/N_1\right)^2/(N_2-1)}$$

Ce qui est assez simple et évident, non ? Peut-être pas ! Ça n'a pas vraiment d'importance pour nous. Ce qui compte, c'est que vous verrez que la valeur « df » qui résulte d'un test de Welch a tendance à être un peu plus petite que celle utilisée pour le test de Student, et il n'est pas nécessaire que ce soit un nombre entier

Faire le test de Welch avec Jamovi

Si vous cochez la case pour le test de Welch dans l'analyse que nous avons faite ci-dessus, alors voici ce qu'il vous donne (Figure 11-12) :

Independent Samples T-Test

Independent Samples T-Test

							95% Confidence Interval		
		statistic	df	p	Mean difference	SE difference	Lower	Upper	Cohen's d
grade	Student's t	2.12	31.00	0.04253	5.48	2.59	0.20	10.76	0.74
	Welch's t	2.03	23.02	0.05361	5.48	2.69	−0.09	11.05	0.74

Figure 11-12 : Résultats montrant le test Welch parallèlement au test t de Student par défaut dans Jamovi

L'interprétation de ce résultat devrait être assez évidente. Vous lirez le résultat du test de Welch de la même manière que vous le feriez pour le test de Student. Vous avez vos statistiques descriptives, les résultats des tests et d'autres informations. C'est donc assez facile.

Sauf que, sauf que… notre résultat n'est plus significatif. Lorsque nous avons effectué le test de Student, nous avons obtenu un effet significatif, mais le test de Welch sur le même ensemble de données ne l'est pas ($t(23.02) = 2{,}03, p = .054$). Qu'est-ce que cela signifie ? Devrions-nous paniquer ? Le ciel brûle-t-il ? Probablement pas. Le fait qu'un test soit significatif et l'autre ne le soit pas ne signifie pas grand-chose en soi, d'autant plus que j'ai en quelque sorte truqué les données pour que cela arrive. En règle générale, ce n'est pas une bonne idée d'essayer d'interpréter ou d'expliquer la différence entre une *valeur p* de .049 et une *valeur p* de.051. Si ce genre de chose se produit dans la vie réelle, la *différence* dans ces *valeurs p* est presque certainement due au hasard. Ce qui importe, c'est que vous preniez un peu de soin à réfléchir au test que vous utilisez. Le test de Student et le test de Welch ont des forces et des faiblesses différentes. Si les deux populations ont réellement des variances égales, alors le test de Student est légèrement plus puissant (taux d'erreur de type II inférieur) que le test de Welch. Cependant, s'ils n'ont pas les mêmes variances, alors les hypothèses du test de Student sont violées et vous pourriez ne pas être en mesure de lui

faire confiance ; vous pourriez vous retrouver avec un taux d'erreur de Type I plus élevé. C'est donc un compromis. Cependant, dans la vie réelle, j'ai tendance à préférer le test de Welch, parce que presque personne ne croit *vraiment* que les écarts-types de population sont identiques.

Hypothèses du test

Les hypothèses du test de Welch sont très semblables à celles du *test t de Student* (voir la section 11.3.7), sauf que le test de Welch ne suppose pas d'homogénéité de la variance. Cela ne laisse que l'hypothèse de normalité et l'hypothèse d'indépendance. Les particularités de ces hypothèses sont les mêmes pour le test de Welch que pour le test de Student.

Le *test t-test des* échantillons appariés

Qu'il s'agisse du test de Student ou du test de Welch, un *test t* pour groupes indépendants est destiné à être utilisé dans une situation où vous avez deux échantillons qui sont indépendants l'un de l'autre. Cette situation se produit naturellement lorsque les participants sont assignés au hasard à l'une des deux conditions expérimentales, mais elle fournit une très mauvaise approximation par rapport à d'autres types de modèles de recherche. En particulier, l'analyse à l'aide de *tests t pour* échantillons indépendants ne convient pas un plan à mesures répétées, dans lequel chaque participant est mesuré (par rapport à la même variable de résultat) dans les deux conditions expérimentales. Par exemple, nous pourrions nous demander si l'écoute de musique réduit la capacité de mémoire de travail des gens. Pour ce faire, nous avons pu mesurer la capacité de mémoire de travail de chaque personne dans deux conditions : avec et sans musique. Dans un plan expérimental comme celui-ci,[Cette conception est très semblable à celle de la section 11.7 qui a motivé le test McNemar. Cela ne devrait pas être une surprise. Dans les deux cas, il s'agit de mesures répétées standard comportant deux mesures. La seule différence est que cette fois-ci, notre variable de résultat est une échelle d'intervalle (capacité de la mémoire de travail) plutôt qu'une variable d'échelle nominale binaire (une question oui ou non).] chaque participant apparaît dans les *deux* groupes. Cela nous oblige à aborder le problème d'une manière différente, en utilisant le **test t pour des échantillons appariés**.

Les données

Les données que nous utiliserons cette fois-ci proviennent de la classe du Dr Chico.[87] Dans sa classe, les élèves passent deux tests majeurs, l'un au début du semestre et l'autre plus tard dans le semestre. A l'entendre, elle dirige un cours très dure, un cours que la plupart des élèves trouvent très difficile. Mais elle soutient qu'en établissant des évaluations rigoureuses, on encourage les élèves à travailler plus fort. Sa théorie est que le premier test est une sorte de « réveil » pour les élèves. Quand ils réaliseront à quel point son cours est vraiment difficile, ils travailleront plus dur pour le deuxième test et obtiendront une

---

[87] Nous accueillons maintenant les Drs Harpo, Chico et Zeppo. Pas de récompense pour qui devine qui est le Dr Groucho.

meilleure note. A-t-elle raison ? Pour tester cela, importons le fichier chico.csv dans Jamovi. Cette fois, Jamovi fait un bon travail lors de l'importation de l'attribution correcte des échelles de mesure. L'ensemble de données chico contient trois variables : une variable id qui identifie chaque élève de la classe, la variable grade_test1 qui enregistre la note de l'élève pour le premier test, et la variable grade_test2 qui a les notes pour le second test.

Si nous regardons la feuille de calcul Jamovi, il semble que la classe est difficile (la plupart des notes se situent entre 50% et 60%), mais il semble y avoir une amélioration entre le premier et le second test.

Si nous examinons rapidement les statistiques descriptives de la Figure 11-13, nous constatons que cette impression semble se confirmer. Pour l'ensemble des 20 élèves, la note moyenne pour le premier test est de 57, mais elle passe à 58 pour le deuxième test. Bien que les écarts-types soient respectivement de 6,6 et 6,4, il commence à sembler que l'amélioration n'est peut-être qu'illusoire ; peut-être une variation aléatoire. Cette impression est renforcée lorsque vous voyez les moyennes et les intervalles de confiance représentés à la Figure 11-14a. Si nous devions nous fier uniquement à ce graphique et examiner l'ampleur de ces intervalles de confiance, nous serions tentés de croire que l'amélioration apparente du rendement des élèves est le fruit du hasard.

Néanmoins, cette impression est fausse. Pour comprendre pourquoi, jetez un coup d'œil au nuage de points des notes de l'épreuve 1 par rapport à celles de l'épreuve 2, comme le montre la Figure 11-14b. Dans ce graphique, chaque point correspond aux deux notes d'un élève donné. Si leur note au test 1 (coordonnée x) est égale à leur note au test 2 (coordonnée y), le point tombe sur la ligne. Les points qui tombent au-dessus de la ligne sont les élèves qui ont obtenu de meilleurs résultats au deuxième test. De façon critique, presque tous les points de données se situent au-dessus de la ligne diagonale : presque tous les élèves semblent avoir amélioré leur note, ne serait-ce que d'une petite quantité. Cela suggère que nous devrions examiner les *améliorations* apportées par chaque élève d'un test à l'autre et considérer cela comme nos données brutes. Pour ce faire, nous devrons créer une nouvelle variable pour l'amélioration apportée par chaque élève, et l'ajouter à l'ensemble de données chico. La façon la plus simple de le faire est de calculer une nouvelle variable, avec l'expression grade test2 – grade test1

Une fois que nous avons calculé cette nouvelle variable d'amélioration, nous pouvons dessiner un histogramme montrant la distribution de ces scores d'amélioration, comme le montre la Figure 11-14c. Quand on regarde l'histogramme, il est très clair qu'il y a une réelle amélioration. La grande majorité des élèves ont obtenu de meilleurs résultats au test 2 qu'au test 1, ce qui se reflète dans le fait que presque tout l'histogramme est au-dessus de zéro.

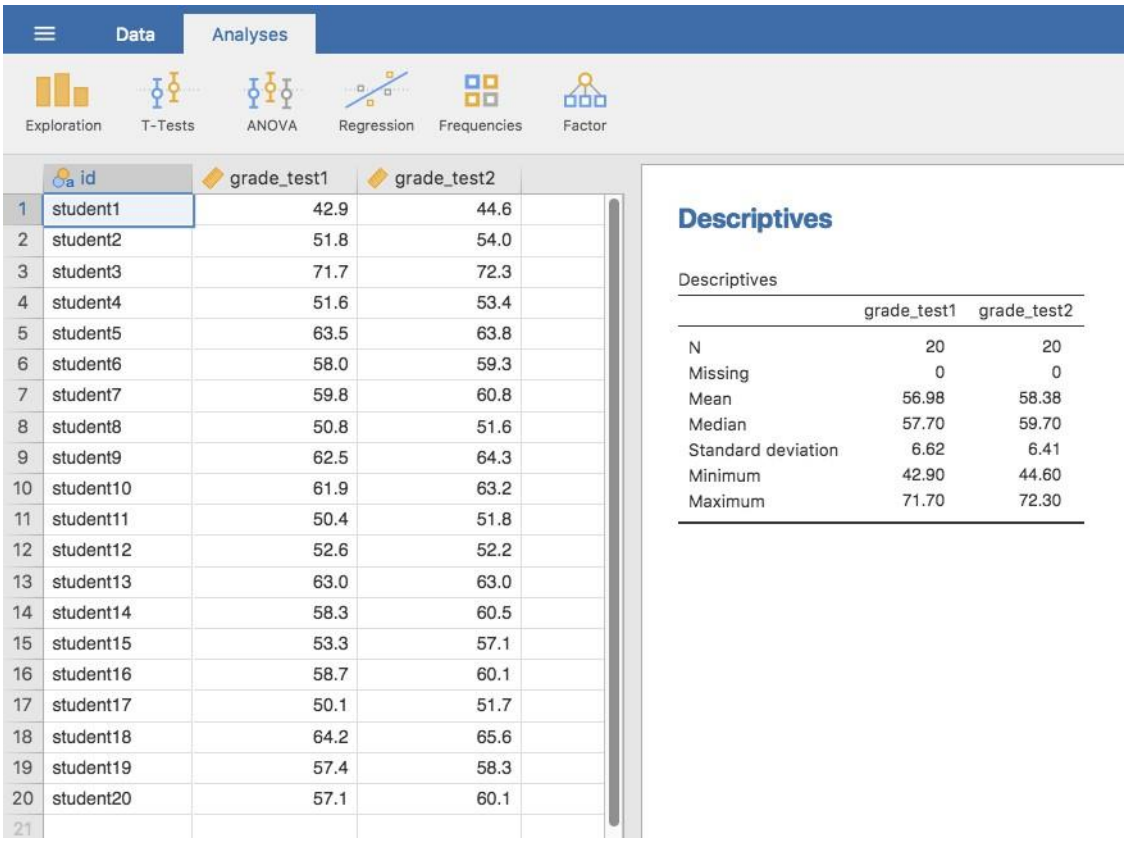


Figure 11-13 : Analyse descriptives des deux variables d'essai de niveau dans l'ensemble de données de chico

Qu'est-ce que le *test t-test des* échantillons appariés ?

A la lumière de l'exploration précédente, réfléchissons à la manière de construire un test *t* approprié. Une possibilité serait d'essayer d'exécuter un *test t d'*échantillons indépendants en utilisant les variables grade_test1 et grade_test2 comme variables d'intérêt. Cependant, c'est clairement la mauvaise chose à faire car le *test t des échantillons indépendants* suppose qu'il n'y a pas de relation particulière entre les deux échantillons. Pourtant, il est clair que ce n'est pas vrai dans ce cas-ci en raison de la structure des mesures répétées dans les données. Pour reprendre le langage que j'ai présenté dans la dernière section, si nous essayions de faire un *test t d'échantillons indépendants*, nous confondrions les différences à **l'intérieur du sujet** (ce que nous voulons tester) avec la variabilité **entre sujets** (ce que nous ne faisons pas).

La solution au problème est évidente, je l'espère, puisque nous avons déjà fait tout le travail difficile dans la section précédente. Au lieu d'exécuter un *test t d'échantillons indépendants* sur grade_test1 et grade_test2, nous exécutons un *test t sur un échantillon* sur la variable de différence intra-sujet, l'amélioration. Pour formaliser un peu, si X_{i1} est le score que le *i-ème* participant a obtenu sur la première variable, et X_{i2} est le score que la même personne a obtenu sur la seconde, alors le score de différence est :

$$D_i = X_{i1} - X_{i2}$$

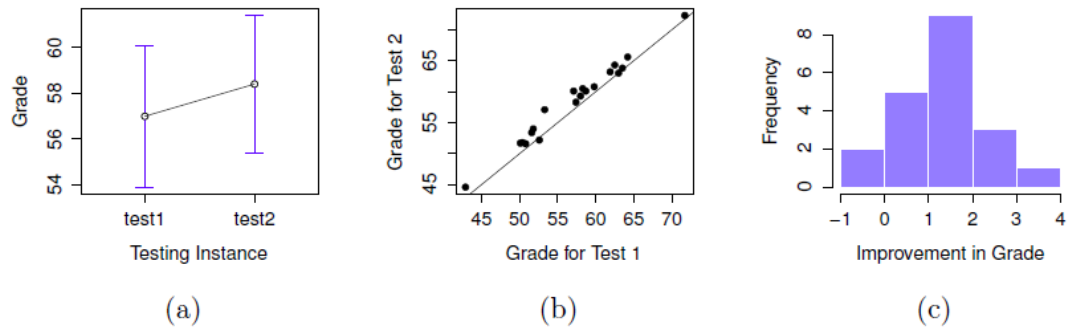


Figure 11-14 : Note moyenne pour l'essai 1 et l'essai 2, avec des intervalles de confiance associés à 95 % (panel a). Diagramme de dispersion montrant les notes individuelles pour l'essai 1 et l'essai 2 (panneau b). Histogramme montrant l'amélioration réalisée par chaque élève de la classe du Dr Chico (panel c). Dans le panneau c, remarquez que la distribution est presque entièrement au-dessus de zéro : la grande majorité des élèves ont amélioré leur performance du premier au deuxième test.

Notez que les scores de différence sont la *variable 1 moins la variable 2* et non l'inverse, donc si nous voulons que l'amélioration corresponde à une différence de valeur positive, nous voulons en fait que « test 2 » soit notre « variable 1 ». De même, nous dirions que $\mu_D = \mu_1 - \mu_2$ est la moyenne de population pour cette variable de différence. Donc, pour convertir ceci en un test d'hypothèse, notre hypothèse nulle est que cette différence moyenne est nulle et l'hypothèse alternative est que ce n'est pas le cas.

$$\text{H}_{0} : \mu_{D}=0\\ \text{H}_{1} : \mu_{D}\neq 0$$

En supposant qu'il s'agit d'un test à deux volets. Ceci est plus ou moins identique à la façon dont nous avons décrit les hypothèses du *test t pour un échantillon*. La seule différence est que la valeur spécifique prédite par l'hypothèse nulle est 0, de sorte que notre *statistique t* est définie à peu près de la même manière. Si nous laissons $\overline{D}$ indiquer la moyenne des scores de différence, alors :

$$t = \frac{\overline{D}}{SE(\overline{D})}$$

Lequel correspond à

$$t = \frac{\overline{D}}{\hat{\sigma}_D / \sqrt{N}}$$

Où $\hat{\sigma}_D$ est l'écart-type des scores de différence. Puisqu'il ne s'agit que d'un *test t* ordinaire, sur un échantillon, sans rien de particulier, les degrés de liberté sont toujours *N - 1*. Et c'est tout. Le test t pour échantillons appariés n'est vraiment pas un nouveau test du tout. Il s'agit d'un test t à un échantillon, mais appliqué à la différence entre deux variables. C'est en fait très simple. La seule raison pour laquelle il mérite une discussion aussi longue que celle que

nous venons de traverser est que vous devez être capable de reconnaître *quand* un test d'échantillons pour échantillons appariés est approprié, et de comprendre *pourquoi* c'est mieux qu'un test t pour échantillons indépendants.

Faire le test avec Jamovi

Comment faire un *t-test d'échantillons appariés* avec Jamovi ? Une possibilité est de suivre le processus que j'ai décrit ci-dessus. C'est-à-dire, créer une variable de « différence » et ensuite exécuter un test t pour un échantillon sur cette variable. Puisque nous avons déjà créé une variable appelée amélioration, faisons cela et voyons ce que nous obtenons, Figure 11-15.

One Sample T-Test

One Sample T-Test

						95% Confidence Interval		
		statistic	df	p	Mean difference	Lower	Upper	Cohen's d
improvement	Student's t	6.48	19.00	<.00001	1.40	0.95	1.86	1.45

Figure 11-15 : Résultats d'un test t d'un échantillon sur des scores de différence appariés

Le résultat illustré à la Figure 11-15 est (évidemment) formaté exactement de la même façon que la dernière fois que nous avons utilisé l'analyse « One Sample T-Test » (section 11.2), et cela confirme notre intuition. Il y a une amélioration moyenne de 1,4 % entre le test 1 et le test 2, ce qui est très différent de 0 ($t(19) = 6{,}48, p < .001$).

Cependant, supposons que vous soyez paresseux et que vous ne vouliez pas faire tout l'effort de créer une nouvelle variable. Ou peut-être voulez-vous simplement que la différence entre les tests à un échantillon et les tests à deux échantillons reste claire dans votre tête. Si c'est le cas, vous pouvez utiliser l'analyse de Jamovi « Paired Samples T-Test » pour obtenir les résultats présentés à la Figure 11-16.

Paired Samples T-Test

Paired Samples T-Test

							95% Confidence Interval		
			statistic	df	p	Mean difference	SE difference	Lower	Upper	Cohen's d
grade_test2	grade_test1	Student's t	6.48	19.00	<.00001	1.40	0.22	0.95	1.86	1.45

Figure 11-16 : Résultats d'un test t d'échantillon apparié. Comparer avec la Figure 11-15

Les chiffres sont identiques à ceux qui proviennent de l'analyse sur un seul échantillon, puisque, nous le soulignons une nouvelle fois, l'*analyse t-test pour échantillons appariés* n'est au fond qu'une analyse sur un seul échantillon.

Tests unilatéraux

Lorsque j'ai présenté la théorie des tests d'hypothèse nulle, j'ai mentionné qu'il y a certaines situations où il est approprié de spécifier un test *unilatéral* (voir la section 9.4.3). Jusqu'à présent, tous les *tests* ont été effectués de façon bliatérale. Par exemple, lorsque nous avons spécifié un *test t sur* un échantillon pour les notes de la classe de M. Zeppo, l'hypothèse nulle était que la vraie moyenne était de 67,5. L'autre hypothèse était que la moyenne réelle était supérieure *ou* inférieure à 67,5. Supposons que nous ne nous intéressions qu'à savoir si la vraie moyenne est supérieure à 67,5, et que nous n'ayons aucun intérêt à tester pour savoir si la vraie moyenne est inférieure à 67,5. Si tel est le cas, notre hypothèse nulle serait que la vraie moyenne est de 67,5 ou moins, et l'hypothèse alternative serait que la vraie moyenne est supérieure à 67,5. Dans Jamovi, pour l'analyse « One Sample T-Test », vous pouvez le spécifier en cliquant sur l'option « > Test Value », sous « Hypothesis ». Une fois que vous avez fait cela, vous obtiendrez les résultats comme indiqué en Figure 11-17.

One Sample T-Test

One Sample T-Test

					95% Confidence Interval		
		statistic	df	p	Lower	Upper	Cohen's d
x	Student's t	2.25	19.00	0.01807	68.62	Inf	0.50

Note. H_a population mean > 67.5

Figure 11-17 : résultats de Jamovi montrant un « test t sur un échantillon »où l'hypothèse réelle est unilatérale, c'est-à-dire que la vraie moyenne est supérieure à 67,5.

Notez qu'il y a quelques changements par rapport au résultat que nous avons vu la dernière fois. Le plus important est le fait que l'hypothèse réelle a changé, pour refléter les différents tests. La deuxième chose à noter est que bien que la *statistique t* et les degrés de liberté n'aient pas changé, la *valeur p* a changé. Cela s'explique par le fait que le test unilatéral a une zone de rejet différente de celle du test bilatéral. Si vous avez oublié pourquoi et ce que cela signifie, vous pouvez relire le chapitre 9, et la section 9.4.3 en particulier. La troisième chose à noter est que l'intervalle de confiance est lui aussi différent : il fait désormais état d'un intervalle de confiance « unilatéral » plutôt que bilatéral. Dans un intervalle de confiance bilatéral, nous essayons de trouver les nombres *a* et *b* de telle sorte que nous sommes convaincus que, si nous devions répéter l'étude plusieurs fois, alors 95% du temps la moyenne se situerait *entre a et b*. Dans un intervalle de confiance unilatéral, nous essayons de trouver un seul nombre *a* tel que nous sommes convaincus que 95% du temps la vraie moyenne serait *supérieure à a* (ou inférieure à *a* si vous sélectionnez Measure 1 < Measure 2 dans la section « Hypothesis »).

C'est donc comme ça qu'on fait un *t-test* unilatéral d'un échantillon. Cependant, toutes les versions du *test t* peuvent être unilatérales. Pour un *test t d'échantillons indépendants*, vous

pourriez avoir un test unilatéral si vous êtes seulement intéressé à vérifier si le groupe A a des scores *plus élevés* que le groupe B, mais n'avez aucun intérêt à savoir si le groupe B a des scores plus élevés que le groupe A. Supposons que, pour la classe du Dr Harpo, vous vouliez voir si les élèves d'Anastasia ont des notes supérieures à celle de Bernadette. Pour cette analyse, dans les options « Hypothèse », précisez que « Groupe 1 < Groupe2. Vous devriez obtenir les résultats illustrés à la Figure 11-18.

Independent Samples T-Test

Independent Samples T-Test

							95% Confidence Interval		
		statistic	df	p	Mean difference	SE difference	Lower	Upper	Cohen's d
grade	Student's t	2.12	31.00	0.02126	5.48	2.59	1.09	Inf	0.74

Note. H_a Anastasia > Bernadette

Figure 11-18 : résultats de Jamovi montrant un 'Test t d'échantillons indépendants » où l'hypothèse réelle est unilatérale, c'est-à-dire que les étudiants d'Anastasia avaient des notes supérieures à ceux de Bernadette.

Encore une fois, le résultat change de façon prévisible. La définition de l'hypothèse alternative a changé, la *valeur p* a changé, et elle fait maintenant état d'un intervalle de confiance unilatéral plutôt que bilatéral.

Qu'en est-il du *test t des échantillons appariés* ? Supposons que nous voulions tester l'hypothèse que les notes *augmentent* du test 1 au test 2 dans la classe de M. Zeppo, et que nous ne sommes pas prêts à considérer l'idée que les notes diminuent. Dans Jamovi vous le feriez en spécifiant, sous l'option « Hypothesis », que grade test2 (« Measure 1 » dans Jamovi, parce que nous l'avons d'abord copié dans la boîte des variables appariées) > grade test1 (« Measure 2 » dans Jamovi). Vous devriez obtenir les résultats illustrés à la Figure 11-19.

Paired Samples T-Test

Paired Samples T-Test

								95% Confidence Interval		
			statistic	df	p	Mean difference	SE difference	Lower	Upper	Cohen's d
grade_test2	grade_test1	Student's t	6.48	19.00	<.00001	1.40	0.22	1.03	Inf	1.45

Note. H_a Measure 1 > Measure 2

Figure 11-19 : résultats de Jamovi montrant un « Test t d'échantillons appariés » où l'hypothèse réelle est unilatérale, c'est-à-dire que le grade test2 (Mesure 1) > grade test1 (Measure 2)

Encore une fois, la production change de façon prévisible. L'hypothèse a changé, la *valeur p* a changé et l'intervalle de confiance est maintenant unilatéral.

Taille de l'effet

La mesure de l'ampleur de l'effet la plus couramment utilisée pour un *test t* est le *d de Cohen** (Cohen 1988). C'est une mesure très simple en principe, avec pas mal de rides quand on commence à creuser dans les détails. Cohen lui-même l'a défini principalement dans le contexte d'un *test t d'échantillons indépendants*, en particulier le test de Student. Dans ce contexte, une façon naturelle de définir la valeur de l'effet est de diviser la différence entre les moyennes par une estimation de l'écart-type. En d'autres termes, nous cherchons à calculer quelque chose dans ce sens :

$$d = \frac{(\text{moyenne 1}) - (\text{moyenne 2})}{\text{std dev}}$$

et il a suggéré un guide approximatif pour l'interprétation *d* dans le Tableau 11-1. On pourrait penser que ce serait assez clair, mais ce n'est pas le cas. C'est en grande partie parce que Cohen n'était pas assez précis sur ce qui, selon lui, devrait être utilisé comme mesure de l'écart-type (pour sa défense, il essayait d'adopter un point de vue plus large dans son livre, et non des détails mineurs). Comme l'ont mentionné McGrath et Meyer (2006), il existe plusieurs versions différentes d'usage courant, et chaque auteur a tendance à adopter une notation légèrement différente. Par souci de simplicité (par opposition à l'exactitude), j'utiliserai *d* pour faire référence à toute statistique calculée à partir de l'échantillon, et j'utiliserai δ pour faire référence à un effet dans la population théorique. Évidemment, cela signifie qu'il y a plusieurs choses différentes, toutes appelées *d*.

Tableau 11-1 : Un guide (très) grossier pour interpréter le d. de Cohen. Ma recommandation personnelle est de ne pas les utiliser à l'aveuglette. La statistique d a une interprétation naturelle en soi. Il décrit la différence entre les moyennes comme étant le nombre d'écarts-types qui sépare ces moyennes. C'est donc généralement une bonne idée de réfléchir à ce que cela signifie en termes pratiques. Dans certains contextes, un « petit » effet pourrait avoir une grande importance pratique. Dans d'autres situations, un effet « important » peut ne pas être très intéressant.

valeur d	interprétation approximative
environ 0,2	«Effet « petit
environ 0,5	«effet « modéré
environ 0,8	«effet « grand

Je pense que le seul cas où vous voudriez utiliser le d de Cohen, c'est lorsque vous exécutez un *test t*, et Jamovi a la possibilité de calculer la taille de l'effet pour toutes les variantes de *test t* qu'il propose.

d de Cohen sur un échantillon

La situation la plus simple à considérer est celle qui correspond à un *test t sur* un échantillon. Dans ce cas, il s'agit de la moyenne $\overline{X}$ *d'*un échantillon et de la moyenne μ_0 d'une population (hypothétique) à laquelle la comparer. De plus, il n'y a qu'une seule façon raisonnable d'estimer l'écart-type de la population. Nous utilisons simplement notre estimateur habituel $\hat{\sigma}$. Par conséquent, nous nous retrouvons avec ce qui suit comme seule façon de calculer *d*

$$d = \frac{\overline{X} - \sigma_0}{\hat{\sigma}}$$

Lorsqu'on examine les résultats de la Figure 11-6, la valeur de taille de l'effet est la valeur du d de Cohen = 0,50. Dans l'ensemble, les étudiants en psychologie de la classe du Dr Zeppo obtiennent donc des notes (moyenne = 72,3) qui sont d'environ 0,5 écarts-types plus élevés que le niveau auquel vous vous attendriez (67,5) s'ils se situaient au même niveau que les autres étudiants. Si l'on en juge par le guide approximatif de Cohen, il s'agit d'un effet de taille modérée.

d de Cohen à partir d'un test t de Student

La majorité des discussions sur le *d de Cohen* se concentrent sur une situation analogue au test *t de Student des échantillons indépendants* , et c'est dans ce contexte que l'histoire devient plus confuse, puisqu'il existe plusieurs versions différentes de *d* que vous pourriez utiliser dans cette situation. Pour comprendre pourquoi il existe plusieurs versions de *d*, il est utile de prendre le temps d'écrire une formule qui correspond à la taille réelle de l'effet de population δ. C'est assez simple,

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

où, comme d'habitude, μ_1 et μ_2 sont les moyennes de population correspondant respectivement au groupe 1 et au groupe 2, et σ est l'écart-type (identique pour les deux populations). La façon évidente d'estimer δ est de faire exactement la même chose que dans le *test t* lui-même, c'est-à-dire d'utiliser les moyennes de l'échantillon dans numérateur et une estimation de l'écart type regroupé dénominateur.

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}_P}$$

Où $\hat{\sigma}_P$ est exactement la même mesure de l'écart-type global qui apparaît dans le *test t*. C'est la version la plus couramment utilisée du d de Cohen lorsqu'il est appliquée aux résultats d'un *test t de Student*, et c'est celle fournie dans Jamovi. On l'appelle parfois la statistique *g de Hedges* (Hedges 1981).

Cependant, il y a d'autres possibilités que je vais décrire brièvement. Premièrement, vous pouvez avoir des raisons de ne vouloir utiliser qu'un seul des deux groupes comme base de calcul de l'écart-type. Cette approche (souvent appelée Δ de Glass, prononcé *delta*) n'a de

sens que lorsque l'on a de bonnes raisons de traiter l'un des deux groupes comme un reflet plus pur de « variation naturelle » que l'autre. Cela peut se produire si, par exemple, l'un des deux groupes est un groupe témoin. Deuxièmement, rappelons que dans le calcul habituel de l'écart-type global, nous divisons par *N-2* pour corriger le biais dans la variance de l'échantillon. Dans une version du d de Cohen, cette correction est omise et nous la divisons par *N*. Cette version a du sens surtout lorsqu'on essaie de calculer la taille de l'effet dans l'échantillon plutôt que d'estimer la taille de l'effet dans la population. Enfin, il existe une version basée sur Hedges et Olkin (2014), qui soulignent l'existence d'un léger biais dans l'estimation habituelle (globale) du d de Cohen. Ils introduisent donc une petite correction en multipliant la valeur habituelle de *d* par *(N–3)/(N-2,25)*.

Laissons de côté toutes ces variations que vous pourriez utiliser si vous le vouliez et jetons un coup d'œil à la version par défaut dans Jamovi. Dans la Figure 11-10 d de Cohen = 0,74, ce qui indique que les notes des élèves de la classe d'Anastasia sont, en moyenne, de 0,74 écart-type supérieurs aux notes des élèves de la classe de Bernadette. Pour un test de Welch, la taille de l'effet estimé est la même (Figure 11-12).

d de Cohen pour un test t sur des échantillons appariés

Enfin, que devrions-nous faire pour un *test t sur des* échantillons appariés ? Dans ce cas, la réponse dépend de ce que vous essayez de faire. Jamovi suppose que vous voulez mesurer l'ampleur de votre effet par rapport à la distribution des scores de différence, et la mesure de *d* que vous calculez est :

$$d = \frac{\overline{D}}{\hat{\sigma}_D}$$

Où $\hat{\sigma}_D$ est l'estimation de l'écart-type des différences. Dans la Figure 11-16, d de Cohen = 1,45, ce qui indique que les notes au temps 2 sont, en moyenne, de 1,45 écart-type plus élevées que les notes au temps 1.

C'est la version du d de Cohen qui est utilisé par l'analyse Jamovi « Paired Samples T-Test ». Le seul problème, c'est de savoir si c'est la mesure que vous voulez ou non. Dans la mesure où vous vous souciez des conséquences pratiques de votre recherche, vous voulez souvent mesurer la taille de l'effet par rapport aux variables *initiales*, et non les scores de *différence* (par exemple, l'amélioration de 1% dans la classe du Dr Chico au fil du temps est assez faible par rapport à la variation des notes entre étudiants), auquel cas vous utiliserez les mêmes versions du d de Cohen's que celles vous utiliseriez pour un test Student ou Welch. Ce n'est pas si simple de le faire en Jamovi ; il faut en fait modifier la structure des données dans la feuille de données alors je vais pas vous en parler ici.[88], mais le d de Cohen pour cette perspective est très différent : il est de 0,22 ce qui est assez petit quand on l'évalue sur l'échelle des variables originales.

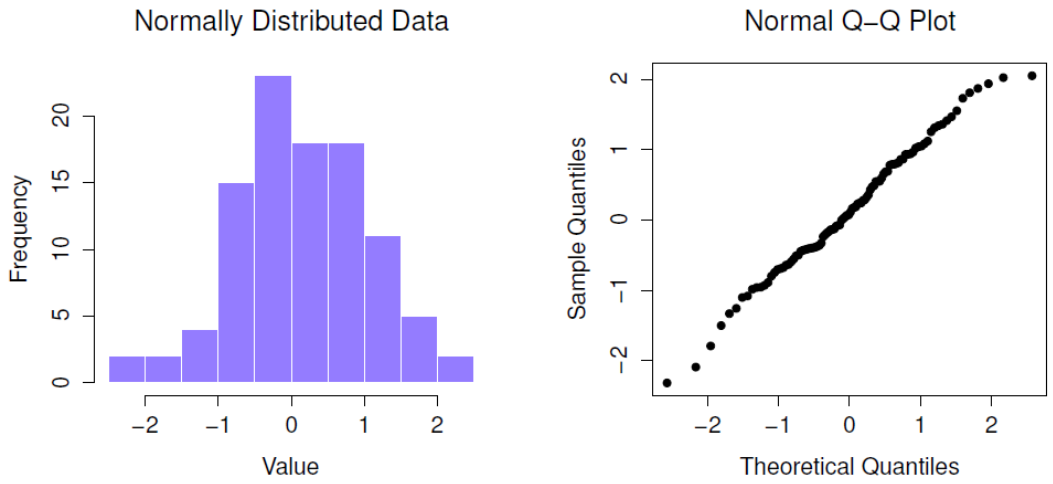
---

[88] Si vous êtes intéressé, vous pouvez voir comment cela a été fait dans le fichier chico2.omv

Vérification de la normalité d'un échantillon

Tous les tests dont nous avons discuté jusqu'ici dans ce chapitre ont supposé que les données sont normalement distribuées. Cette hypothèse est souvent tout à fait raisonnable, car le théorème central limite (section 8.3.3) tend à garantir que de nombreuses quantités du monde réel sont normalement réparties. Chaque fois que vous soupçonnez que votre variable est en *fait* une moyenne de beaucoup de choses différentes, il y a de fortes chances qu'elle soit normalement distribuée, ou au moins assez proche de la normale pour que vous puissiez vous en tirer avec *tests t*. Cependant, la vie n'offre pas de garanties, et il y a d'ailleurs de nombreuses situations où on se retrouve avec des variables qui sont complètement anormales. Par exemple, chaque fois que vous pensez que votre variable est en fait le minimum de beaucoup de choses différentes, il y a de très bonnes chances qu'elle finisse de façon assez asymétrique. En psychologie, les données sur le temps de réponse (TR) en sont un bon exemple. Si vous supposez qu'il y a beaucoup de choses qui pourraient déclencher une réponse d'un sujet humain, alors la réponse réelle se produira la première fois qu'un de ces événements déclencheurs se produit.[89] Cela signifie que les données des TR sont systématiquement non normales. Bien, donc si la normalité est assumée par tous les tests, et qu'elle est satisfaite la plupart du temps mais pas toujours (au moins approximativement) par les données du monde réel, comment pouvons-nous vérifier la normalité d'un échantillon ? Dans cette section, j'aborde deux méthodes : les diagrammes QQ et le test Shapiro-Wilk.

Les diagrammes QQ



[89] Il s'agit d'une sursimplification massive.

Figure 11-20 : Histogramme (figure a) et graphique QQ normal (figure b) de données normales, un échantillon normalement distribué avec 100 observations. La statistique Shapiro-Wilk associée à ces données est W = .99, ce qui indique qu'aucun écart significatif par rapport à la normalité n'a été détecté (p = .73)

Une façon de vérifier si un échantillon viole l'hypothèse de normalité est de dessiner un **« graphique QQ »** (graphique Quantile-Quantile). Cela vous permet de vérifier visuellement si vous constatez des violations systématiques. Dans un graphique QQ, chaque observation est tracée comme un point unique. La coordonnée x est le quantile théorique dans lequel l'observation devrait se situer si les données étaient normalement distribuées (avec la moyenne et la variance estimées à partir de l'échantillon), et sur la coordonnée y est le quantile réel des données dans l'échantillon. Si les données sont normales, les points doivent former une ligne droite. Par exemple, voyons ce qui se passe si nous générons des données en échantillonnant à partir d'une distribution normale, puis en dessinant un graphique QQ. Les résultats sont présentés à la Figure 11-20. Comme vous pouvez le constater, ces données forment une ligne assez droite ; ce qui n'est pas surprenant étant donné que nous les avons échantillonnées à partir d'une distribution normale ! Par contre, examinez les deux ensembles de données illustrés à la Figure 11-21. Les panneaux supérieurs montrent l'histogramme et un graphique QQ pour un ensemble de données très asymétrique : le tracé QQ se courbe vers le haut. Les figures inférieures montrent les mêmes tracés pour un ensemble de données à forte kurtose : dans ce cas, le tracé QQ s'aplatit au milieu et s'incurve fortement à chaque extrémité.

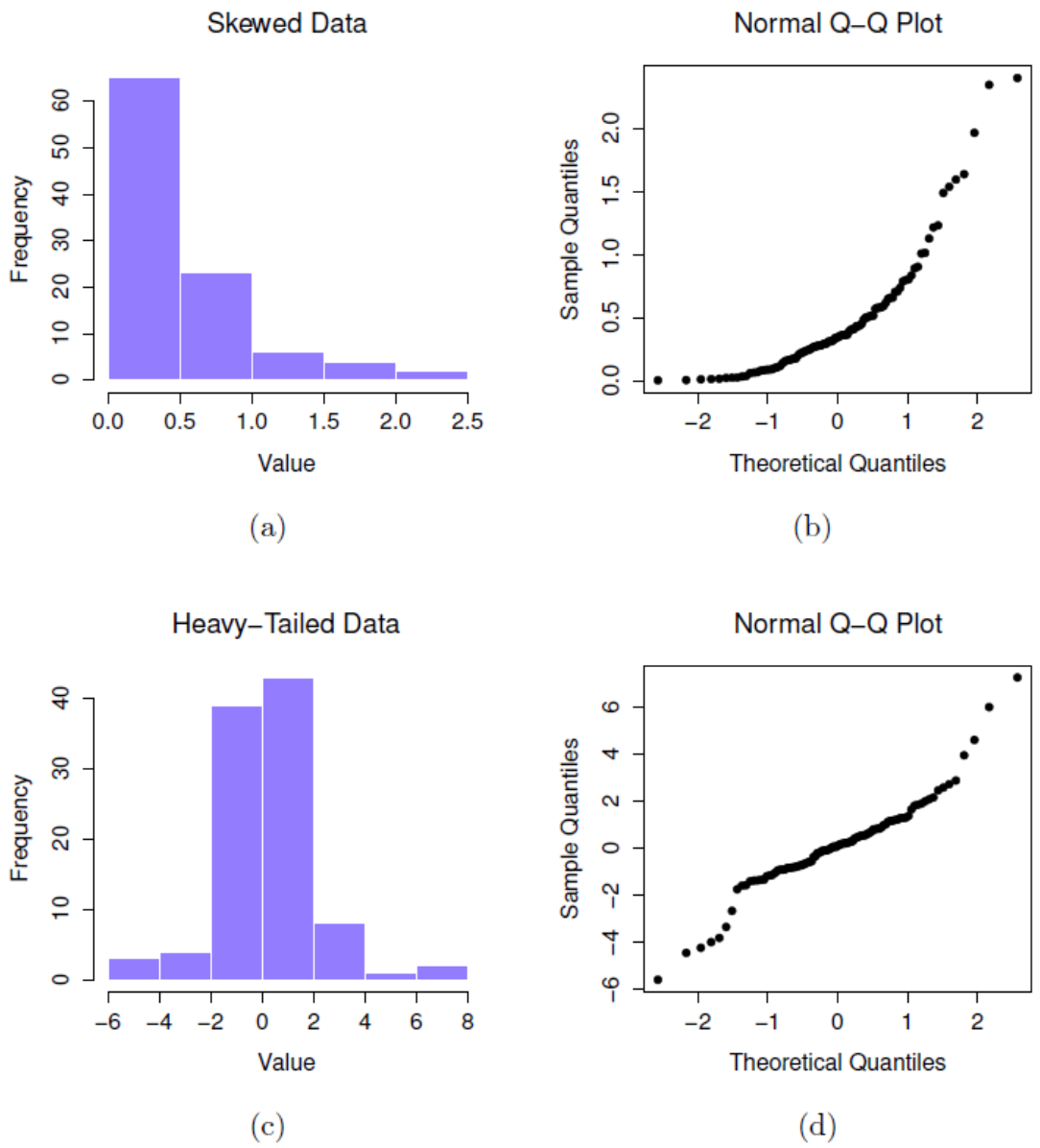


Figure 11-21 : Dans la rangée du haut, un histogramme (figure a) et un graphique QQ normal (figure b) des 100 observations d'un ensemble de données skewed.data. L'asymétrie des données ici est de 1,94 et se reflète dans un graphique QQ qui s'incurve vers le haut. En conséquence, la statistique de Shapiro-Wilk est W = .80, ce qui reflète un écart important par rapport à la normalité ($p < .001$). La rangée du bas montre les mêmes graphiques pour un ensemble de données à queues pointues, qui se compose de 100 observations. Dans ce cas, les queues pointues dans les données produisent une kurtosis élevée (2,80), et provoquent l'aplatissement de la courbe QQ au milieu, et une forte courbe de chaque côté. La statistique Shapiro-Wilk résultante est W = .93, ce qui reflète une fois de plus une non-normalité importante ($p < .001$).

Tests Shapiro-Wilk

Les graphiques QQ fournissent un bon moyen de vérifier de façon informelle la normalité de vos données, mais parfois vous voudrez faire quelque chose d'un peu plus formel et le **test Shapiro-Wilk** (Shapiro and Wilk 1965) est probablement ce que vous recherchez.[90] Comme on peut s'y attendre, l'hypothèse nulle testée est qu'un ensemble de *N* observations est normalement distribué.

La statistique du test qu'il calcule est conventionnellement désignée par *W*, et elle est calculée comme suit. Tout d'abord, nous trions les observations par ordre croissant de taille, X_1 étant la plus petite valeur de l'échantillon, X_2 étant la deuxième plus petite et ainsi de suite. Ensuite, la valeur de *W* est donnée par :

$$W = \frac{(\sum_{i=1}^{N} a_i X_i)^2}{\sum_{i=1}^{N} (X_i - \overline{X})^2}$$

où $\overline{X}$ est la moyenne des observations, et les valeurs de ai sont … heu … quelque chose de compliqué qui dépasse un peu la portée d'un texte d'introduction.

Comme il est un peu difficile d'expliquer les mathématiques qui se cachent derrière la statistique *W*, il est préférable de donner une description générale de la façon dont elle se comporte. Contrairement à la plupart des statistiques de test que nous rencontrerons dans ce livre, ce sont en fait de *petites* valeurs de *W* qui indiquent un écart par rapport à la normalité. La statistique *W* a une valeur maximale de 1, ce qui se produit lorsque les données semblent « parfaitement normales ». Plus la valeur de *W* est petite, moins les données sont normales. Cependant, la distribution d'échantillonnage pour *W*, qui n'est pas l'une des distributions standard dont j'ai parlé au chapitre 7 et qui est en fait un casse-tête, dépend de la taille de l'échantillon *N*. Pour vous donner une idée de ce à quoi ressemblent ces distributions, j'en ai représenté trois à la Figure 11-22. Notez que, à mesure que la taille de l'échantillon commence à augmenter, la distribution d'échantillonnage devient très serrée près de *W=1*, et par conséquent, pour les échantillons plus grands, *W* n'a pas besoin d'être beaucoup plus petit que 1 pour que le test soit significatif.

Pour obtenir la statistique Shapiro-Wilk dans les tests t avec Jamovi, cochez l'option « Normalité » sous « Hypothesis ». Dans les données échantillonnées au hasard (N = 100) que nous avons utilisées pour la courbe QQ, la valeur de la statistique du test de normalité de Shapiro-Wilk était W = 0,99 avec une *valeur p* de 0,69. Il n'est donc pas surprenant que nous n'ayons aucune preuve que ces données s'écartent de la normalité. Lorsque vous rapportez les résultats d'un test Shapiro-Wilk, vous devriez (comme d'habitude) vous

---

[90] Vous pouvez aussi utiliser le test de Kolmogorov-Smirnov, qui est probablement plus traditionnel que le Shapiro-Wilk. Bien que la plupart de ce que j'ai pu lire semblent suggérer que Shapiro-Wilk est le meilleur test de normalité, le Kolomogorov-Smirnov est un test général d'équivalence distributive qui peut être adapté pour traiter d'autres types de tests de distribution. Dans Jamovi, le test Shapiro-Wilk est préférable.

assurer d'inclure la statistique du test *W* et la valeur *p*, mais la distribution d'échantillonnage dépend tellement de *N* qu'il serait probablement bienvenu d'inclure également *N*.

Exemple

En attendant, cela vaut probablement la peine de vous montrer un exemple de ce qui arrive au graphiques QQ et au test de Shapiro-Wilk quand les données ne sont pas normales.

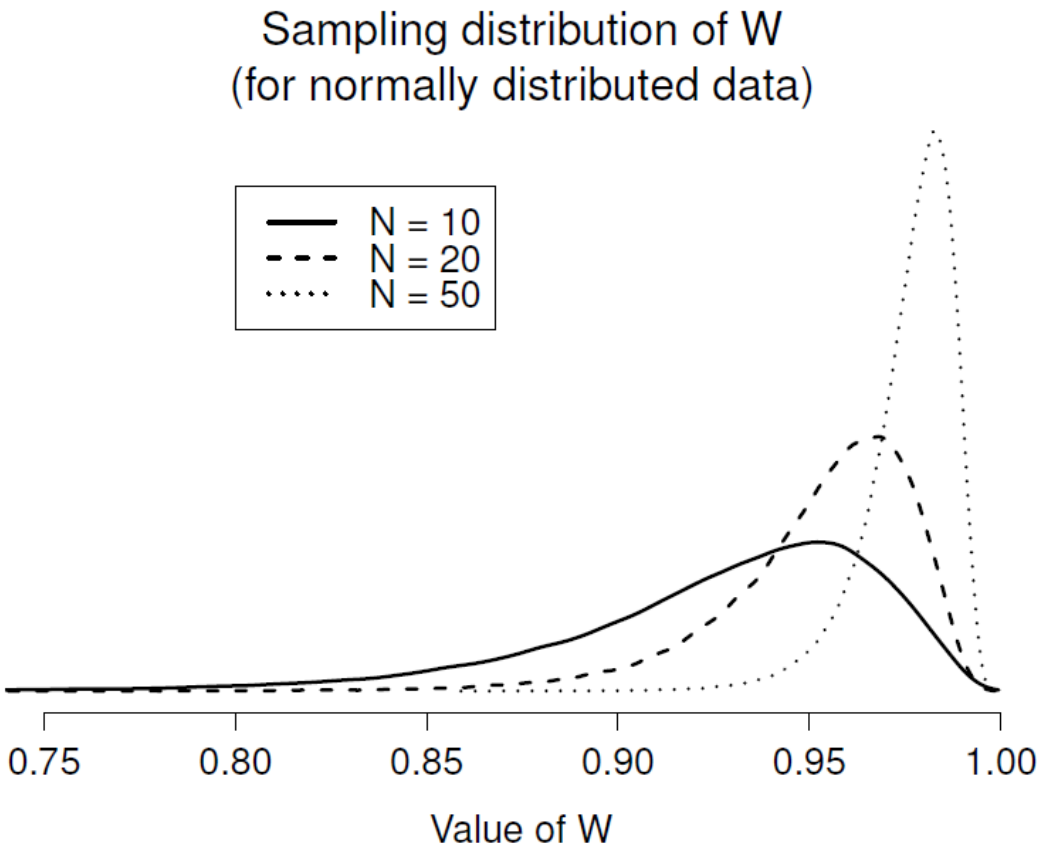
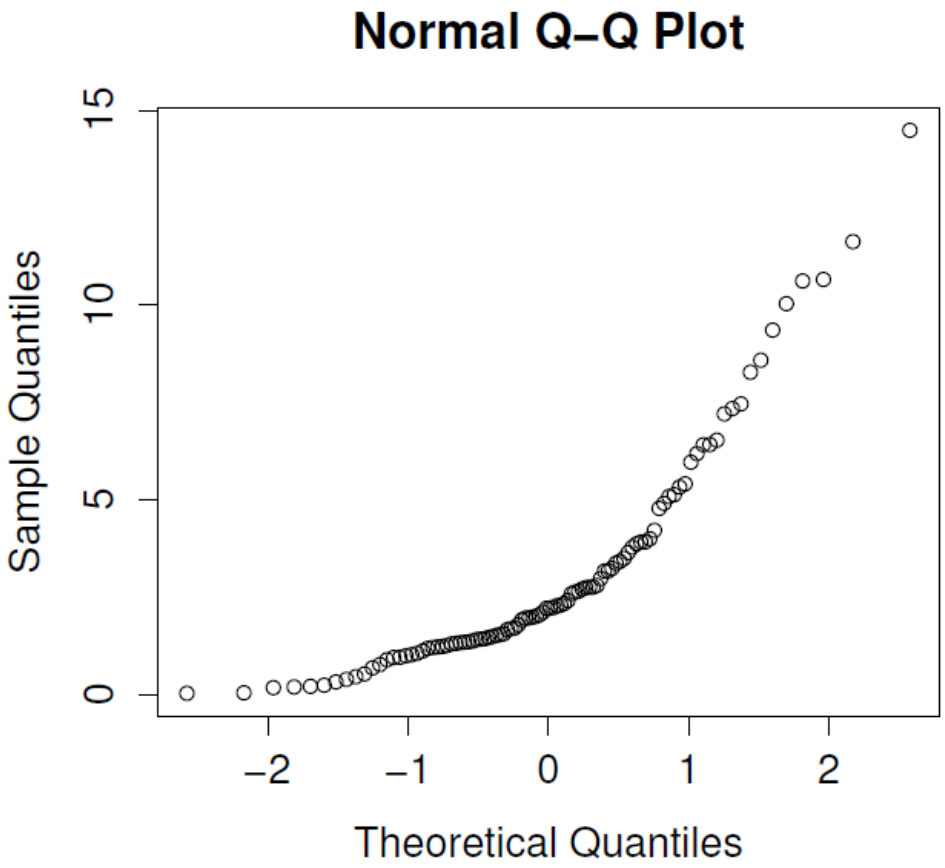


Figure 11-22 : Distribution d'échantillonnage de la statistique Shapiro-Wilk W, sous l'hypothèse nulle que les données sont normalement distribuées, pour des échantillons de taille 10, 20 et 50. Notez que les petites valeurs de W indiquent un écart par rapport à la normalité.

Pour cela, regardons la distribution de nos données de AFL winning margins, qui, si vous vous souvenez bien, au chapitre 4, ne semblaient pas du tout provenir d'une distribution normale. Voici ce qui arrive au graphique QQ :



Et lorsque nous exécutons le test Shapiro-Wilk sur les données des AFL margins, nous obtenons une valeur pour la statistique du test de normalité Shapiro-Wilk de W = 0,94, et une *valeur p* = 9,481e-07. Un effet clairement significatif !

Tester des données non-normales avec des tests de Wilcoxon

Bien, supposons que vos données s'avèrent être substantiellement non-normales, mais que vous voulez quand même faire un test t? Cette situation se produit souvent dans la vie réelle. Pour les données AFL winning margins, par exemple, le test Shapiro-Wilk a montré très clairement que l'hypothèse de normalité est violée. C'est la situation dans laquelle vous voulez utiliser les tests Wilcoxon.

Comme le *test t*, le test de Wilcoxon se présente sous deux formes, pour un échantillon et pour deux échantillons, et ils sont utilisés dans plus ou moins les mêmes situations que les *tests t* correspondants. Contrairement au *test t*, le test de Wilcoxon ne suppose pas la normalité, ce qui est bien. En fait, ils ne font aucune supposition quant au type de distribution en cause. Dans le jargon statistique, ce sont en fait des **tests non paramétriques**. Bien qu'éviter l'hypothèse de normalité soit une bonne chose, il y a un inconvénient : le test de Wilcoxon est habituellement moins puissant que le *test t* (c.-à-d. un taux d'erreur de type II plus élevé). Je ne parlerai pas des tests de Wilcoxon aussi en détail que des *tests t*, mais je vais vous en donner un bref aperçu.

Test Mann-Whitney U à deux échantillons

Je commencerai par décrire le **test Mann-Whitney U**, puisqu'il est en fait plus simple que la version à échantillon unique. Supposons que nous ayons les résultats de 10 personnes à un test. Puisque mon esprit m'a complètement déçu, faisons comme s'il s'agissait d'un « test de génialité » et que nous ayons deux groupes de personnes, « A » et « B ». Je suis curieux de savoir quel groupe est le plus génial. Les données sont incluses dans le fichier awesome.csv, et il y a deux variables en dehors de la variable ID habituelle : scores et groupe.

Tant qu'il n'y a pas d'égalité (c.-à-d. des personnes ayant exactement le même score de génialité), le test que nous voulons faire est étonnamment simple. Tout ce que nous avons à faire est de construire un tableau qui compare chaque observation du groupe *A* à chaque observation du groupe *B*. Lorsque le point de référence du groupe *A* est plus grand, nous plaçons une coche dans le tableau :

		group B				
		14.5	10.4	12.4	11.7	13.0
group A	6.4	.	.	.	.	.
	10.7	.	✓	.	.	.
	11.9	.	✓	.	✓	.
	7.3	.	.	.	.	.
	10.0	.	.	.	.	.

Nous comptons ensuite le nombre de cases à cocher. Voici notre statistique de test, *W*[91]. La distribution d'échantillonnage réelle pour *W* est quelque peu compliquée, et je vais sauter les détails. Pour nos besoins, il suffit de noter que l'interprétation de *W* est qualitativement la même que l'interprétation de *t* ou *z*. Autrement dit, si nous voulons un test bilatéral, nous rejetons l'hypothèse nulle lorsque *W* est très grand ou très petit, mais si nous avons une hypothèse orientée (c'est-à-dire unilatérale), nous utilisons seulement l'une ou l'autre.

Dans Jamovi, si nous exécutons un « T-Test d'échantillons indépendants » avec scores comme variable dépendante et un group comme variable de regroupement, et que nous cochons sous « tests » l'option « Mann-Whitney U », nous aurons des résultats montrant que U = 3 (c'est-à-dire le même nombre de marques de contrôle comme montré ci-dessus) et une *valeur p=0,05556*.

---

[91] En fait, il existe deux versions différentes de la statistique de test qui diffèrent l'une de l'autre par une valeur constante. La version que j'ai décrite est celle que Jamovi calcule

test de Wilcoxon pour un échantillon

Qu'en est-il du **test de Wilcoxon sur un échantillon** (ou de son équivalent, le test de Wilcoxon sur des échantillons appariés) ? Supposons que je sois intéressé à savoir si le fait de suivre un cours de statistique a un effet sur le bonheur des élèves. Mes données sont dans le fichier happiness.csv. Ce que j'ai mesuré ici, c'est le bonheur de chaque élève avant et après le cours, et le score de changement est la différence entre les deux. Tout comme nous l'avons vu avec le *test t*, il n'y a pas de différence fondamentale entre faire un test sur des échantillons appariés en utilisant les scores avant et après, et faire un test pour un échantillon unique en utilisant les scores de changement. Comme auparavant, la façon la plus simple de penser au test est de construire un tableau. La façon de procéder cette fois-ci consiste à prendre les scores de changement qui sont des différences positives et de les comparer à l'ensemble de l'échantillon. Ce que vous obtenez, c'est une table qui ressemble à ceci :

		all differences									
		-24	-14	-10	7	-6	-38	2	-35	-30	5
positive differences	7	.	.	.	✓	✓	.	✓	.	.	✓
	2	.	.	.	.	.	.	✓	.	.	.
	5	.	.	.	.	.	.	✓	.	.	✓

En comptant les marques cette fois-ci, nous obtenons une statistique de test de $W = 7$. Comme précédemment, si notre test est bilatéral, nous rejetons l'hypothèse nulle lorsque W est très grand ou très petit. Pour ce qui est de le calculer avec Jamovi, il est facile de savoir à quoi on peut s'attendre. Pour la version à un échantillon, vous spécifiez l'option « Wilcoxon rank » sous « Tests » dans la fenêtre d'analyse « One Sample T-Test ». Cela donne Wilcoxon $W = 7, p = 0{,}03711$. Comme on peut le constater, nous avons un effet significatif. Évidemment, suivre un cours de statistique a un effet sur votre bonheur. Bien sûr, utiliser une version pour échantillons appariés du test ne nous donnera pas une réponse différente; voir Figure 11-23.

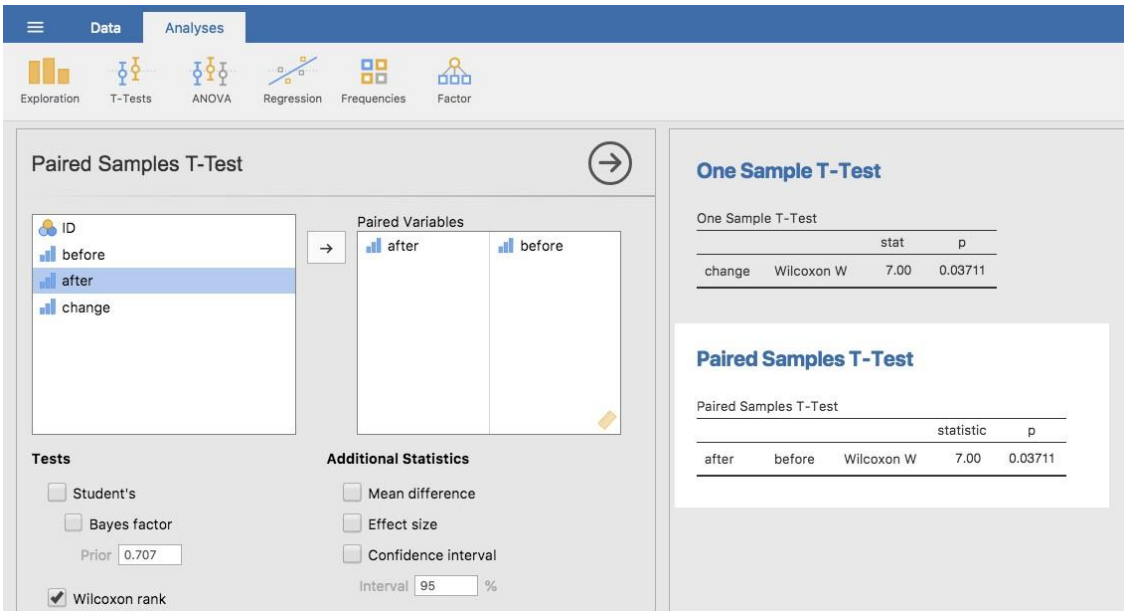


Figure 11-23 : Copie d'écran Jamovi montrant les résultats pour le test non paramétrique W de Wilcoxon sur un échantillon et des échantillons appariés

Résumé

- *Un test t sur* un échantillon permet de comparer la moyenne d'un seul échantillon à une valeur hypothétique de la moyenne de la population. (Section 11.2)
- Un *test t d'*échantillons indépendants est utilisé pour comparer les moyennes de deux groupes et tester l'hypothèse nulle qu'ils ont la même moyenne. Il se présente sous deux formes : le test de Student (section 11.3) suppose que les groupes ont le même écart-type, mais pas le test de Welch (section 11.4).
- Un *test t par* paires d'échantillons est utilisé lorsque vous avez deux scores de chaque personne, et que vous voulez tester l'hypothèse nulle que les deux scores ont la même moyenne. Cela équivaut à prendre la différence entre les deux scores pour chaque personne, puis à effectuer *test t sur un échantillon* sur les scores de différence. (Section 11.5)
- Les tests unilatéraux sont parfaitement légitimes à condition qu'ils soient planifiés à l'avance. (Section 11.6)
- La taille de l'effet pour la différence entre les moyennes peut être calculée à l'aide de la statistique *d de Cohen*. (Section 11.7).
- Vous pouvez vérifier la normalité d'un échantillon à l'aide des graphiques QQ (non disponibles actuellement dans Jamovi) et du test Shapiro-Wilk. (Section 11.8)
- Si vos données sont anormales, vous pouvez utiliser les tests de Mann-Whitney ou de Wilcoxon au lieu des *tests t*. (Section 11.9)

Corrélation et régression linéaire

Le but de ce chapitre est d'introduire la **corrélation** et la **régression linéaire**. Ce sont les outils standard sur lesquels les statisticiens s'appuient pour analyser la relation entre les prédicteurs continus et les résultats continus.

Corrélations

Dans cette section, nous verrons comment décrire les relations *entre les* variables des données. Pour ce faire, nous voulons surtout parler de la **corrélation** entre les variables. Mais d'abord, il nous faut des données.

Les données

Tableau 12-1 Statistiques descriptives pour les données sur la parentalité (parenthood data)

variable	min	max	mean	median	Std. dev	IQR
Grinchiosité de Dan	41	91	63,71	62	10,05	14
Les heures de sommeil de Dan	4,84	9,00	6,97	7,03	1,02	1,45
Les heures de sommeil du fils de Dan	3,25	12,07	8,05	7,95	2,07	3,21

Parlons d'un sujet qui tient à cœur à tous les parents : le sommeil. L'ensemble de données que nous utiliserons est fictif, mais basé sur des événements réels. Supposons que je sois curieux de savoir dans quelle mesure les habitudes de sommeil de mon fils influent sur mon humeur. Disons que je peux évaluer mon caractère grincheux très précisément, sur une échelle allant de 0 (pas du tout grincheux) à 100 (grincheux comme un vieil homme ou une vieille femme très, très grincheux). Et supposons aussi que je mesure ma mauvaise humeur, mes habitudes de sommeil et celles de mon fils depuis un certain temps déjà. Disons, pour 100 jours. Et, étant un intello, j'ai sauvegardé les données dans un fichier appelé parenthood.csv. Si nous chargeons les données, nous pouvons voir que le fichier contient quatre variables : dan.sleep, baby.sleep, dan.grump et day. Notez que lorsque vous chargez pour la première fois cet ensemble de données, Jamovi n'a peut-être pas deviné correctement le type de données pour chaque variable, auquel cas vous devrez le corriger : dan.sleep, baby.sleep, dan.grump et day peuvent être spécifiés comme variables continues, et ID est une variable nominale (entier).[92]

---

[92] J'ai remarqué que dans certaines versions de Jamovi vous pouvez aussi spécifier un type de variable'ID', mais pour nos besoins, peu importe comment nous spécifions la variable ID car nous ne l'inclurons dans aucune analyse.

Ensuite, j'examinerai quelques statistiques descriptives de base et, pour donner une représentation graphique de ce à quoi ressemble chacune des trois variables d'intérêt, la Figure 12-1 présente les histogrammes. Une chose à noter : ce n'est pas parce que Jamovi peut calculer des douzaines de statistiques différentes que vous devez les rapporter toutes. Si je devais rédiger ce rapport, je choisirais probablement les statistiques qui m'intéressent le plus (et qui intéressent mon lectorat), puis je les regrouperais dans un tableau simple et agréable comme celui du Tableau 12-1[93]. Remarquez que lorsque je l'ai mis dans un tableau, j'ai donné à tous des noms « en langage naturel ». C'est toujours une bonne pratique. Remarquez aussi que je ne dors pas assez. Ce n'est pas une bonne pratique, mais d'autres parents me disent que c'est la norme.

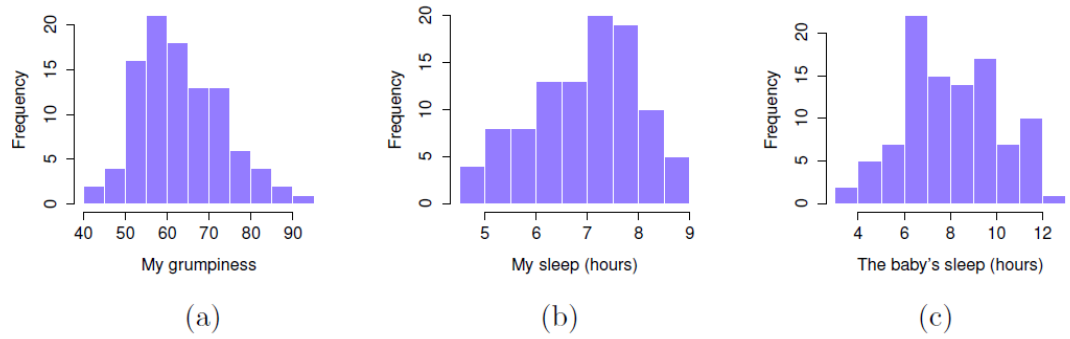


Figure 12-1 : Histogrammes pour les trois variables d'intérêt de l'ensemble de données sur la parentalité.

La force et l'orientation d'une relation

Nous pouvons dessiner des diagrammes de dispersion pour nous donner une idée générale du degré de corrélation entre deux variables. Idéalement, cependant, nous voudrions peut-être en dire un peu plus à ce sujet. Par exemple, comparons la relation entre dan.sleep et dan.grump (Figure 12-2, à gauche) avec celle entre baby.sleep et dan.grump (Figure 12-2, à droite). En regardant ces deux figures côte à côte, il est clair que la relation est *qualitativement* la même dans les deux cas : plus de sommeil égale moins de grinchiosité ! Cependant, il est aussi assez évident que la relation entre dan.sleep et dan.grump est *plus forte* que la relation entre baby.sleep et dan.grump. La figure de gauche est « plus nette » que celle de droite. Il me semble que si vous vouliez prédire mon humeur, cela vous aiderait un peu de savoir combien d'heures mon fils a dormi, mais ce serait *plus* utile de savoir combien d'heures j'ai dormi.

---

[93] En fait, même cette table, c'est plus que ce que j'aurais voulu. Dans la pratique, la plupart des gens choisissent *une* mesure de la tendance centrale et *une* mesure de la variabilité seulement.

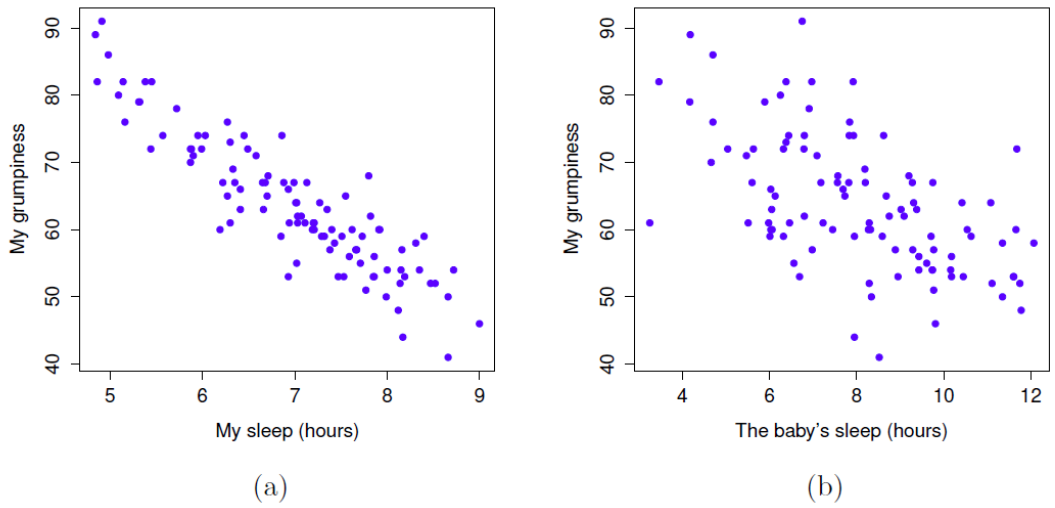


Figure 12-2 : Diagrammes de dispersion montrant la relation entre dan.sleep et dan.grump (à gauche) et la relation entre baby.sleep et dan.grump (à droite)

En revanche, considérons les deux diagrammes de dispersion illustrés à la Figure 12-3. Si l'on compare le nuage de points de « baby.sleep et dan.grump » (à gauche) au nuage de points de « baby.sleep et dan.sleep » (à droite), la force globale de la relation est la même, mais la direction est différente. C'est-à-dire, si mon fils dort plus, je dors *plus* (relation positive, côté droit), mais s'il dort plus, je suis *moins* grincheux (relation négative, côté gauche).

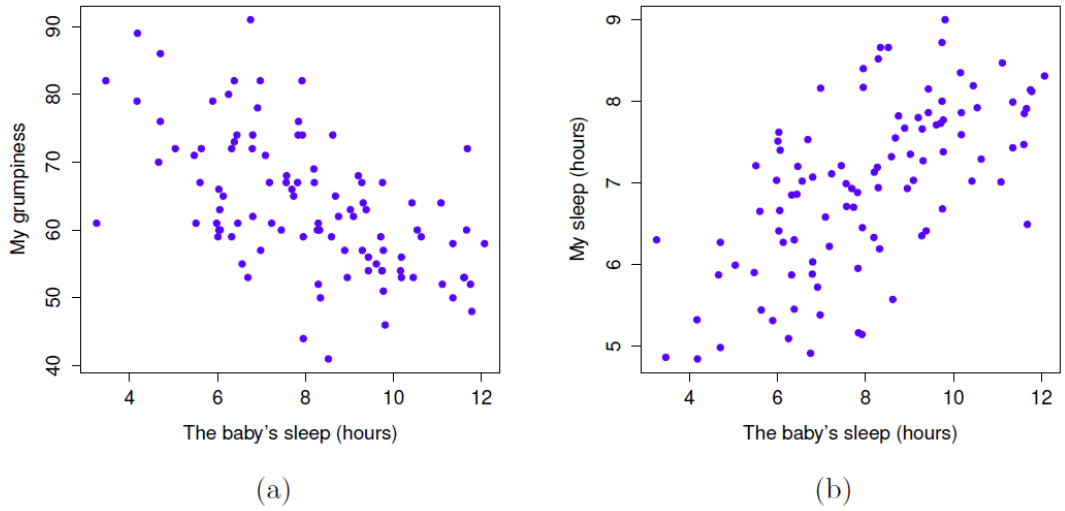


Figure 12-3 : Diagrammes de dispersion montrant la relation entre baby.sleep et dan.grump (à gauche), par rapport à la relation entre baby.sleep et dan.sleep (à droite).

Le coefficient de corrélation

Nous pouvons rendre ces idées un peu plus explicites en introduisant l'idée d'un **coefficient de corrélation** (ou, plus précisément, le coefficient de corrélation de Pearson), traditionnellement appelé *r*. Le coefficient de corrélation entre deux variables *X* et *Y* (parfois appelé r_{XY}), que nous allons définir plus précisément dans la section suivante, est une mesure qui varie entre -1 et 1. Quand *r=-1* signifie que nous avons une relation négative parfaite, et quand *r=1* signifie que nous avons une relation positive parfaite. Quand *r=0*, il n'y a pas de relation du tout. Si vous regardez la Figure 12-4, vous pouvez voir plusieurs graphiques montrant à quoi ressemblent différentes corrélations.

La formule du coefficient de corrélation de Pearson peut être écrite de plusieurs façons. Je pense que la façon la plus simple d'écrire la formule est de la diviser en deux étapes. Tout d'abord, introduisons l'idée d'une **covariance**. La covariance entre deux variables *X* et *Y* est une généralisation de la notion de variance et est un moyen mathématiquement simple de décrire la relation entre deux variables qui n'est pas très instructive pour l'homme.

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

Puisque nous multiplions (c'est-à-dire que nous prenons le « produit ») d'une quantité qui dépend de *X* par une quantité qui dépend de *Y* et que nous calculons ensuite la moyenne[94], vous pouvez considérer la formule de la covariance comme un « produit croisé moyen » entre *X* et *Y*.

[94] Tout comme nous l'avons vu avec la variance et l'écart-type, en pratique nous divisons par *N-1* au lieu de *N*

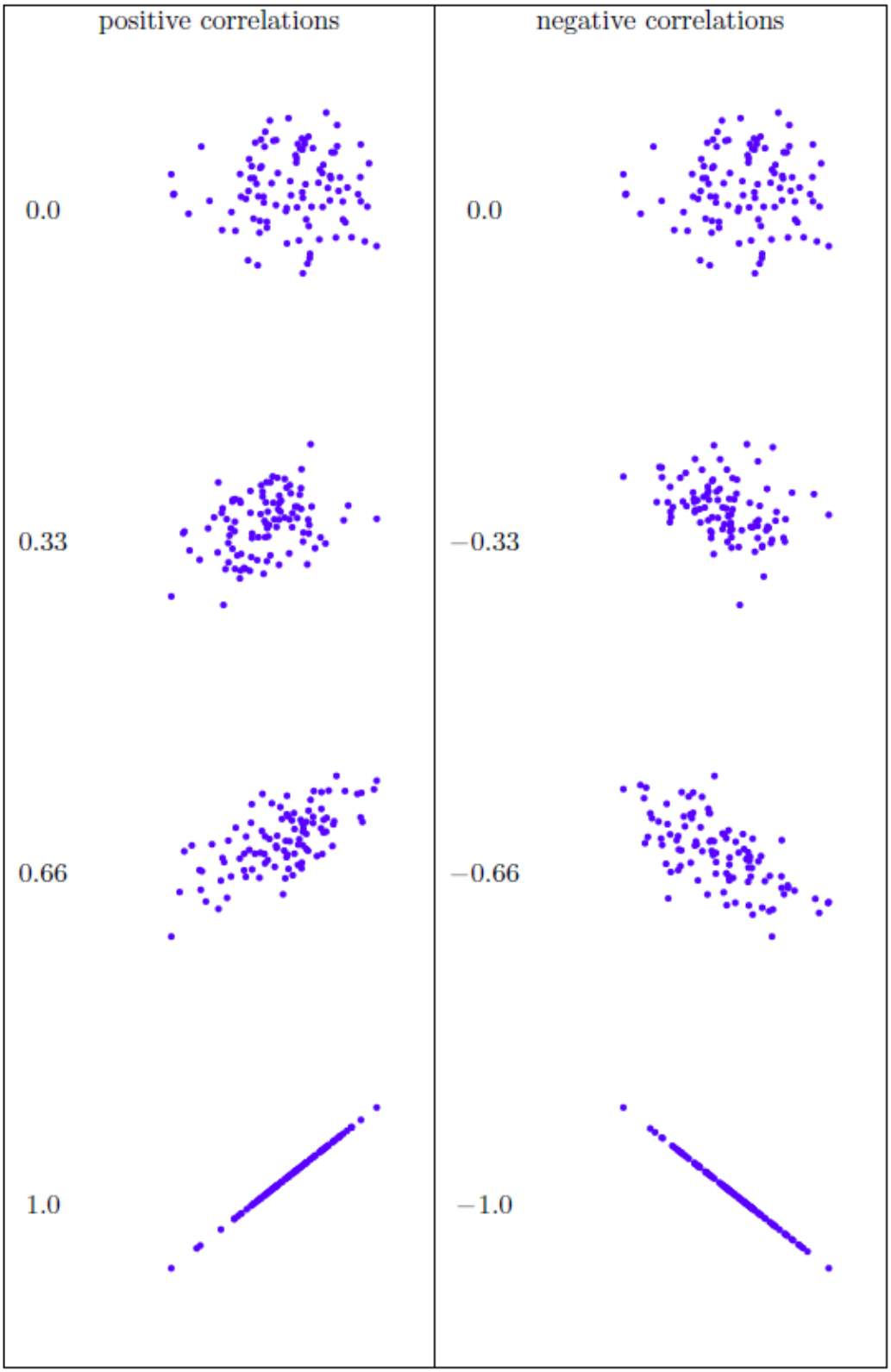

positive correlations
negative correlations
0.0
0.0
0.33
−0.33
0.66
−0.66
1.0
−1.0

Figure 12-4 : Illustration de l'effet de la variation de l'intensité et de la direction d'une corrélation. Dans la colonne de gauche, les corrélations sont 0; 0,33; 0,66 et 1 ; dans la colonne de droite, les corrélations sont 0; -0,33; -0,66 et -1.

La covariance a la propriété intéressante que, si *X* et *Y* sont entièrement indépendants, alors la covariance est exactement nulle. Si la relation entre eux est positive (au sens de la Figure 12-4), la covariance est également positive, et si la relation est négative, la covariance est également négative. En d'autres termes, la covariance saisit l'idée qualitative fondamentale de corrélation. Malheureusement, l'amplitude brute de la covariance n'est pas facile à interpréter car elle dépend des unités dans lesquelles *X* et *Y* sont exprimés et, pire encore, les unités réelles dans lesquelles la covariance elle-même est exprimée sont vraiment étranges. Par exemple, si *X* fait référence à la variable dan.sleep (unités : heures) et *Y* à la variable dan.grump (unités : grinchiosité), alors les unités pour leur covariance sont « heures X grinchiosité ». Je n'ai aucune idée de ce que ça peut vouloir dire.

Le coefficient de corrélation de Pearson *r* corrige ce problème d'interprétation en normalisant la covariance, à peu près de la même manière que le *score z* normalise un score brut, en divisant par l'écart type. Cependant, comme nous avons deux variables qui contribuent à la covariance, la standardisation ne fonctionne que si nous divisons par les deux écarts-types[95]. En d'autres termes, la corrélation entre *X* et *Y* peut être écrite comme suit :

$$r_{XY} = \frac{Cov(X,Y)}{\hat{\sigma}_X \hat{\sigma}_y}$$

En standardisant la covariance, non seulement nous conservons toutes les belles propriétés de la covariance discutée précédemment, mais les valeurs réelles de *r* sont sur une échelle significative : *r=1* implique une relation positive parfaite et *r=-1* implique une relation négative parfaite. Je reviendrai un peu plus en détail sur ce point plus tard, à la section 12.1.5 Mais avant, voyons comment calculer les corrélations dans Jamovi.

Calcul des corrélations dans Jamovi

Le calcul des corrélations dans Jamovi peut se faire en cliquant sur le menu « Regression » - « Correlation Matrix ». Transférez les quatre variables continues dans la boîte de droite pour obtenir le résultat de la Figure 12-5.

[95] C'est une simplification exagérée, mais ça fera l'affaire pour nous.

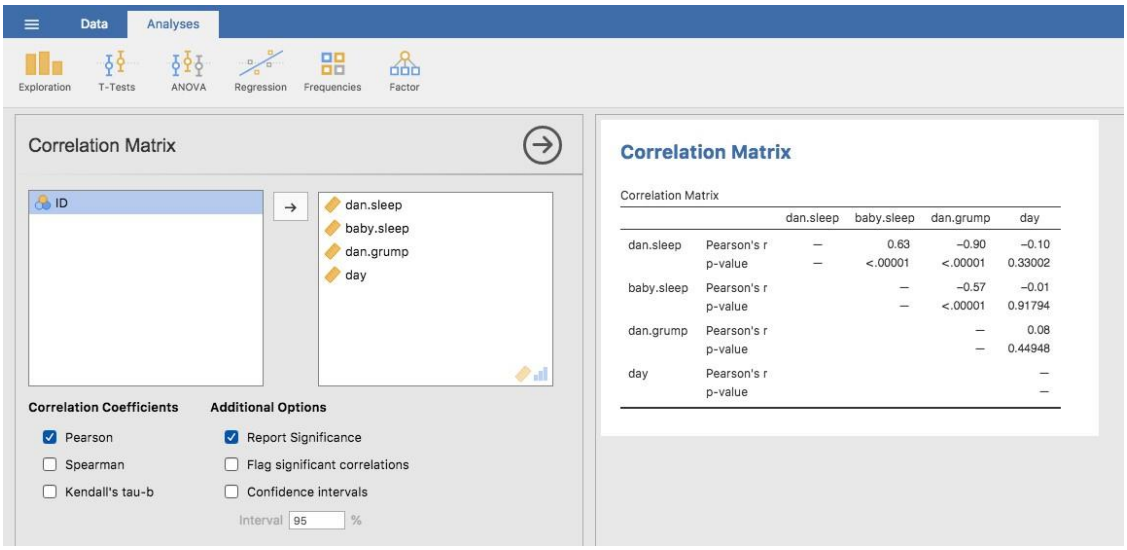


Figure 12-5 : Une capture d'écran Jamovi montrant les corrélations entre les variables du fichierparenthood.csv

Interpréter une corrélation

Naturellement, dans la vie réelle, on ne voit pas beaucoup de corrélations de 1. Alors comment interpréter une corrélation de, disons, *r=0,4* ? La réponse honnête est que cela dépend vraiment de la raison pour laquelle vous voulez utiliser les données et de l'importance des corrélations dans votre domaine. Un de mes amis en ingénierie m'a dit un jour que toute corrélation inférieure à *0,95* est complètement inutile (je pense qu'il exagérait, même pour l'ingénierie). D'un autre côté, il y a des cas réels, même en psychologie, où il faut vraiment s'attendre à des corrélations aussi fortes. Par exemple, l'un des ensembles de données de référence utilisés pour tester les théories sur la façon dont les gens jugent les similarités est si spécifique que toute théorie qui ne peut pas atteindre une corrélation d'au moins *0,9* n'est pas vraiment considéré comme un succès. Cependant, lorsque vous recherchez (disons) des corrélats élémentaires d'intelligence (p. ex. temps d'inspection, temps de réponse), si vous obtenez une corrélation supérieure à *0,3*, c'est un bon résultat. Bref, l'interprétation d'une corrélation dépend beaucoup du contexte. Cela dit, le guide approximatif du Tableau 12-2 est assez typique.

Tableau 12-2 : Un guide approximatif pour interpréter les corrélations. Notez que je dis un guide approximatif. Il n'y a pas de règles strictes pour ce qui est des relations fortes ou faibles. Cela dépend du contexte.

Corrélation	Force	Direction
-1,0 à -0,9	Très fort	Négatif
-0,9 à -0,7	Fort	Négatif

-0,7 à -0,4	Modéré	Négatif
-0,4 à -0,2	Faible	Négatif
-0,2 à 0	Négligeable	Négatif
0 à 0,2	Négligeable	Positif
0,2 à 0,4	Faible	Positif
0,4 à 0,7	Modéré	Positif
0,7 à 0,9	Fort	Positif
0,9 à 1,0	Très fort	Positif

Cependant, ce qu'on ne soulignera jamais assez, c'est qu'il faut *toujours* regarder le nuage de points avant de faire toute interprétation des données. Une corrélation peut ne pas signifier ce que vous pensez qu'elle signifie. L'illustration classique est fourni par « le Quatuor d'Anscombe » (Anscombe 1973), une collection de quatre ensembles de données. Chaque ensemble de données comporte deux variables, un *X* et un *Y*. Pour les quatre données la valeur moyenne pour *X* à 9 et la moyenne pour *Y* à 7,5. Les écarts-types pour toutes les variables *X* sont presque identiques, tout comme ceux des variables *Y.* Et dans chaque cas, la corrélation entre *X* et *Y* est *r=0,816*. Vous pouvez le vérifier vous-même, car il se trouve que je l'ai sauvegardé dans un fichier appelé anscombe.csv.

On pourrait penser que ces quatre ensembles de données se ressemblent beaucoup. Ce n'est pas le cas. Si nous dessinons des diagrammes de dispersion de *X* par rapport à *Y* pour les quatre variables, comme le montre la Figure 12-6, nous constatons que ces quatre variables sont *très* différentes les unes des autres. La leçon ici, que tant de gens semblent oublier dans la vie réelle, est qu'il faut *toujours tracer un graphique de vos données brutes* (chapitre 5).

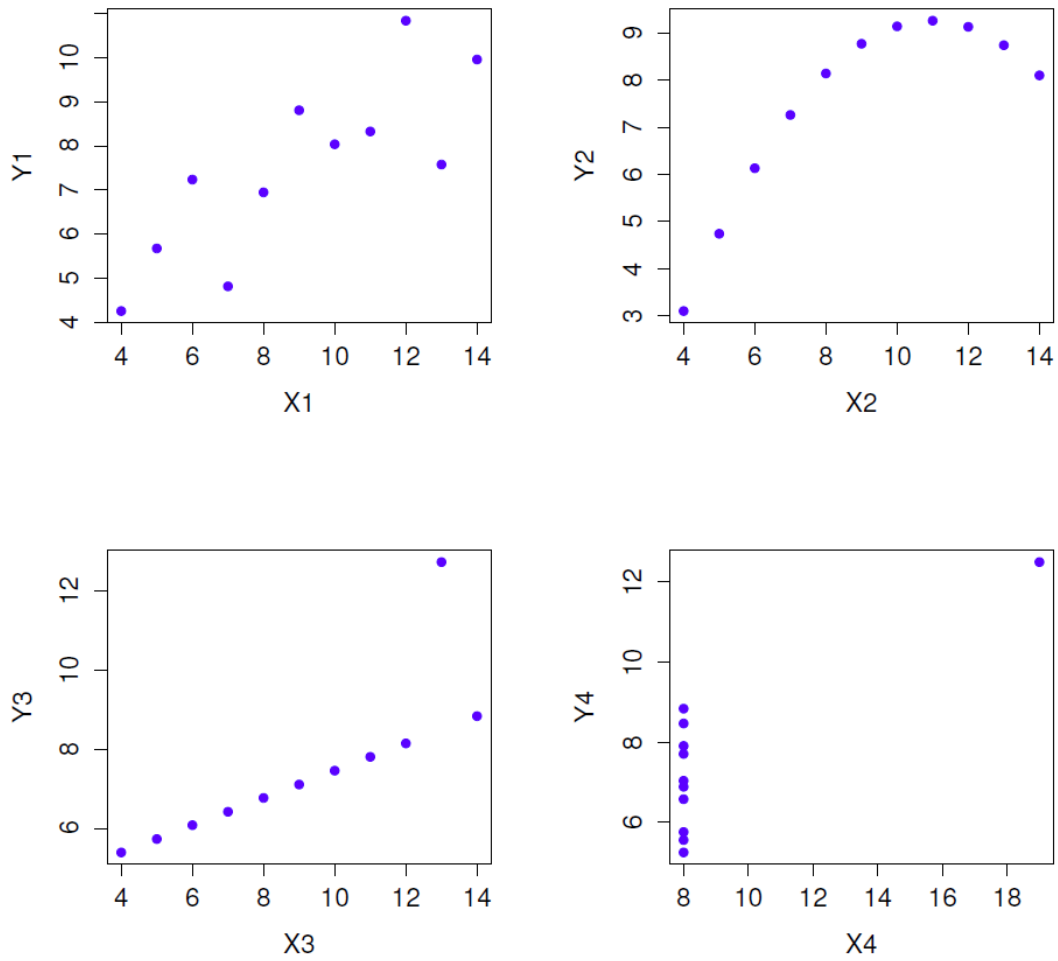


Figure 12-6 : Le quatuor d'Anscombe. Ces quatre ensembles de données ont une corrélation de Pearson de r = 0,816, mais ils sont qualitativement différents les uns des autres.

Corrélations de rang de Spearman

Le coefficient de corrélation de Pearson est utile pour beaucoup de choses, mais il comporte des lacunes. Une question en particulier ressort : ce qu'il mesure en fait, c'est la force de la relation *linéaire* entre deux variables. En d'autres termes, ce qu'il vous donne est une mesure de la tendance des données à toutes s'aligner sur une seule ligne parfaitement droite. Souvent, il s'agit d'une assez bonne approximation de ce que nous entendons par « relation », et il est pertinent de calculer la corrélation de Pearson. Mais parfois, ça ne l'est pas.

Une situation très courante où la corrélation de Pearson n'est pas tout à fait ce qu'il faut utiliser survient lorsqu'une augmentation d'une variable *X* se reflète réellement dans une augmentation d'une autre variable *Y*, mais que la nature de la relation n'est pas nécessairement linéaire. Un exemple de cela pourrait être la relation entre l'effort et la récompense lorsqu'on étudie pour un examen. Si vous ne faites aucun effort (*X*) pour apprendre une matière, vous devez vous attendre à obtenir une note de 0 % (*Y*). Cependant,

un peu d'effort entraînera une amélioration *massive*. Le simple fait d'assister à des cours magistraux vous permet d'apprendre pas mal de choses, et si vous venez en classe et que vous griffonnez quelques notes, votre score pourrait atteindre 35 %, le tout sans beaucoup d'efforts. Cependant, vous n'obtenez pas le même effet à l'autre bout de l'échelle. Comme tout le monde le sait, il faut *beaucoup* plus d'efforts pour obtenir une note de 90 % que pour obtenir une note de 55 %. Ce que cela signifie, c'est que, si j'ai des données sur l'effort d'étude et les notes, il y a de fortes chances que les corrélations de Pearson soient trompeuses.

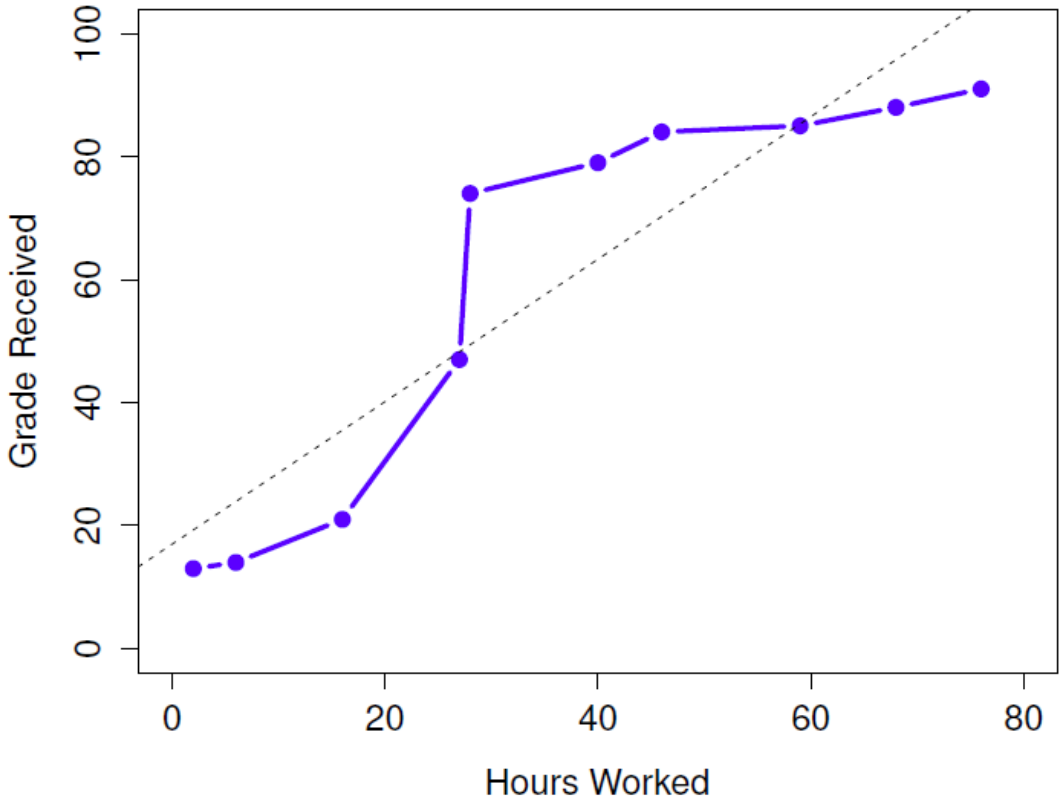


Figure 12-7 : La relation entre les heures travaillées et la note reçue pour un ensemble de données fictif composé de seulement 10 élèves (chaque point correspond à un élève). La ligne en pointillés au milieu montre la relation linéaire entre les deux variables. Ceci produit une forte corrélation de Pearson de *r=0,91*. Cependant, ce qu'il est intéressant de noter ici, c'est qu'il y a en fait… une relation monotone parfaite entre les deux variables. Dans cet exemple fictif, l'augmentation des heures travaillées accroît toujours la note reçue, comme l'illustre le trait plein. Cela se reflète dans une corrélation Spearman de 1. Avec un si petit ensemble de données, cependant, c'est une question ouverte que de savoir quelle version décrit le mieux la relation réelle impliquée.

A titre d'exemple, examinons les données illustrées à la Figure 12-7, qui montrent la relation entre les heures travaillées et la note reçue pour 10 élèves qui suivent un cours. Ce qui est curieux dans cet ensemble de données (hautement fictif), c'est que le fait d'augmenter votre effort augmente *toujours* votre note. Il peut être de beaucoup ou d'un peu, mais un effort accru ne diminuera jamais votre note. Si nous utilisons une corrélation standard de Pearson, elle montre une forte relation entre les heures travaillées et la note reçue, avec un coefficient de corrélation de 0,91. Toutefois, cela ne tient pas compte du fait que l'augmentation du nombre d'heures travaillées fait *toujours* augmenter la note. Il y a ici un sens dans lequel nous voulons pouvoir dire que la corrélation est *parfaite* mais pour une notion quelque peu différente de ce qu'est une « relation ». Ce que nous recherchons, c'est quelque chose qui saisit le fait qu'il y a une **relation ordinale** parfaite ici. Autrement dit, si l'élève 1 travaille plus d'heures que l'élève 2, nous pouvons garantir que l'élève 1 obtiendra la meilleure note. Ce n'est pas du tout ce que signifie une corrélation de *r=0,91*.

Comment devrions-nous aborder cette question ? En fait, c'est très facile. Si nous recherchons des relations ordinales, tout ce que nous avons à faire est de traiter les données comme s'il s'agissait d'une échelle ordinale ! Ainsi, au lieu de mesurer l'effort en termes « d'heures travaillées », classons les 10 étudiants par ordre d'heures travaillées. C'est-à-dire que l'élève 1 a fait le moins de travail de n'importe qui (2 heures), de sorte qu'il obtient le rang le plus bas (rang = 1). L'étudiant 4 était le deuxième plus paresseux, n'ayant travaillé que 6 heures sur l'ensemble du semestre, de sorte qu'il obtient le deuxième rang le plus bas (rang = 2). Notez que j'utilise « rang = 1 » pour signifier « rang faible ». Parfois, dans le langage courant, on parle de « rang = 1 » pour signifier « rang le plus élevé » plutôt que « rang le plus bas ». Soyez donc prudent, vous pouvez classer « de la plus petite valeur à la plus grande valeur » (c.-à-d. petit égal 1) ou « de la plus grande valeur à la plus petite valeur » (c.-à-d., grand égal 1). Dans ce cas, je classe du plus petit au plus grand, mais comme il est vraiment facile d'oublier la façon dont vous organisez les choses, il faut faire un effort d'attention !

Jetons donc un coup d'œil à nos élèves quand nous les classons du pire au meilleur en termes d'effort et de récompense :

	rang (heures travaillées)	rang (note reçue)
élève 1	1	1
élève 2	10	10
élève 3	6	6
élève 4	2	2
élève 5	3	3
élève 6	5	5
élève 7	4	4

élève 8	8	8
élève 9	7	7
étudiant 10	9	9

Ils sont *pareils*. L'étudiant qui a fait le plus d'efforts a obtenu la meilleure note, l'étudiant qui a fait le moins d'efforts a obtenu la pire note, etc. Comme le montre le tableau ci-dessus, ces deux classements sont identiques, donc si nous les calculons une corrélation maintenant, nous obtenons une relation parfaite, avec une corrélation de 1.

Ce que nous venons de réinventer, c'est **la corrélation par rang de Spearman**, habituellement désignée ρ pour la distinguer de la corrélation de Pearson r. Nous pouvons calculer la corrélation de Spearman ρ en utilisant Jamovi simplement en cochant la case « Spearman » dans « Correlation Matrix ».

Diagrammes de dispersion

Les nuages de points sont un outil simple mais efficace pour visualiser la relation entre *deux* variables, comme nous l'avons vu avec les chiffres de la section sur la corrélation (Section 12.1). C'est cette dernière application que nous avons généralement en tête lorsque nous utilisons le terme « nuage de points ». Dans ce type de graphique, chaque observation correspond à un point. L'emplacement horizontal du point trace la valeur de l'observation sur une variable et l'emplacement vertical affiche sa valeur sur l'autre variable. Dans de nombreuses situations, vous n'avez pas vraiment d'opinion claire sur la nature de la relation *causale* (p. ex., est-ce que A cause B, ou est-ce que B cause A, ou est-ce qu'une autre variable C contrôle à la fois A et B). Si c'est le cas, peu importe quelle variable vous tracez sur l'axe des x et celle que vous tracez sur l'axe des y. Cependant, dans de nombreuses situations, vous avez une idée assez précise de la variable que vous pensez être la plus susceptible d'être causale, ou du moins vous avez quelques soupçons pour cela. Si c'est le cas, il est conventionnel de tracer la variable cause sur l'axe des x, et la variable effet sur l'axe des y. En gardant cela à l'esprit, voyons comment dessiner des nuages de points dans Jamovi, en utilisant le même ensemble de données sur la parentalité (i.e. parenthood.csv) que celui que j'ai utilisé pour introduire les corrélations.

Supposons que mon but est de dessiner un nuage de points montrant la relation entre la quantité de sommeil que j'obtiens (dan.sleep) et à quel point je suis grincheux le jour suivant (dan.grump). Il y a deux façons différentes d'utiliser Jamovi pour obtenir le graphique que nous recherchons. La première façon est d'utiliser l'option « Tracé » sous le bouton « Régression » - « Matrice de corrélation », ce qui nous donne la sortie indiquée dans la Figure 12-8. Notez que Jamovi trace une ligne à travers les points, nous y reviendrons un peu plus loin dans la section 12.3. Le tracé d'un nuage de points de cette façon vous permet également de spécifier des « densités pour les variables » et cette option ajoute une courbe de densité montrant comment les données de chaque variable sont réparties.

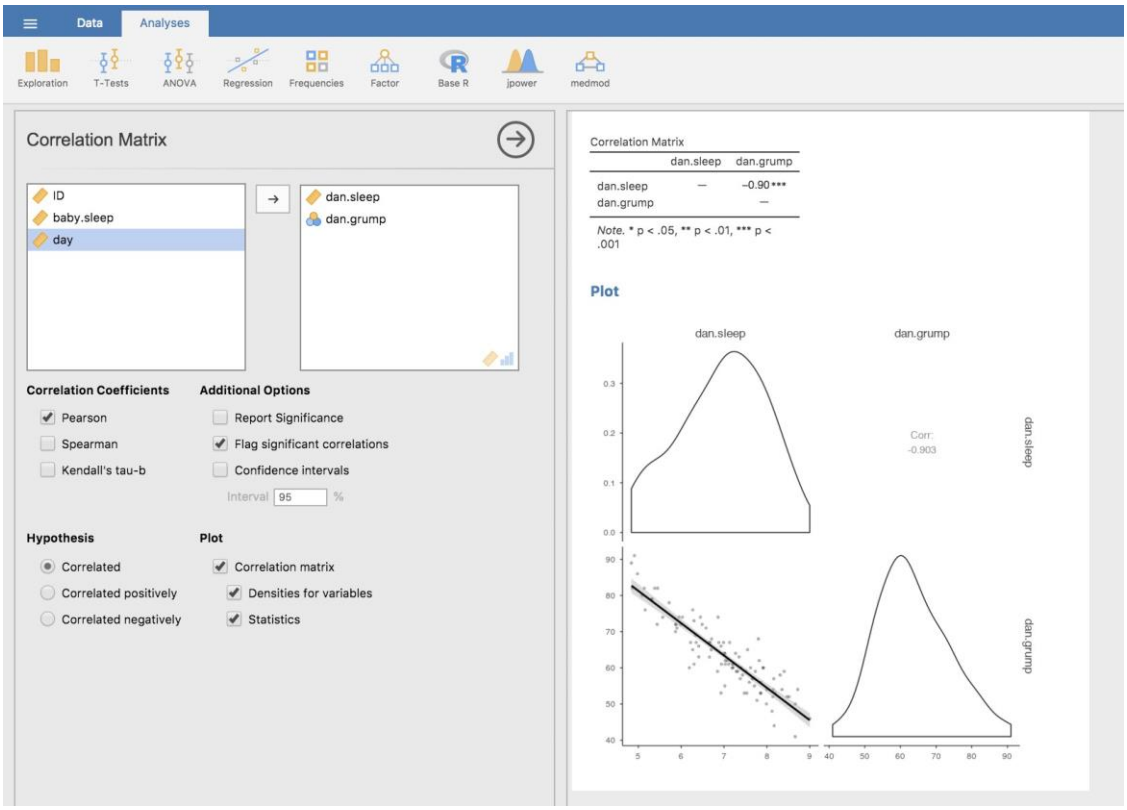


Figure 12-8 : Diagramme de dispersion via la commande « Correlation Matrix » dans Jamovi

La deuxième façon de faire est d'utiliser l'un des modules complémentaires Jamovi. Ce module s'appelle « Scatr » et vous pouvez l'installer en cliquant sur la grande icône « + » en haut à droite de l'écran Jamovi, en ouvrant la bibliothèque Jamovi, en faisant défiler vers le bas jusqu'à trouver « scatr » et en cliquant sur « Install ». Une fois que vous avez fait cela, vous trouverez une nouvelle commande « Scatterplot » disponible sous le bouton « Exploration ». Ce graphique est un peu différent de la première façon, voir la Figure 12-9, mais l'information importante est la même.

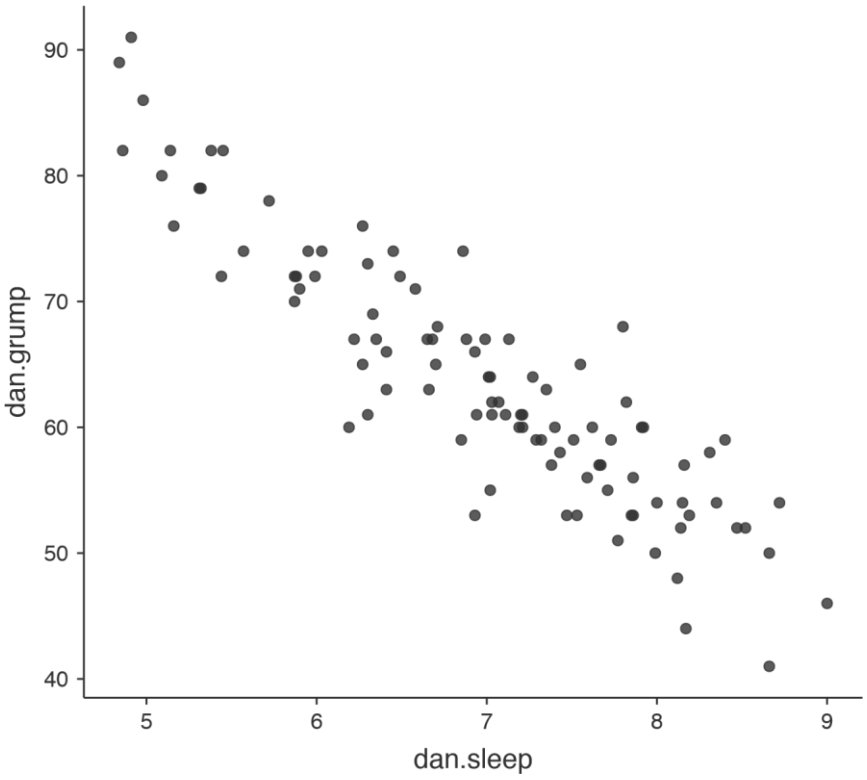


Figure 12-9 : Diagramme de dispersion via le module additionnel « Scatr » dans Jamovi

Des options plus élaborées

Souvent, vous voudrez examiner les relations entre plusieurs variables à la fois, en utilisant une **matrice de nuage de points** (dans Jamovi via la commande « Correlation Matrix » - « Plot »). Ajoutez simplement une autre variable, par exemple baby.sleep à la liste des variables à corréler, et Jamovi créera une matrice de nuage de points pour vous, tout comme celle de la Figure 12-10.

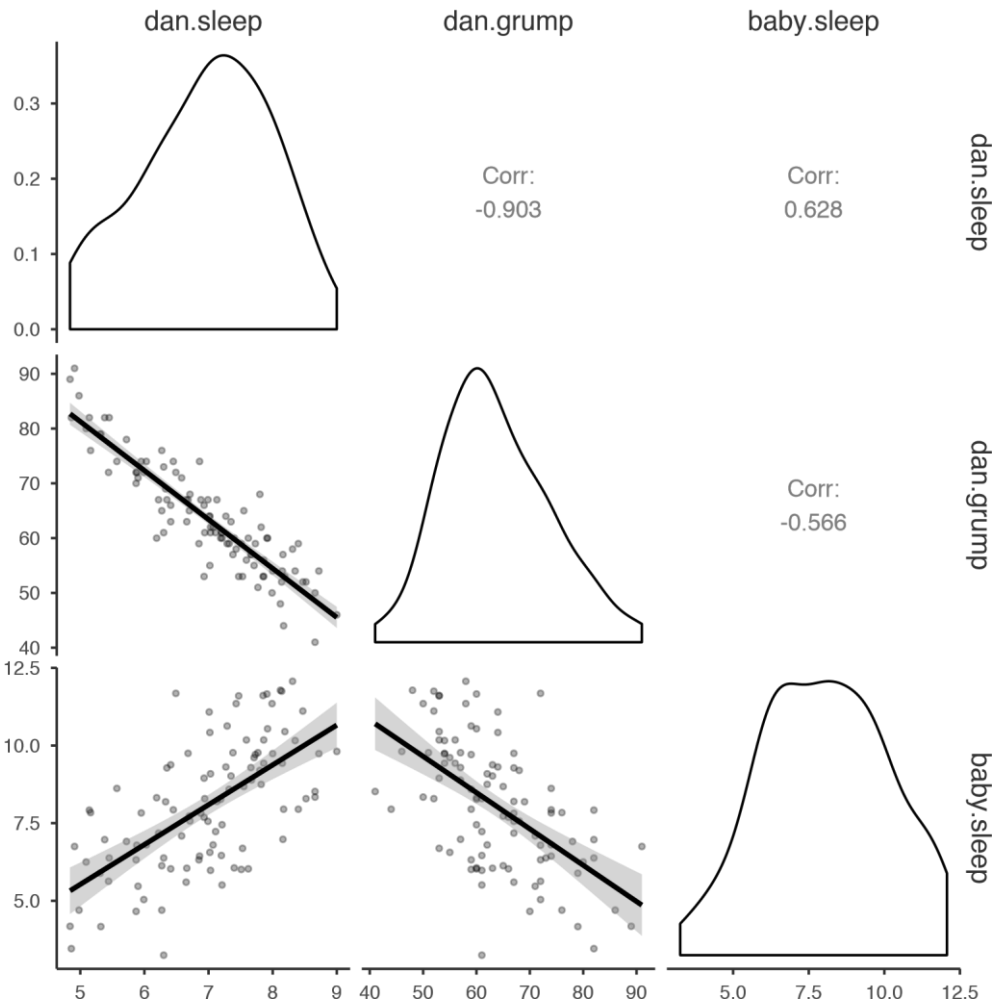


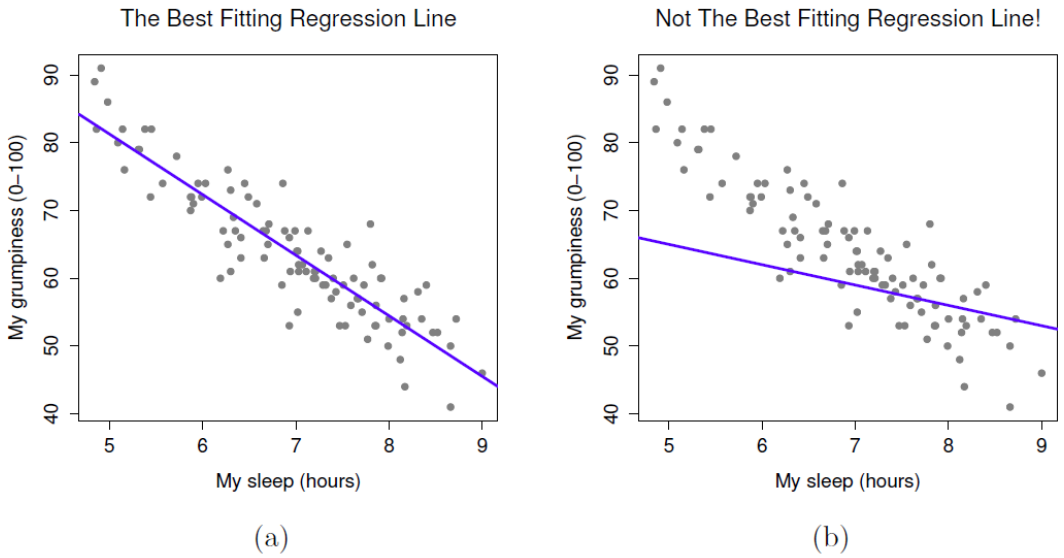
Figure 12-10 : Une matrice de nuages de points réalisé à l'aide de Jamovi.

Qu'est-ce qu'un modèle de régression linéaire ?

Les modèles de régression linéaire, réduits à l'essentiel, sont essentiellement une version légèrement plus sophistiquée de la corrélation de Pearson (section 12.1), bien que, comme nous le verrons plus loin, les modèles de régression soient des outils beaucoup plus puissants.

Puisque les idées de base de la régression sont étroitement liées à la corrélation, nous reviendrons au fichier parenthood.csv que nous utilisions pour illustrer le fonctionnement des corrélations. Rappelons que, dans cet ensemble de données, nous essayions de découvrir pourquoi Dan est si grincheux tout le temps et notre hypothèse de travail était que je ne dors pas assez. Nous avons dessiné quelques nuages de points pour nous aider à examiner la relation entre la quantité de sommeil que j'obtiens et ma mauvaise humeur le lendemain, comme dans la Figure 12-9, et comme nous l'avons vu précédemment cela correspond à une corrélation de *r=-0,90*, mais ce que nous imaginons secrètement est quelque chose qui ressemble davantage à la Figure 12-11a. C'est-à-dire que nous traçons mentalement une ligne droite au milieu des données. En statistique, cette ligne que nous

traçons s'appelle une **ligne de régression**. Notez que, puisque nous ne sommes pas des idiots, la ligne de régression passe au milieu des données. Nous ne nous trouvons pas à imaginer quoi que ce soit qui ressemble à l'intrigue plutôt stupide illustrée à la Figure 12-11b.



(a) (b)

Figure 12-11: La figure (a) montre le diagramme de dispersion du sommeil de la Figure 12-9 avec la ligne de régression la mieux adaptée tracée au-dessus du sommet. Comme on pouvait s'y attendre, la ligne passe au milieu des données. En revanche, la figure(b) montre les mêmes données, mais avec un très mauvais choix de ligne de régression tracée par-dessus.

Ce n'est pas très surprenant. La ligne que j'ai tracée dans la Figure 12-11b ne correspond pas très bien aux données, donc il n'est pas très logique de la proposer comme moyen de résumer les données. C'est une observation très simple à faire, mais elle s'avère très puissante lorsqu'on commence à essayer de l'entourer d'un peu de mathématiques. Pour ce faire, commençons par un rafraîchissement de quelques notions mathématiques du secondaire. La formule pour une ligne droite est généralement écrite comme ceci

$$y = a + bx$$

Ou, du moins, c'est ce que c'était quand je suis allé au lycée, il y a longtemps. Les deux variables sont x et y, et nous avons deux *coefficients*, a et b.[96] Le coefficient a représente l'intersection y de la ligne, et le coefficient b représente la *pente* de la ligne. En remontant plus loin dans nos souvenirs délabrés du lycée (désolé, pour certains d'entre nous, le lycée c'était il y a longtemps), nous nous souvenons que l'intersection est interprétée comme « la

[96] Également parfois écrit comme y = mx + c où m est le coefficient de pente et c est le coefficient d'intersection (constante).

valeur de *y* que l'on obtient quand *x=0* »[97]. De même, une pente de *b* signifie que si vous augmentez la *valeur x de* 1 unité, alors la *valeur y* augmente de *b* unités, et une pente négative signifie que la *valeur y* diminuerait au lieu de monter. Mais bien sûr, tout me revient. Maintenant que nous nous sommes rappelés qu'il ne faut pas s'étonner de découvrir que nous utilisons exactement la même formule pour une droite de régression. Si *Y* est la variable résultat (la VD) et *X* est la variable prédictive (la VI), alors la formule qui décrit notre régression s'écrit comme suit

$$\hat{Y}_i = b_0 + b_1 X_i$$

Bien. Ça ressemble à la même formule, mais il y a quelques petits morceaux de plus dans cette version. Assurons-nous de les comprendre. Tout d'abord, remarquez que j'ai écrit X_i et Y_i plutôt que simplement *X* et *Y*. C'est parce que nous voulons nous rappeler qu'il s'agit de données réelles. Dans cette équation, X_i est la valeur de la variable prédictive pour la i-ième observation (c.-à-d. le nombre d'heures de sommeil que j'ai eues le *premier* jour de ma petite étude) et Y_i est la valeur correspondante de la variable résultat (c.-à-d. ma mauvaise humeur ce jour-là). Et bien que je ne l'ai pas dit aussi explicitement dans l'équation, ce que nous supposons, c'est que cette formule fonctionne pour toutes les observations de l'ensemble des données (c.-à-d. pour tout *i*). Deuxièmement, remarquez que j'ai écrit $\hat{Y}_i$ et non Y_i. C'est parce que nous voulons faire la distinction entre les *données réelles* Y_i et l'*estimation* $\hat{Y}_i$ (c.-à-d. la prévision que fait notre ligne de régression). Troisièmement, j'ai changé les lettres utilisées pour décrire les coefficients de *a* et *b* à *b0* et *b1*. C'est exactement la façon dont les statisticiens aiment se référer aux coefficients dans un modèle de régression. Je ne sais pas pourquoi ils ont choisi *b*, mais c'est ce qu'ils ont fait. Dans tous les cas, b_0 se réfère toujours au terme d'intersection et b_1 à la pente.

Excellent !. Ensuite, je ne peux m'empêcher de remarquer, qu'il s'agisse de la bonne ou de la mauvaise ligne de régression, que les données ne tombent pas parfaitement sur la ligne. Ou, pour le dire autrement, les données Y_i ne sont pas identiques aux prédictions du modèle de régression $\hat{Y}_i$. Puisque les statisticiens aiment attacher des lettres, des noms et des chiffres à tout, considérons la différence entre la prédiction du modèle et les données réelles *résidu*, et nous l'appellerons ϵ_i. En écriture mathématique, les résidus sont définis comme suit

$$\epsilon_i = Y_i - \hat{Y}_i$$

ce qui signifie que nous pouvons écrire le modèle de régression linéaire complet comme suit

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

---

[97] NdT C'est ce qu'on appelle ordonnée à l'origine en géométrie. Pour faire court ce point sera désigné, comme l'on fait les auteurs par intersction dans le reste de l'ouvrage.

Estimation d'un modèle de régression linéaire

Bien, maintenant redessinons nos graphiques mais cette fois j'ajouterai quelques lignes pour montrer la taille du résidu pour toutes les observations. Lorsque la ligne de régression est bonne, nos résidus (la longueur des lignes noires pleines) semblent tous assez petits, comme le montre la Figure 12-12a, mais lorsque la ligne de régression est mauvaise, les résidus sont beaucoup plus grands, comme vous pouvez le voir à la Figure 12-12b.

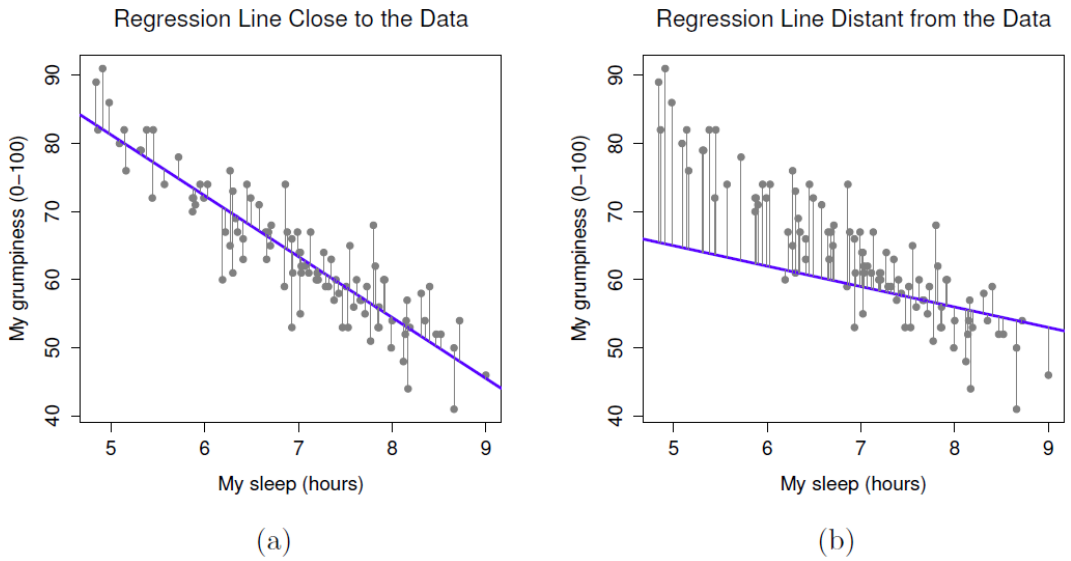


Figure 12-12 : Représentation des résidus associés à la droite de régression la mieux adaptée (figure a) et des résidus associés à une droite de régression faible (figure b). Les résidus sont beaucoup plus petits pour la bonne ligne de régression. Encore une fois, ce n'est pas surprenant étant donné que la bonne ligne est celle qui passe en plein milieu des données.

Bon. Peut-être que ce que nous « voulons » dans un modèle de régression, ce sont de *petits* résidus. Oui, cela semble logique. En fait, je pense aller jusqu'à dire que la ligne de régression « la mieux adaptée » est celle qui a le moins de résidus. Ou, mieux encore, puisque les statisticiens semblent aimer prendre des carrés de tout, pourquoi ne pas le dire :

Les coefficients de régression estimés, $\hat{b}_0$ et $\hat{b}_1$, sont ceux qui minimisent la somme des résidus au carré, que l'on peut écrire soit $\sum_i \left(Y_i - \hat{Y}_i\right)^2$, soit $\sum_i \epsilon_i^2$.

Bien, ça sonne encore mieux. Et comme je l'ai détaillé comme ça, ça veut probablement dire que c'est la bonne réponse. Et comme c'est la bonne réponse, il vaut probablement la peine de noter que nos coefficients de régression sont des *estimations* (nous essayons de deviner les paramètres qui décrivent une population !), c'est pourquoi j'ai ajouté les petits chapeaux, pour obtenir $\hat{b}_0$ et $\hat{b}_1$ plutôt que b_0 et b_1. Enfin, comme il existe en fait plus d'une

façon d'estimer un modèle de régression, le nom plus technique de ce processus d'estimation est **régression des moindres carrés ordinaires (MCO)**.

A ce stade, nous avons maintenant une définition concrète de ce qui constitue notre « meilleur » choix de coefficients de régression, $\hat{b}_0$ et $\hat{b}_1$. La question naturelle à se poser ensuite est la suivante : si nos coefficients de régression optimaux sont ceux qui minimisent la somme des résidus au carré, comment pouvons-nous *trouver* ces merveilleux nombres ? La réponse à cette question est compliquée et ne vous aide pas à comprendre la logique de la régression.[98] Cette fois, je vais vous laisser tranquille. Au lieu de vous montrer le chemin long et fastidieux d'abord et de « révéler » ensuite le merveilleux raccourci que fournit Jamovi, allons droit au but et utilisons juste Jamovi pour faire tout le travail lourd.

Régression linéaire dans Jamovi

Pour exécuter ma régression linéaire, ouvrez l'analyse « Regression » - « Linear Regression » dans Jamovi, en utilisant le fichier de données parenthood.csv. Spécifiez ensuite dan.grump comme variable dépendante et dan.sleep comme variable entrée dans la case « Covariables ». Ceci donne les résultats illustrés à la Figure 12.13, montrant une interception $\hat{b}_0 = 125{,}96$ et la pente $\hat{b}_1 = -8{,}94$. En d'autres termes, la ligne de régression la mieux adaptée que j'ai tracée à la figure 12.11 a cette formule :

$$\hat{Y}_i = 125{,}96 + (-8{,}94 X_i)$$

Interprétation du modèle estimé

La chose la plus importante à comprendre est de savoir comment interpréter ces coefficients. Commençons par $\hat{b}_1$, la pente. Si l'on se souvient de la définition de la pente, un coefficient de régression de $\hat{b}_1 = -8{,}94$ signifie que si j'augmente X_i de 1, alors je diminue Y_i de 8,94. C'est-à-dire, chaque heure supplémentaire de sommeil que je améliorer mon humeur en réduisant ma grinchiosité de 8,94 points. Et l'intersection ? Eh bien, puisque $\hat{b}_0$ correspond à « la valeur attendue de Y_i quand X_i est égal à 0 », c'est assez simple. Cela

---

[98] Ou du moins, je suppose que ça n'aide pas la plupart des gens. Mais au cas où quelqu'un lisant ceci est un véritable maître de kung-fu de l'algèbre linéaire (et pour être juste, j'ai toujours quelques-uns de ces gens dans ma classe d'introduction aux stats), cela *vous* aidera *à* savoir que la solution au problème d'estimation se révèle être $\hat{b} = (X'X)^{-1}X'y$,où $\hat{b}$ est un vecteur contenant les coefficients estimés, X est la « matrice du plan » qui contient les variables de prévision (plus une colonne supplémentaire contenant toutes celles ; strictement X est une matrice des variables explicatives, mais je n'ai pas encore discuté de la distinction), et *y* est un vecteur contenant la variable résultat. Pour tout le monde, ce n'est pas vraiment utile et peut être carrément effrayant. Cependant, puisque beaucoup de choses en régression linéaire peuvent être écrites en termes d'algèbre linéaire, vous verrez un tas de notes de bas de page comme celle-ci dans ce chapitre. Si vous pouvez suivre ces maths, c'est génial. Si ce n'est pas le cas, ignorez-les.

implique que si j'obtiens zéro heure de sommeil ($X_i=0$) alors ma grinchiosité va monter en flèche, à une valeur folle de ($Y_i=125,96$). Il vaut mieux l'éviter, je crois.

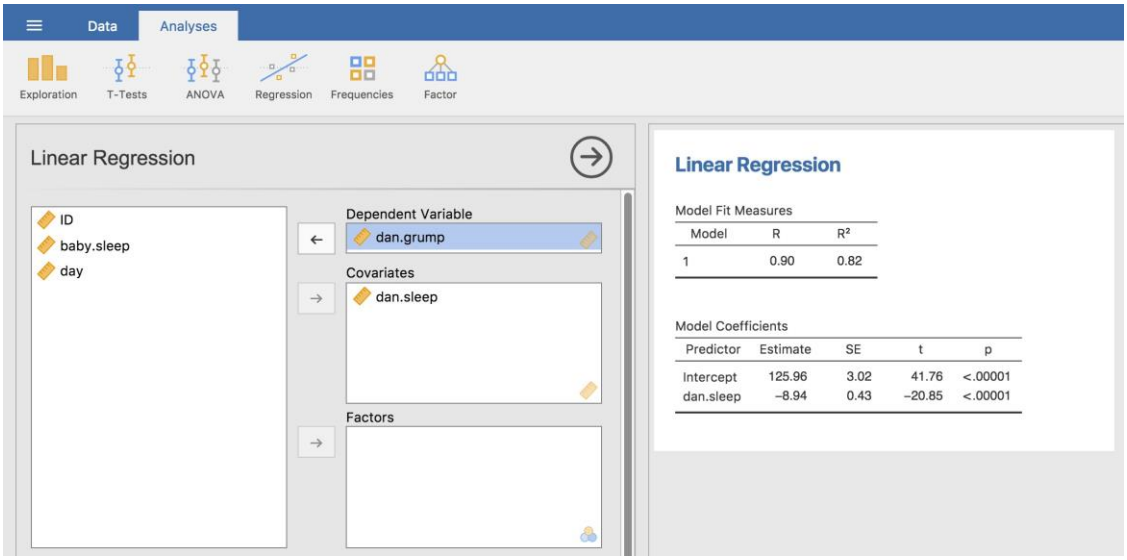


Figure 12-13 : Une capture d'écran de Jamovi montrant une analyse de régression linéaire simple.

Régression linéaire multiple

Le modèle de régression linéaire simple dont nous avons discuté jusqu'à présent suppose qu'il y a une seule variable prédictive qui vous intéresse, en l'occurrence dan.sleep. En fait, jusqu'à maintenant, *tous les* outils statistiques dont nous avons parlé ont supposé que votre analyse utilise une variable prédictive et une variable résultat. Cependant, dans de nombreux projets de recherche (peut-être la plupart), vous avez en fait de multiples prédicteurs que vous voulez examiner. Dans l'affirmative, il serait bon de pouvoir étendre le cadre de régression linéaire pour pouvoir inclure de multiples prédicteurs. Peut-être un modèle de **régression multiple** serait-il nécessaire ?

La régression multiple est conceptuellement très simple. Tout ce que nous faisons, c'est d'ajouter d'autres termes à notre équation de régression. Supposons que nous ayons deux variables qui nous intéressent ; peut-être voulons-nous utiliser dan.sleep et baby.sleep pour prédire la variable dan.grump. Comme avant, nous laissons Y_i se représenter à ma grinchiosité le *i-ème* jour. Mais maintenant nous avons deux variables X : la première correspondant à la quantité de sommeil que j'ai eu et la seconde correspondant à la quantité de sommeil que mon fils a eu. Nous avons donc X_{i1} qui correspond aux heures pendant lesquelles j'ai dormi le *i-ème* jour et X_{i2} aux heures pendant lesquelles le bébé a dormi ce jour-là. Si c'est le cas, alors nous pouvons écrire notre modèle de régression comme ceci :

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \epsilon_i$$

Comme précédemment, ϵ_i est le résidu associé à la *i-ème* observation, $\epsilon_i = Y_i - \hat{Y}_i$. Dans ce modèle, nous avons maintenant trois coefficients à estimer : b_0 est l'interception, b_1 est le coefficient associé à mon sommeil et b_2 est le coefficient associé au sommeil de mon fils. Cependant, bien que le nombre de coefficients à estimer ait changé, l'idée de base du fonctionnement de l'estimation reste inchangée : nos coefficients estimés $\hat{b}_0$, $\hat{b}_1$et $\hat{b}_2$ sont ceux qui minimisent la somme au carré des résidus.

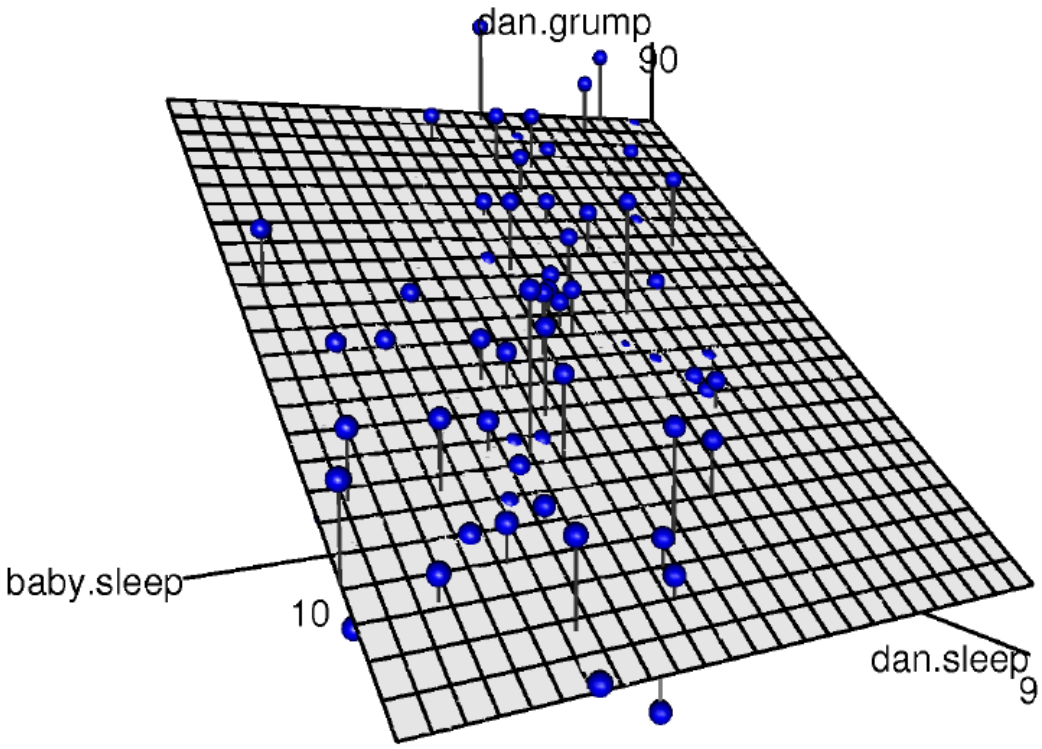


Figure 12-14 : Visualisation 3D d'un modèle de régression multiple. Il y a deux prédicteurs dans le modèle, dan.sleep et baby.sleep, et la variable de résultat est dan.grump. Ensemble, ces trois variables forment un espace 3D. Chaque observation (point) est un point dans cet espace. De la même manière qu'un modèle de régression linéaire simple forme une ligne dans un espace 2D, ce modèle de régression multiple forme un plan dans un espace 3D. Lorsque nous estimons les coefficients de régression, nous essayons de trouver un plan aussi proche que possible de tous les points bleus.

Le faire avec Jamovi

La régression multiple dans Jamovi n'est pas différente de la régression simple. Tout ce que nous avons à faire est d'ajouter des variables supplémentaires à la boîte « Covariables » dans Jamovi. Par exemple, si nous voulons utiliser dan.sleep et baby.sleep comme prédicteurs dans notre tentative d'expliquer pourquoi je suis si grincheux, alors déplacez baby.sleep dans la case « Covariables » à côté de dan.sleep. Par défaut, Jamovi suppose que

le modèle doit inclure une intersection. Les coefficients que nous obtenons cette fois-ci sont :

(Intersection)	dan.sleep	baby.sleep
125.97	-8.95	.01

Le coefficient associé à dan.sleep est assez élevé, ce qui suggère que chaque heure de sommeil que je perds me rend beaucoup plus grincheux. Cependant, le coefficient pour baby.sleep est très faible, ce qui suggère que le temps de sommeil de mon fils n'a pas vraiment d'importance. Ce qui compte pour moi, c'est de savoir combien de temps je dors. Pour avoir une idée de ce à quoi ressemble ce modèle de régression multiple, la Figure 12-14 montre un graphique 3D qui trace les trois variables, ainsi que le modèle de régression lui-même.

Formule pour le cas général

L'équation que j'ai donnée ci-dessus vous montre à quoi ressemble un modèle de régression multiple lorsque vous incluez deux prédicteurs. Il n'est donc pas surprenant que si vous voulez plus de deux prédicteurs, il vous suffit d'ajouter plus de termes X et plus de coefficients b. En d'autres termes, si vous avez K prédicteur dans le modèle, l'équation de régression se présente comme suit

$$Y_i = b_0 + \left(\sum_{k=1}^{k} b_k X_{\text{ik}} \right) + \epsilon_i$$

Quantifier l'ajustement du modèle de régression

Nous savons donc maintenant comment estimer les coefficients d'un modèle de régression linéaire. Le problème est que nous ne savons pas encore si ce modèle de régression est bon. Par exemple, le modèle de régression.1 prétend que chaque heure de sommeil améliorera considérablement mon humeur, mais il se peut que ce soit de la foutaise. Rappelez-vous que le modèle de régression ne produit qu'une prédiction $\hat{Y}_i$ sur mon humeur, mais mon humeur réelle est Y_i. Si ces deux éléments sont très proches, le modèle de régression marche bien. S'ils sont très différents, c'est qu'il ne marche pas.

La valeur R^2

Encore une fois, enveloppons d'un peu de mathématiques cela. Tout d'abord, nous avons la somme des résidus au carré dont nous espérons qu'il sera plutôt petit.

$$\text{SS}_{\text{res}} = \sum_{i} \left(Y_i - \hat{Y}_i \right)^2$$

Plus précisément, ce que nous aimerions, c'est qu'il soit très faible par rapport à la variabilité totale de la variable résultats

$$SS_{tot} = \sum_i (Y_i - \overline{Y})^2$$

Pendant que nous y sommes, calculons ces valeurs, mais pas à la main. Utilisons quelque chose comme Excel ou un autre tableur standard. J'ai fait ceci en ouvrant le fichier parenthood.csv dans Excel et en l'enregistrant comme parenthood rsquared.xls pour que je puisse travailler dessus. La première chose à faire est de calculer les valeurs de $\hat{Y}$, et pour le modèle simple qui n'utilise qu'un seul prédicteur, nous ferions ce qui suit :

1. créer une nouvelle colonne appelée « Y.pred » en utilisant la formule « = 125,97 + (-8,94 * dan.sleep) ».

Bien, maintenant que nous avons une variable qui stocke les prédictions du modèle de régression afin de savoir à quel point je serai grincheux un jour donné, calculons notre somme des résidus au carré. Pour ce faire, nous utiliserions la formule suivante :

1. calculer la SSres en créant une nouvelle colonne appelée « (Y-Y.pred)^2 » en utilisant la formule « = (dan.grump - Y.pred)^2 ».

2. Ensuite, au bas de cette colonne, calculez la somme de ces valeurs, i.e.'sum(((((Y-Y.pred)^2)).

Ceci devrait vous donner une valeur de « 1838.722 ». Merveilleux. Un gros chiffre qui ne veut pas dire grand-chose. Néanmoins, allons quand même de l'avant avec audace et calculons aussi la somme totale des carrés. C'est aussi assez simple. Calculez la SS(tot) par :

4. Au bas de la colonne dan.grump, calculez la valeur moyenne pour dan.grump (NB Excel utilise le mot « MOYENNE » dans sa fonction).

5. Créez ensuite une nouvelle colonne, appelée « (Y - mean(Y))^2) » en utilisant la formule « = (dan.grump - AVERAGE(dan.grump))^2 ».

6. Ensuite, au bas de cette colonne, calculez la somme de ces valeurs, C'est-à-dire « somme((Y - moyenne(Y))^2) ».

Ceci devrait vous donner une valeur de « 9998.59 ». Bien, il s'agit d'un nombre beaucoup plus élevé que le précédent, ce qui donne à penser que notre modèle de régression fait de bonnes prédictions. Mais ce n'est pas très interprétable.

On peut peut-être arranger ça. Ce que nous aimerions faire, c'est convertir ces deux chiffres plutôt insignifiants en un seul chiffre. Un joli chiffre interprétable, que nous appellerons R^2 sans raison particulière. Ce que nous aimerions, c'est que la valeur de R^2 soit égale à 1 si le modèle de régression ne fait aucune erreur de prédiction des données. En d'autres termes, s'il s'avère que les erreurs résiduelles sont nulles. De même, si le modèle est complètement inutile, nous voudrions que R^2 soit égal à 0. Qu'est-ce que j'entends par « inutile » ? Aussi tentant que cela puisse paraître d'exiger que le modèle de régression quitte la maison, se coupe les cheveux et trouve un vrai emploi, je vais probablement devoir choisir une définition plus pratique. Dans ce cas, tout ce que je veux dire, c'est que la somme résiduelle des carrés n'est pas inférieure à la somme totale des carrés, *SSres = SStot*. Attendez,

pourquoi ne pas faire exactement ça ? La formule qui nous fournit notre valeur R^2 est assez simple à écrire,

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

et tout aussi simple à calculer dans Excel :

7. Calculez R.au carré en tapant dans une case vide ce qui suit : « =1 - (SS(resid) / SS(tot)) ».

Ceci donne une valeur pour R^2 de 0,8161018. La valeur R^2, parfois appelée **coefficient de détermination,**[99] a une interprétation simple : c'est la *proportion* de la variance de la variable résultat qui peut être prise en compte par le prédicteur. Donc, dans ce cas, le fait que nous ayons obtenu $R^2=0,816$ signifie que le prédicteur (my.sleep) explique 81,6% de la variance du résultat (my.grump).

Naturellement, vous n'avez pas besoin de taper toutes ces commandes vous-même dans Excel si vous voulez obtenir la valeur R^2 pour votre modèle de régression. Comme nous le verrons plus loin dans Section 12.7.3, tout ce que vous avez à faire est de le spécifier comme option dans Jamovi. Cependant, mettons cela de côté pour l'instant. Il y a une autre propriété de R^2 que je tiens à souligner.

La relation entre régression et corrélation

À ce stade, nous pouvons revenir sur mon affirmation antérieure selon laquelle la régression, sous cette forme très simple dont j'ai parlé jusqu'ici, est essentiellement la même chose qu'une corrélation. Auparavant, nous utilisions le symbole *r* pour désigner une corrélation de Pearson. Pourrait-il y avoir une relation entre la valeur du coefficient de corrélation *r* et la valeur R^2 de la régression linéaire ? Bien sûr qu'il y en a : la corrélation au carré r^2 est identique à la valeur R^2 pour une régression linéaire avec un seul prédicteur. En d'autres termes, l'exécution d'une corrélation de Pearson équivaut plus ou moins à l'exécution d'un modèle de régression linéaire qui utilise une seule variable prédictive.

La valeur R^2 ajustée

Une dernière chose à souligner avant de passer à autre chose. Il est assez courant pour les gens de rapporter une mesure légèrement différente de la performance du modèle, connue sous le nom de « R^2 ajusté ». La motivation qui sous-tend le calcul de la valeur corrigée de R^2 est l'observation que l'ajout d'autres variables prédictrices dans le modèle entraînera *toujours* une augmentation (ou au moins pas de diminution) de la valeur de R^2.

---

[99] Et par « parfois » je veux dire « presque jamais ». Dans la pratique, tout le monde l'appelle simplement « *R-carré* ».

La valeur R^2 ajustée introduit une légère modification dans le calcul, comme suit. Pour un modèle de régression avec K prédicteurs, ajusté à un ensemble de données contenant N observations, le R^2 ajusté est :

$$\text{adj. } R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} \times \frac{N-1}{N-K-1}$$

Cet ajustement est une tentative de prise en compte des degrés de liberté. Le grand avantage de la valeur corrigée de R^2 est que lorsque vous ajoutez plus de prédicteurs au modèle, la valeur corrigée de R^2 n'augmentera que si les nouvelles variables améliorent la performance du modèle plus que vous ne l'attendriez par hasard. Le grand inconvénient est que la valeur R^2 ajustée *ne peut pas* être interprétée de la même manière élégante que R^2. R^2 a une interprétation simple, c'est la proportion de variance dans la variable résultats qui s'explique par le modèle de régression. A ma connaissance, il n'existe pas d'interprétation équivalente pour le R^2 ajusté.

Une question évidente est donc de savoir si vous devez déclarer R^2 ou R^2 ajusté. C'est probablement une question de préférence personnelle. Si vous vous souciez davantage de l'interprétabilité, alors R^2 est meilleur. Si vous vous souciez davantage de corriger les biais, alors R^2 ajusté est probablement mieux. Pour ma part, je préfère R^2. J'ai l'impression qu'il est plus important d'être capable d'interpréter la mesure de la performance de votre modèle. De plus, comme nous le verrons à la section 12.7, si vous craignez que l'amélioration de R^2 que vous obtenez en ajoutant un prédicteur ne soit due qu'au hasard et non à un meilleur modèle, nous avons des tests d'hypothèse pour cela.

Tests d'hypothèse pour les modèles de régression

Jusqu'à présent, nous avons parlé de ce qu'est un modèle de régression, de la façon dont les coefficients d'un modèle de régression sont estimés et de la façon dont nous quantifions le rendement du modèle (le dernier de ces éléments, soit dit en passant, est essentiellement notre mesure de la valeur de l'effet). La prochaine chose dont nous devons parler, c'est des tests d'hypothèse. Il y a deux types de tests d'hypothèse différents (mais apparentés) dont nous devons parler : ceux dans lesquels nous vérifions si le modèle de régression dans son ensemble donne de bien meilleurs résultats qu'un modèle basé sur l'hypothèse nulle, et ceux dans lesquels nous vérifions si un coefficient de régression particulier est sensiblement différent de zéro.

Tester le modèle dans son ensemble

Supposons que vous ayez estimé votre modèle de régression. Le premier test d'hypothèse que vous pouvez essayer est l'hypothèse nulle qu'il n'y a *pas de relation* entre les prédicteurs et le résultat, et l'hypothèse alternative que *les données sont distribuées exactement comme le modèle de régression le prévoit.*

Formellement, notre « modèle correspondant à l'hypothèse nulle » correspond au modèle de « régression » assez trivial dans lequel nous incluons 0 prédicteurs et n'incluons que le terme d'intersection b_0 :

$$H_0: Y_i = b_0 + \epsilon_i$$

Si notre modèle de régression a des *k* prédicteurs, le « modèle alternatif » est décrit à l'aide de la formule habituelle pour un modèle de régression multiple :

$$H_1: Y_i = b_0 + \left(\sum_{k=1}^{K} b_k X_{\text{ik}} \right) + \epsilon_i$$

Comment tester ces deux hypothèses l'une par rapport à l'autre ? L'astuce est de comprendre qu'il est possible de diviser la variance totale SS_{tot} en la somme de la variance résiduelle SS_{res} et la variance du modèle de régression SS_{mod}. Je vais sauter les détails techniques, puisque nous y reviendrons plus tard lorsque nous examinerons l'analyse de variance au chapitre 13. Mais notez juste que

$$\text{SS}_{\text{mod}} = \text{textSS}_{tot} - \text{SS}_{\text{res}}$$

Et nous pouvons convertir les sommes des carrés en carrés moyens en divisant par les degrés de liberté.

$$\text{MS}_{\text{mod}} = \frac{\text{SS}_{\text{mod}}}{\text{ddl}_{\text{mod}}}$$

$$\text{MS}_{\text{res}} = \frac{\text{SS}_{\text{res}}}{\text{ddl}_{\text{res}}}$$

Alors, combien de degrés de liberté avons-nous ? Comme vous pouvez vous y attendre, le *df* associé au modèle est étroitement lié au nombre de prédicteurs que nous avons inclus. En fait, il s'avère que $df_{mod}=K$. Pour les résidus, le degré total des degrés de liberté est $df_{res}=N–K-1$.

Maintenant que nous avons nos carrés moyens, nous pouvons calculer une statistique *F* comme ceci :

$$F = \frac{\text{MS}_{\text{mod}}}{\text{MS}_{\text{res}}}$$

Nous verrons beaucoup plus de statistiques *F* au chapitre 13, mais pour l'instant, sachez simplement que nous pouvons interpréter de grandes valeurs *F* comme indiquant que l'hypothèse nulle fonctionne mal par rapport à l'hypothèse alternative. Dans un instant, je vais vous montrer comment faire le test avec Jamovi de la manière la plus simple, mais voyons d'abord les tests pour les coefficients de régression individuels.

Essais pour les coefficients individuels

Le *test F* que nous venons de présenter est utile pour vérifier que le modèle dans son ensemble fonctionne mieux que le hasard. Si votre modèle de régression ne produit pas un résultat significatif pour le test *F*, vous n'avez probablement pas un très bon modèle de régression (ou, très probablement, vous n'avez pas de très bonnes données). Cependant, bien que l'échec à ce test soit un indicateur assez fort que le modèle a des problèmes, le fait

de réussir le test (c.-à-d. de rejeter l'hypothèse nulle) ne signifie pas que le modèle est bon ! Vous vous demandez peut-être pourquoi ? La réponse à cette question peut être trouvée en examinant les coefficients du modèle de régression multiple que nous avons déjà regardés à la section 12.5 ci-dessus, où nous avons obtenu les coefficients :

(Intersection)	dan.sleep	baby.sleep
125,97	-8,950	0,01

Je ne peux m'empêcher de remarquer que le coefficient de régression estimé pour la variable baby.sleep est minuscule (0,01), par rapport à la valeur que nous obtenons pour dan.sleep (-8,95). Étant donné que ces deux variables sont absolument sur la même échelle (elles sont toutes deux mesurées en « heures de sommeil »), je trouve cela éclairant. En fait, je commence à soupçonner que c'est seulement la quantité de sommeil que j'obtiens qui compte pour prédire ma grinchiosité.

Nous pouvons réutiliser un test d'hypothèse dont nous avons parlé plus tôt, le *test t*. Le test qui nous intéresse a pour hypothèse nulle que le vrai coefficient de régression est zéro ($b = 0$), qui doit être testé contre l'hypothèse alternative qu'il n'est pas ($b \neq 0$). C'est-à-dire :

$$ \text{H}_{0} : b=0\\ \text{H}_{1} : b\neq0 $$

Comment pouvons-nous le tester ? Eh bien, si le théorème central limite est correct, nous pouvons deviner que la distribution d'échantillonnage de $\hat{b}$, le coefficient de régression estimé, est une distribution normale avec une moyenne centrée sur b. Cela signifie est que si l'hypothèse nulle était vraie, alors la distribution d'échantillonnage de $\hat{b}$ a une moyenne nulle et un écart type inconnu. En supposant que nous puissions obtenir une bonne estimation de l'erreur type du coefficient de régression, $\text{SE}\left(\hat{b}\right)$, alors nous avons de la chance. C'est *exactement* la situation pour laquelle nous avons introduit le *test t* sur un échantillon au chapitre 11. Définissons donc une statistique t comme ceci

$$t = \frac{\hat{b}}{\text{SE}\left(\hat{b}\right)}$$

Je vais sauter les justifications, mais nos degrés de liberté dans ce cas sont *df=N-K-1*.

Il est irritant de constater que l'estimation de l'erreur type du coefficient de régression, $\text{SE}\left(\hat{b}\right)$, n'est pas aussi facile à calculer que l'erreur type de la moyenne que nous avons utilisée pour les tests t plus simples au chapitre 11. En fait, la formule est quelque peu laide et n'est pas très utile à regarder.[100] Pour nos besoins, il suffit de souligner que l'erreur type

---

[100] Pour les lecteurs avancés seulement. Le vecteur des résidus est $\epsilon = y - X\,\hat{b}$. Pour K prédicteurs plus l'intersection, la variance résiduelle estimée est $\hat{\sigma}^2 = \epsilon'\epsilon/(N - K - 1)$. La

du coefficient de régression estimé dépend à la fois du prédicteur et des variables de résultat, et qu'elle est quelque peu sensible aux violations de l'hypothèse d'homogénéité de la variance (examinée plus loin).

En tout état de cause, cette statistique *t* peut être interprétée de la même manière que les statistiques *t* dont nous avons parlé au chapitre 11. En supposant que vous avez une hypothèse alternative bilatérale (c'est-à-dire que vous ne vous souciez pas vraiment de savoir si $b < 0$ ou $b > 0$), alors ce sont les valeurs extrêmes de *t* (c'est-à-dire beaucoup moins que zéro ou beaucoup plus que zéro) qui suggèrent que vous devez rejeter l'hypothèse nulle.

Exécuter les tests d'hypothèse dans Jamovi

Pour calculer toutes les statistiques dont nous avons parlé jusqu'à présent, tout ce que vous avez à faire est de vous assurer que les options pertinentes sont cochées dans Jamovi et ensuite exécuter la régression. Si nous faisons cela, comme le montre la Figure 12-15, nous obtenons toute une série de résultats utiles.

Les « coefficients du modèle » au bas des résultats de l'analyse de Jamovi présentés à la Figure 12-15 fournissent les coefficients du modèle de régression. Chaque ligne de ce tableau fait référence à l'un des coefficients du modèle de régression. La première ligne est le terme d'intersection, et les dernières lignes regardent chacun des prédicteurs. Les colonnes vous donnent toutes les informations pertinentes. La première colonne est l'estimation réelle de *b* (p. ex. 125,97 pour l'intersection et -8,95 pour le prédicteur dan.sleep). La deuxième colonne est l'estimation de l'erreur type $\hat{\sigma}_b$. Les troisième et quatrième colonnes fournissent les valeurs inférieures et supérieures pour l'intervalle de confiance à 95 % autour de l'estimation *b* (nous y reviendrons plus loin).

---

matrice de covariance estimée des coefficients est $\hat{\sigma}^2 \ (X'X)^{-1}$, dont la diagonale principale est SE $\left(\hat{b}\right)$, notre erreur type estimée.

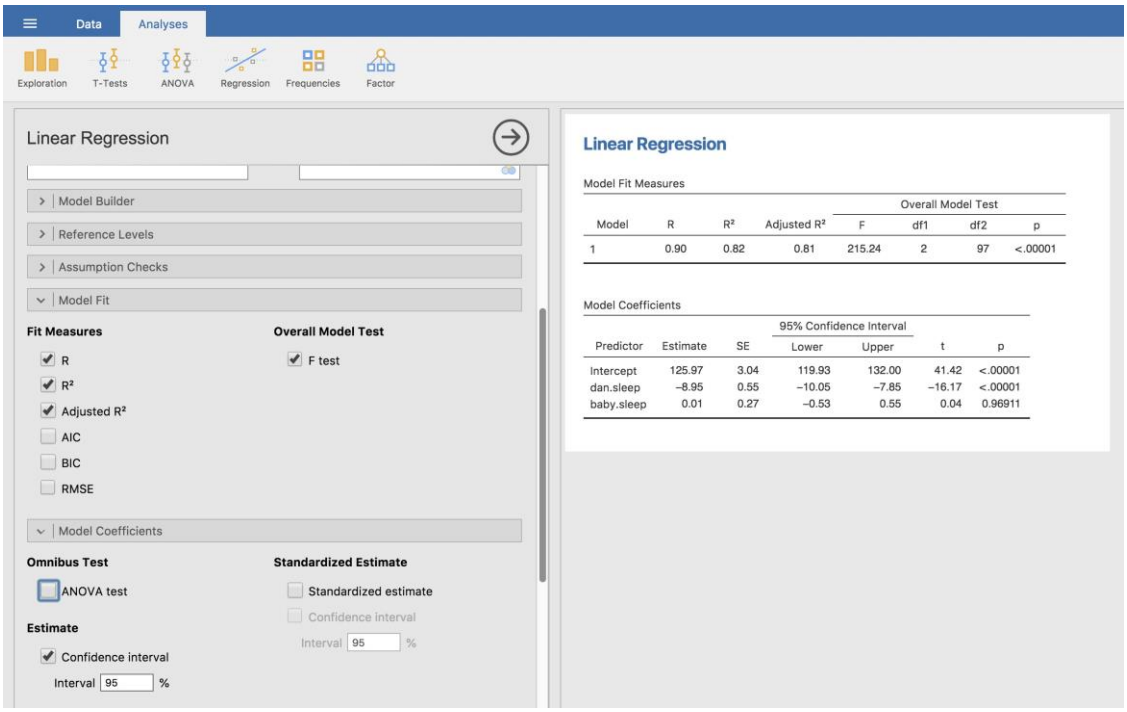


Figure 12-15 : Une capture d'écran de Jamovi montrant une analyse de régression linéaire multiple, avec quelques options utiles cochées.

La cinquième colonne vous donne la statistique *t*, et il est intéressant de noter que dans ce tableau $t = \hat{b} / SE\left(\hat{b}\right)$ à chaque fois. Enfin, la dernière colonne vous donne la *valeur p* réelle de chacun de ces tests.[101]

La seule chose que le tableau des coefficients lui-même n'énumère pas est le degré de liberté utilisé dans le *test t*, qui est toujours *N -K-1* et qui est énuméré dans le tableau en haut de la sortie, intitulé « Model Fit Measures ». On peut voir dans ce tableau que le modèle est bien plus performant que ce à quoi on pourrait s'attendre par hasard ($F_{(2,97)}=215,24$, $p<.001$), ce qui n'est pas si surprenant : le $R^2 =0,81$ indiquent que le modèle de régression représente 81 % de la variabilité de la mesure des résultats (et 82 % pour le R^2 ajusté). Cependant, lorsque nous examinons les tests *t* pour chacun des coefficients individuels, nous avons des preuves assez solides que la variable baby.sleep n'a pas d'effet significatif. Tout le travail dans ce modèle est effectué par la variable dan.sleep. Pris ensemble, ces résultats suggèrent que ce modèle de régression n'est pas le bon modèle pour les données. Vous feriez probablement mieux de laisser tomber le prédicteur de baby.sleep. En d'autres

[101] Notez que, bien que Jamovi a fait plusieurs tests ici, il n'a pas fait de correction de Bonferroni ou autre. Il s'agit de tests *t* standard à un échantillon avec une alternative bilatérale. Si vous voulez faire des corrections pour plusieurs tests, vous devez le faire vous-même.

termes, le modèle de régression simple avec lequel nous avons commencé est le meilleur modèle.

A propos des coefficients de régression

Avant de discuter des hypothèses qui sous-tendent la régression linéaire et de ce que vous pouvez faire pour vérifier si elles sont respectées, j'aimerais aborder brièvement deux autres sujets, qui ont tous deux trait aux coefficients de régression. La première chose dont il faut parler est le calcul des intervalles de confiance pour les coefficients. Ensuite, j'aborderai la question quelque peu floue de savoir comment déterminer quel prédicteur est le plus important.

Intervalles de confiance pour les coefficients

Comme tout paramètre de population, les coefficients de régression b ne peuvent être estimés avec une précision complète à partir d'un échantillon de données ; c'est en partie pourquoi nous avons besoin de tests d'hypothèse. Par conséquent, il est très utile de pouvoir rapporter des intervalles de confiance qui expriment notre incertitude au sujet de la valeur réelle de b. Cela est particulièrement utile lorsque la question de recherche est fortement axée sur une tentative d'établir la relation entre la variable X et la variable Y, puisque, dans ces situations, l'intérêt se porte principalement sur le coefficient de régression b.

Heureusement, les intervalles de confiance pour les coefficients de régression peuvent être construits de la manière habituelle

$$\mathrm{CI}(b) = \hat{b} \pm \left(t_{\mathrm{crit}} \times SE(\hat{b}) \right)$$

où $SE(\hat{b})$ est l'erreur type du coefficient de régression et t_{crit} est la valeur critique pertinente de la distribution t appropriée. Par exemple, si c'est un intervalle de confiance à 95 % que nous voulons, alors la valeur critique est le quantile 97,5 d'une distribution t avec $N–K-1$ degrés de liberté. En d'autres termes, il s'agit essentiellement de la même méthode de calcul des intervalles de confiance que celle que nous avons utilisée jusqu'ici.

Dans Jamovi, nous avions déjà spécifié l'intervalle de confiance à 95 %, comme le montre la Figure 12-15, bien que nous aurions pu facilement choisir une autre valeur, par exemple un intervalle de confiance à 99 %, si nous l'avion voulu.

Calcul des coefficients de régression standardisés

Vous voudrez peut-être également calculer des coefficients de régression « standardisés », souvent désignés par β. La logique derrière les coefficients standardisés est la suivante. Dans de nombreuses situations, vos variables se situent sur des échelles fondamentalement différentes. Supposons, par exemple, que mon modèle de régression vise à prédire le quotient intellectuel des gens en utilisant leur niveau de scolarité (nombre d'années de scolarité) et leur revenu comme prédicteurs. De toute évidence, le niveau d'instruction et le revenu ne sont pas sur la même échelle. Le nombre d'années de scolarité ne peut varier que de 10 d'années, alors que le revenu peut varier de 10 000 dollars (ou plus). Les unités de

mesure ont une grande influence sur les coefficients de régression. Les coefficients *b* n'ont de sens que lorsqu'ils sont interprétés à la lumière des unités, tant pour les prédicteurs que pour la variable résultat. Il est donc très difficile de comparer les coefficients des différents prédicteurs. Pourtant, il y a des situations où l'on souhaite vraiment faire des comparaisons entre différents coefficients. Plus précisément, vous pouvez vouloir une sorte de mesure standard pour déterminer quels prédicteurs ont la relation la plus forte avec la variable résultat. C'est ce que les **coefficients standardisés** visent à faire.

L'idée de base est assez simple ; les coefficients standardisés sont les coefficients que vous auriez obtenus si vous aviez converti toutes les variables en *z-scores* avant de calculer la régression[102]. L'idée ici est qu'en convertissant tous les prédicteurs en *z-scores*, ils entrent tous dans la régression sur la même échelle, ce qui élimine le problème d'avoir des variables sur différentes échelles. Quelle que soit la variable initiale, une valeur β de 1 signifie qu'une augmentation du prédicteur d'un écart-type produira une augmentation correspondante d'un écart-type sur la variable résultat. Par conséquent, si la variable A a une valeur absolue de β supérieure à celle de la variable B, on considère qu'elle a une relation plus forte avec le résultat. C'est l'idée du moins. Il faut cependant être un peu prudent ici, car cela repose en grande partie sur l'hypothèse selon laquelle « un changement d'écart type de 1 » est fondamentalement le même genre de chose pour toutes les variables. Il n'est pas toujours évident que ce soit vrai.

Si l'on laisse de côté les questions d'interprétation, examinons la façon dont elle est calculée. Ce que vous pourriez faire, c'est standardiser toutes les variables vous-même, puis effectuer une régression, mais il y a un moyen beaucoup plus simple de le faire. Il s'avère que le coefficient β pour un prédicteur X et un résultat Y a une formule très simple, à savoir

$$\beta_X = b_X \times \frac{\sigma_X}{\sigma_Y}$$

où σ_X est l'écart-type du prédicteur et σ_Y est l'écart-type de la variable résultat Y. Cela rend les choses beaucoup plus simples.

Pour rendre les choses encore plus simples, Jamovi a une option qui calcule les coefficients de β pour vous en utilisant la case à cocher « Standardized estimate » dans les options « Model Coefficients », voir les résultats à la Figure 12-16.

Ceci montre clairement que la variable dan.sleep a un effet beaucoup plus fort que la variable baby.sleep. Cependant, c'est l'exemple parfait d'une situation où il serait probablement plus judicieux d'utiliser les coefficients originaux *b* plutôt que les coefficients standardisés β. Après tout, mon sommeil et le sommeil du bébé sont *déjà* sur la même

---

[102] Strictement, vous normalisez tous les *variables explicatives*. C'est-à-dire, chaque "chose" à laquelle est associé un coefficient de régression dans le modèle. Pour les modèles de régression dont j'ai parlé jusqu'à présent, chaque variable prédictive est représentée sur un seul régresseur, et vice versa. Cependant, ce n'est pas vraiment vrai en général et nous en verrons quelques exemples au chapitre 14. Mais, pour l'instant, nous n'avons pas besoin de trop nous soucier de cette distinction.

échelle : le nombre d'heures de sommeil. Pourquoi compliquer les choses en les convertissant en *z-scores* ?

Linear Regression

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.90	0.82	0.81	215.24	2	97	< .00001

Model Specific ResultsModel 1

Model 1

Model Coefficients

			95% Confidence Interval					95% Confidence Interval	
Predictor	Estimate	SE	Lower	Upper	t	p	Stand. Estimate	Lower	Upper
Intercept	125.97	3.04	119.93	132.00	41.42	< .00001			
dan.sleep	-8.95	0.55	-10.05	-7.85	-16.17	< .00001	-0.90	-1.02	-0.79
baby.sleep	0.01	0.27	-0.53	0.55	0.04	0.96911	0.00	-0.11	0.11

Figure 12-16 : Coefficients normalisés avec des intervalles de confiance à 95 %, pour la régression linéaire multiple

Hypothèses de régression

Le modèle de régression linéaire dont j'ai parlé repose sur plusieurs hypothèses. Dans la Section 12.10, nous développerons la façon de vérifier ces hypothèses, mais examinons d'abord chacune d'entre elles.

- *Normalité*. Comme beaucoup de modèles statistiques, la régression linéaire simple ou multiple de base repose sur une hypothèse de normalité. Plus précisément, il suppose que les *résidus* sont normalement répartis. En fait, il n'y a pas de problème si les prédicteurs *X* et le résultat *Y* sont anormaux, pourvu que les résidus soient normaux. Voir section 12.10.3.

- *Linéarité*. Une hypothèse assez fondamentale du modèle de régression linéaire est que la relation entre *X* et *Y* est linéaire ! Qu'il s'agisse d'une régression simple ou d'une régression multiple, nous supposons que les relations impliquées sont linéaires.

- *Homogénéité de la variance*. Strictement parlant, le modèle de régression suppose que chaque résidu *i* est généré à partir d'une distribution normale avec une moyenne 0, et (ce qui est plus important pour les besoins actuels) avec un écart type identique pour chaque résidu. Dans la pratique, il est impossible de vérifier l'hypothèse selon laquelle chaque résidu est distribué de façon identique. Au lieu de cela, ce qui nous importe, c'est que l'écart-type du résidu soit le même pour toutes les valeurs de $\hat{Y}$, et (si nous

sommes particulièrement paranoïaques) toutes les valeurs de chaque prédicteur *X* dans le modèle.

- *Prédicteurs non corrélés*. L'idée ici est que, dans un modèle de régression multiple, vous ne voulez pas que vos prédicteurs soient trop fortement corrélés les uns aux autres. « Techniquement », ce n'est pas une hypothèse du modèle de régression, mais dans la pratique, c'est nécessaire. Des prédicteurs trop fortement corrélés les uns aux autres (ce qu'on appelle la « colinéarité ») peuvent poser des problèmes lors de l'évaluation du modèle. Voir la section 12.10.4.

- *Les résidus sont indépendants les uns des autres*. Il ne s'agit en fait que d'une hypothèse « passe-partout », avec la supposition que « il n'y a rien d'autre de bizarre dans les résidus ». S'il se passe quelque chose de bizarre (p. ex., les résidus dépendent tous fortement d'une autre variable non mesurée), cela pourrait tout gâcher.

Pas de valeurs aberrantes. Encore une fois, il ne s'agit pas en fait d'une hypothèse technique du modèle (ou plutôt d'une hypothèse implicite de tous les autres), mais il y a une hypothèse implicite que votre modèle de régression n'est pas trop fortement influencé par une ou deux données aberrantes qui pourraient remettre en cause la pertinence du modèle et la fiabilité des données dans certains cas. Voir section 12.10.2.

Vérification du modèle

L'objectif principal de cette section est la **vérification du modèle de régression**, un terme qui fait référence à l'art de vérifier que les hypothèses de votre modèle de régression ont été respectées, de trouver comment corriger le modèle si les hypothèses ne sont pas satisfaites, et généralement de vérifier qu'il n'y a rien de « bizarre ». C'est ce que j'appelle, à juste titre, « l'art » de la vérification des modèles. Ce n'est pas facile, et bien qu'il y ait beaucoup d'outils assez standardisés que vous puissiez utiliser pour diagnostiquer et peut-être même résoudre les problèmes qui affectent votre modèle (s'il y en a !), vous devez vraiment faire preuve d'un certain jugement en faisant cela. Il est facile de se perdre dans tous les détails de la vérification de telle ou telle caractéristique, et c'est assez épuisant d'essayer de se rappeler ces différentes composantes. Cela a pour effet secondaire très désagréable que beaucoup de gens sont frustrés lorsqu'ils essaient d'apprendre tous ces outils, ce qui les pousse à plutôt de ne pas faire de vérification du modèle. C'est un peu inquiétant !

Dans cette section, je décris plusieurs contrôles différents que vous pouvez faire pour vous assurer que votre modèle de régression fait ce qu'il est censé faire. Cela ne couvre pas tout ce que vous pourriez faire, mais c'est beaucoup plus détaillé que ce que je vois beaucoup de gens faire dans la pratique, et même moi je ne présente généralement pas tout cela dans mon cours d'introduction aux statistiques non plus. Toutefois, je pense qu'il est important que vous sachiez quels outils sont à votre disposition, alors je vais essayer d'en présenter quelques-uns ici. Enfin, je dois noter que cette section s'inspire beaucoup du texte de Fox et Weisberg (2011), le livre associé au package car qui est utilisé pour effectuer l'analyse de régression dans R. Le package car est réputé pour fournir d'excellents outils de diagnostic de régression, et le livre lui-même en parle d'une manière admirablement claire. Je ne veux

pas paraître trop brusque, mais je pense que Fox et al (2011) vaut la peine d'être lu, même si certaines des techniques de diagnostic avancées ne sont disponibles que sous R et non dans Jamovi.

Trois types de résidus

La majorité des diagnostics de régression tournent autour de l'examen des résidus, et maintenant vous avez probablement forgé une théorie assez pessimiste des statistiques pour pouvoir deviner que, précisément parce que nous tenons beaucoup aux résidus, il existe en fait différentes catégories de résidus que nous pouvons considérer. En particulier, les trois types de résidus suivants sont mentionnés dans la présente section : les "résidus ordinaires", les "résidus normalisés" et les "résidus studentisés". Il y en a un quatrième type dont il est question dans certaines des figures, et c'est le résidu de Pearson. Toutefois, pour les modèles dont il est question dans ce chapitre, le résidu de Pearson est identique au résidu ordinaire.

Le premier et le plus simple des résidus dont nous nous soucions sont les **résidus ordinaires**. Ce sont les résidus bruts dont j'ai parlé tout au long de ce chapitre jusqu'à maintenant. La valeur résiduelle ordinaire est juste la différence entre la valeur ajustée $\hat{Y}_i$ et la valeur observée Y_i. J'ai utilisé l'indice *i* pour faire référence au *i-ème* résidu ordinaire, et je vais m'y tenir. En gardant à l'esprit que nous avons l'équation très simple suivante

$$\epsilon_i = Y_i - \hat{Y}_i$$

C'est bien sûr ce que nous avons vu plus tôt, et à moins que je ne fasse spécifiquement référence à un autre type de résidu, c'est de celui-ci dont je parle. Il n'y a donc rien de nouveau ici. Je voulais juste me répéter. L'un des inconvénients de l'utilisation des résidus ordinaires est qu'ils sont toujours sur une échelle différente, selon la variable résultats et la qualité du modèle de régression. Autrement dit, à moins que vous n'ayez décidé d'exécuter un modèle de régression sans intersection, les résidus ordinaires auront une moyenne de 0, mais la variance est différente pour chaque résidus. Dans de nombreux contextes, en particulier lorsque l'on ne s'intéresse qu'à *l'évolution* des résidus et non à leurs valeurs réelles, il est commode d'estimer les **résidus normalisés** de manière à avoir un écart type 1.

La façon dont nous les calculons est de diviser le résidu ordinaire par une estimation de l'écart-type (de la population) de ces résidus. Pour des raisons techniques, la formule est la suivante

$$\epsilon_i' = \frac{\epsilon_i}{\hat{\sigma}\sqrt{1-h_i}}$$

où $\hat{\sigma}$ dans ce contexte est l'écart-type de population estimé des résidus ordinaires, et h_i est la « valeur chapeau » de la i-ième observation. Je ne vous ai pas encore expliqué ces valeurs

(mais n'ayez crainte[103], ça va venir bientôt), donc ça n'aura pas beaucoup de sens. Pour l'instant, il suffit d'interpréter les résidus standardisés comme si nous avions converti les résidus ordinaires en *z-scores*. En fait, c'est plus ou moins la vérité, c'est juste que c'est un peu plus chic.

Le troisième type de résidus sont les **résidus Studentisés** (aussi appelés « résidus en portefeuille[104] ») et ils sont encore plus fantaisistes que les résidus standardisés. Encore une fois, l'idée est de prendre le résidu ordinaire et de le diviser par une certaine quantité afin d'estimer une notion standardisée du résidu.

La formule pour faire les calculs cette fois-ci est subtilement différente

$$\epsilon_i^* = \frac{\epsilon_i}{\hat{\sigma}_{(-1)}\sqrt{1-h_i}}$$

Notez que notre estimation de l'écart-type ici est écrite $\hat{\sigma}_{(-1)}$. Cela correspond à l'estimation de l'écart-type résiduel que *vous auriez obtenu* si vous veniez de supprimer la ième observation de l'ensemble de données. Cela ressemble à un cauchemar à calculer, puisque cela suggère qu'il faut exécuter *N* nouveaux modèles de régression (même un ordinateur moderne pourrait râler pour cela, surtout si vous avez un grand ensemble de données). Heureusement, une personne très intelligente a montré que cette estimation de l'écart-type est en fait donnée par l'équation suivante :

$$\sigma_{(-1)} = \hat{\sigma}\sqrt{\frac{N-K-1-\epsilon_i'^2}{N-K-2}}$$

Ce n'est pas une pépite ?

Avant de poursuivre, je dois souligner que vous n'avez pas souvent besoin d'obtenir vous-même ces résidus, même s'ils sont au cœur de presque tous les diagnostics de régression. La plupart du temps, les différentes options qui fournissent les diagnostics, ou les vérifications des hypothèses, s'occuperont de ces calculs pour vous. Malgré tout, il est toujours agréable de savoir comment s'en emparer au cas où vous auriez besoin de faire quelque chose de non standard.

Trois types de données anormales

Un danger que vous pouvez rencontrer avec les modèles de régression linéaire est que votre analyse pourrait être disproportionnellement sensible à un petit nombre d'observations « inhabituelles » ou « anormales ». J'ai déjà discuté de cette idée à la section 5.2.3 dans le contexte des valeurs aberrantes qui sont automatiquement identifiées avec en

---

[103] Ou ne l'esperez pas, selon le cas.

[104] Ndt : Je conserve une traduction littérale, mais j'avoue n'avoir jamais rencontré cette appelation.

traçant un boxplot sous « Exploration » -« Descriptives », mais cette fois nous devons être beaucoup plus précis. Dans le contexte de la régression linéaire, il y a trois façons conceptuellement distinctes d'envisager une observation « anormale ». Tous les trois sont intéressants, mais elles ont des implications assez différentes pour votre analyse.

Le premier type d'observation inhabituelle est une observation **aberrante**. La définition d'une valeur aberrante (dans ce contexte) est une observation très différente de ce que le modèle de régression prévoit. La Figure 12-17 en donne un exemple. En pratique, nous opérationnalisons ce concept en disant qu'une valeur aberrante est une observation qui a un très grand résidu de Studentised, ϵ_i. Les valeurs aberrantes sont intéressantes : une valeur aberrante importante *peut* correspondre à des données inutiles, par exemple, les variables peuvent avoir été enregistrées incorrectement dans l'ensemble de données, ou un autre défaut peut être détectable. Notez que vous ne devriez pas jeter une observation simplement parce qu'il s'agit d'une observation aberrante. Mais le fait qu'il s'agisse d'une valeur aberrante indique souvent qu'il faut examiner ce cas de plus près et essayer de comprendre pourquoi il est si différent.

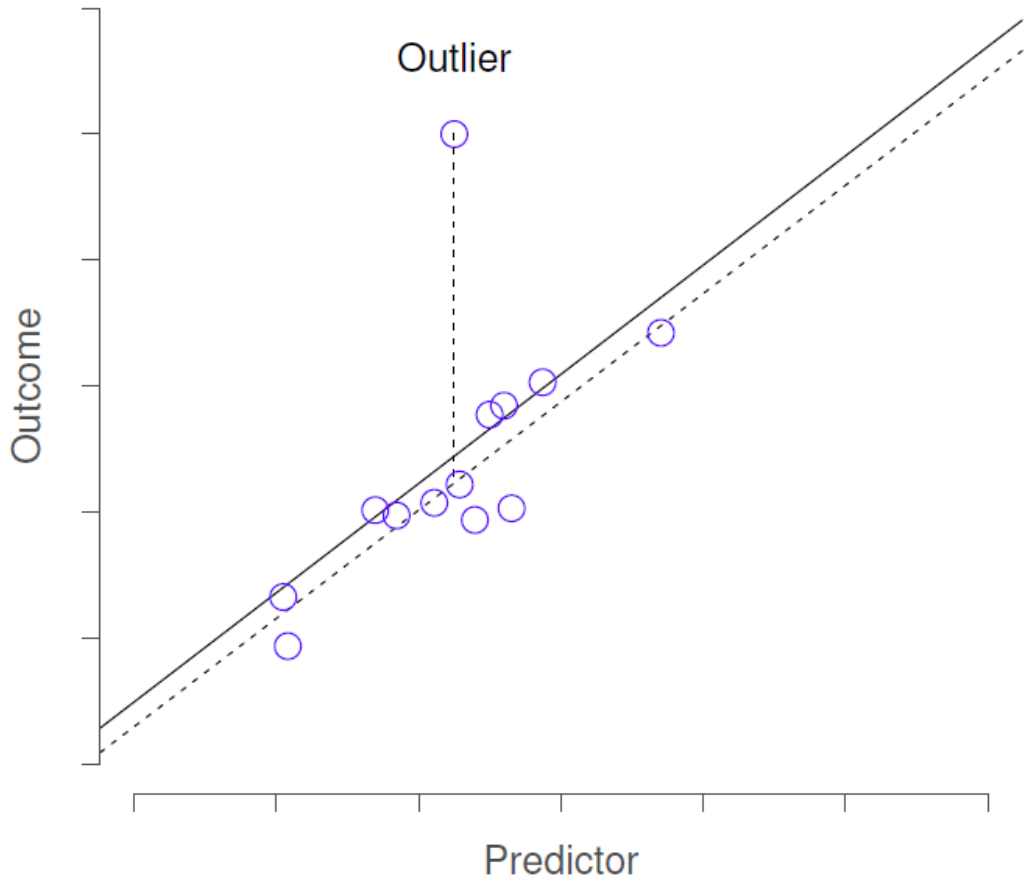


Figure 12-17 : Une illustration des valeurs aberrantes. Les lignes pointillées représentent la ligne de régression qui aurait été estimée sans l'observation anormale incluse, et le résidu correspondant (c.-à-d. le résidu de Studentisé). La ligne continue montre la ligne de régression avec l'observation anormale incluse. La valeur aberrante a une valeur inhabituelle sur la variable résultat (emplacement de l'axe des y), mais pas sur le prédicteur (emplacement de l'axe des x), et se trouve loin de la ligne de régression.

La deuxième façon pour une observation d'être inhabituelle est d'avoir un **effet de levier** élevé, ce qui se produit lorsque l'observation est très différente de toutes les autres observations. Cela ne doit pas nécessairement correspondre à un résidu important. Si l'observation se révèle inhabituelle exactement de la même façon pour toutes les variables, elle peut en fait se situer très près de la ligne de régression. La Figure 12-18 en donne un exemple. L'effet de levier d'une observation est opérationnalisé par sa *valeur chapeau*, généralement écrite en *hi*. La formule de la valeur du chapeau est assez compliquée[105], mais son interprétation ne l'est pas : *hi* est une évaluation de « l'influence » de la *i-ème* observation sur l'orientation finale de la ligne de régression.

En général, si une observation est éloignée des autres en termes de variables prédictrices, elle aura une grande valeur chapeau (à titre indicatif, un effet de levier élevé est obtenu lorsque la valeur de chapeau est plus de 2-3 fois la moyenne ; notez également que la somme des valeurs chapeau est obligatoirement égale à *K + 1*).

---

[105] Encore une fois, pour les fanatiques de l'algèbre linéaire : la "matrice chapeau" est définie comme étant la matrice **H** qui convertit le vecteur des valeurs observées y en un vecteur des valeurs ajustées $\hat{y}$, tel que $\hat{y} = H_y$. Le nom vient du fait que c'est la matrice qui « met un chapeau sur y ». La *valeur* chapeau de la *i-ème* observation est le *i-ème* élément diagonal de cette matrice (donc techniquement je devrais l'écrire comme h_{ii} plutôt que h_i). Et au cas où ça vous intéresserait, voici comment ça se calcule : $H = X(X'X)^{-1}X'$. N'est-ce pas joli ?

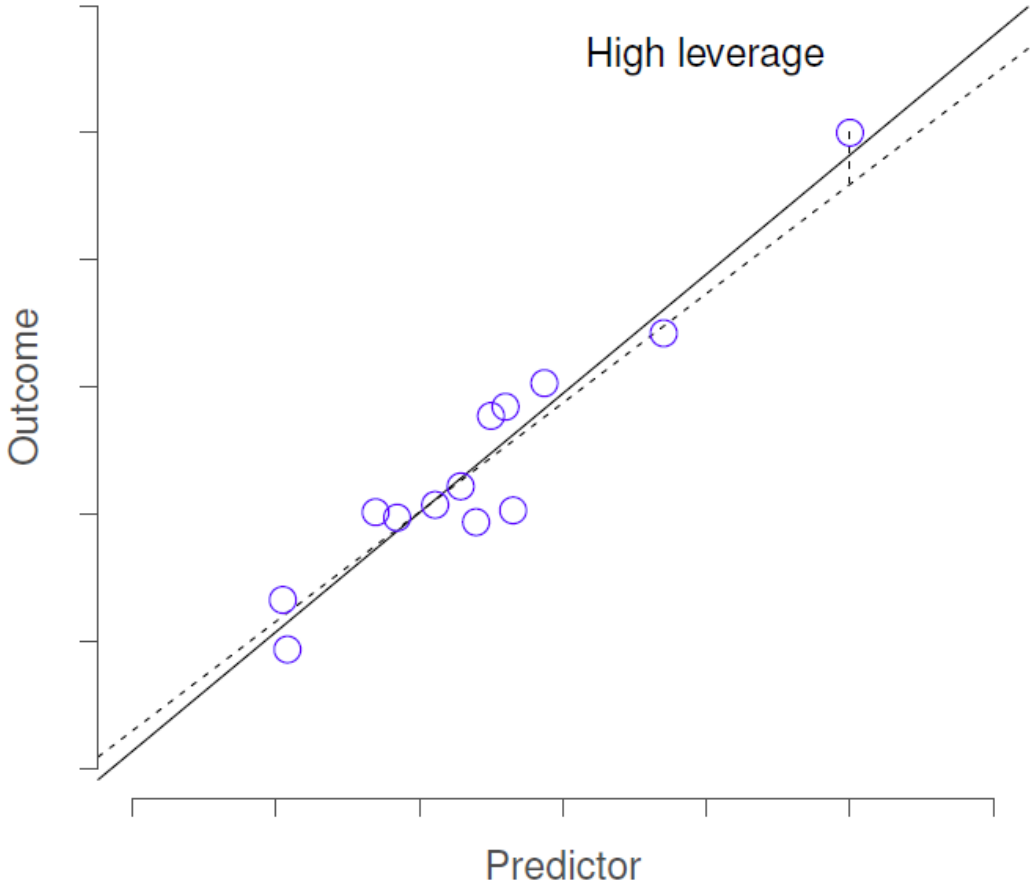


Figure 12-18 : Illustration des points à fort effet de levier. L'observation anormale dans ce cas est inhabituelle, tant du point de vue du prédicteur (axe des x) que du résultat (axe des y), mais cette écart est très cohérent avec le modèle de corrélations qui existe entre les autres observations. L'observation est très proche de la ligne de régression et ne la déforme pas.

Les points à fort effet de levier valent également la peine d'être examinés plus en détail, mais ils sont beaucoup moins susceptibles d'être une source de préoccupation à moins qu'ils ne soient aussi des valeurs aberrantes.

Cela nous amène à notre troisième mesure de l'originalité, **l'influence d'une observation**. Une observation à forte influence est une valeur aberrante qui a un effet de levier élevé. C'est-à-dire qu'il s'agit d'une observation qui est très différente de toutes les autres à certains égards, et qui se situe aussi très loin de la ligne de régression. Ceci est illustré à la Figure 12-19. Remarquez le contraste avec les deux types de valeurs précédents. Les valeurs aberrantes ne déplacent pas beaucoup la ligne de régression et les points de levier élevés non plus. Mais une donnée qui est à la fois une valeur aberrante et qui a un effet de levier élevé a un effet important sur la droite de régression. C'est pourquoi nous parlons pour ces points d'une grande influence, et c'est pourquoi ils nous préoccupent plus. Nous opérationnalisons l'influence en fonction d'une mesure appelée **distance de Cook**.

$$D_i = \frac{\epsilon_i^{*2}}{K+1} \times \frac{h_i}{1-h_1}$$

Notez qu'il s'agit d'une multiplication d'une mesure de la valeur aberrante de l'observation (la partie à gauche), d'une mesure de l'effet de levier de l'observation (la partie à droite).

Afin d'avoir une grande distance de Cook, une observation doit être une valeur aberrante assez importante et avoir un effet de levier élevé. A titre indicatif, la distance de Cook supérieure à 1 est souvent considérée comme grande (c'est ce que j'utilise généralement comme une règle rapide et approximative).

Dans Jamovi, l'information sur la distance du Cook peut être calculée en cliquant sur la case à cocher « Cook's Distance » dans les options « Assumption Checks » - « Data Summary ». Dans le cas du modèle de régression multiple que nous avons utilisé comme exemple dans ce chapitre, vous obtenez les résultats présentés à la Figure 12-20.

Vous pouvez voir que, dans cet exemple, la valeur moyenne de la distance de Cook est de 0,01, et la plage est de 0,0000000262 à 0,11, donc c'est un peu loin de la règle empirique mentionnée ci-dessus une distance de Cook supérieure à 1 est considérée comme grande.

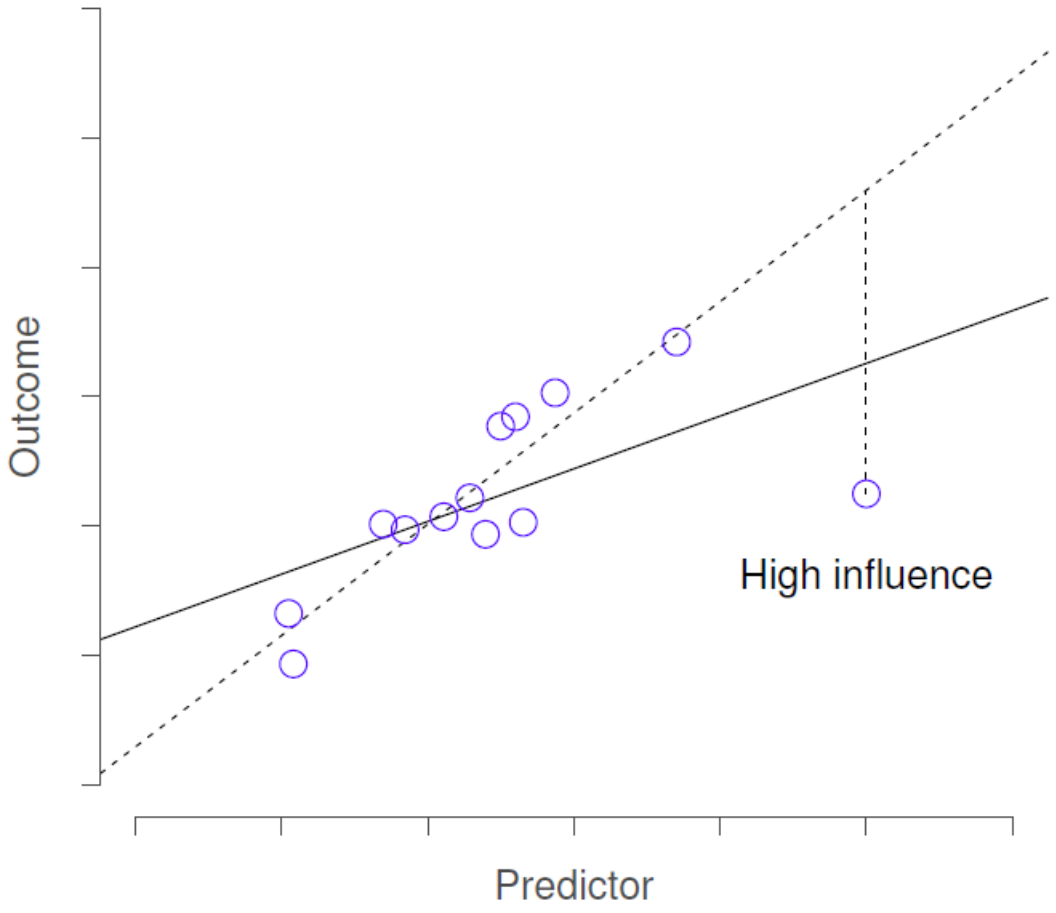


Figure 12-19 : Illustration des points d'influence élevés. Dans ce cas, l'observation anormale est très inhabituelle sur la variable prédictive (axe des *x*) et se situe très loin de la ligne de régression. Par conséquent, la ligne de régression est fortement déformée, même si (dans ce cas) l'observation anormale est tout à fait typique en termes de variable résultat (axe des *y*).

Data Summary

Cook's Distance

			Range	
Mean	Median	SD	Min	Max
0.01	0.00	0.02	2.62e–5	0.11

Figure 12-20 : Copie d'écran Jamovi montrant le tableau pour les statistiques de la distance du Cook

Une question évidente à se poser ensuite est la suivante : si vous avez de grandes valeurs pour la distance de Cook, que devez-vous faire ? Comme toujours, il n'y a pas de règle absolue. La première chose à faire est probablement d'essayer d'exécuter la régression avec la valeur aberrante dont la plus grande distance de Cook est exclue[106] et de voir ce qui arrive au modèle et aux coefficients de régression. S'ils sont vraiment très différents, il est temps de commencer à fouiller dans votre ensemble de données et les notes que vous que vous avez sans aucun doute gribouillée lorsque vous avez mené votre étude. Essayez de comprendre *pourquoi* c'est si différent. Si cela vous conduit à penser que cette données fausse gravement vos résultats, vous pouvez envisager de l'exclure, mais c'est loin d'être idéal, à moins que vous n'ayez une explication solide sur le fait que ce cas particulier est qualitativement différent des autres et mérite donc d'être traité séparément.

Vérification de la normalité des résidus

Comme bon nombre des outils statistiques dont nous avons discuté dans ce livre, les modèles de régression reposent sur une hypothèse de normalité. Dans ce cas, nous supposons que les résidus sont normalement répartis. La première chose que nous pouvons

[106] Parce qu'il n'y a actuellement pas de moyen très facile de le faire avec Jamovi, un programme de régression plus puissant tel que le package car de R serait préférable pour cette analyse plus avancée.

faire est de tracer un graphique QQ via l'option « Assumption Checks » - « Q-Q plot of residuals ».

La sortie est illustrée à la Figure 12-21, montrant les résidus standardisés représentés sur le graphique en fonction de leurs quantiles théoriques selon le modèle de régression.

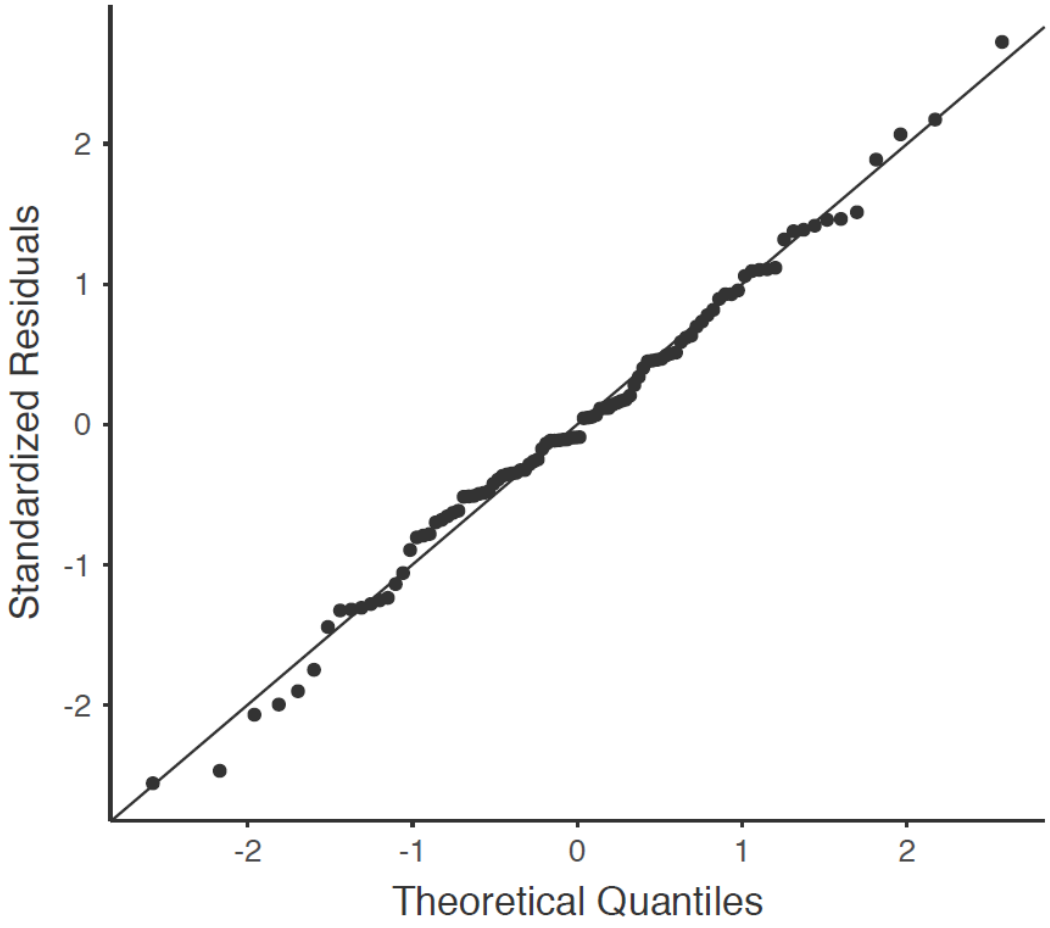


Figure 12-21 : Représentation graphique des quantiles théoriques suivant le modèle, par rapport aux quantiles des résidus normalisés, produits avec Jamovi

Nous devrions vérifier également la relation entre les valeurs ajustées et les résidus eux-mêmes. Pour ce faire, Jamovi peut utiliser l'option « Residuals Plots », qui fournit un diagramme de dispersion pour chaque variable prédictive, entre la variable résultat et les valeurs ajustées par rapport aux résidus, voir la Figure 12-22. Dans ces figures, nous recherchons une distribution assez uniforme des « points » sans regroupement ni structuration claire. Si l'on regarde ces figures, il n'y a rien de particulièrement inquiétant puisque les points sont répartis de façon assez uniforme sur l'ensemble de la figure. Il peut y avoir un peu de non-uniformité dans le graphique (b), mais ce n'est pas un écart important et il ne vaut probablement pas la peine de s'en inquiéter.

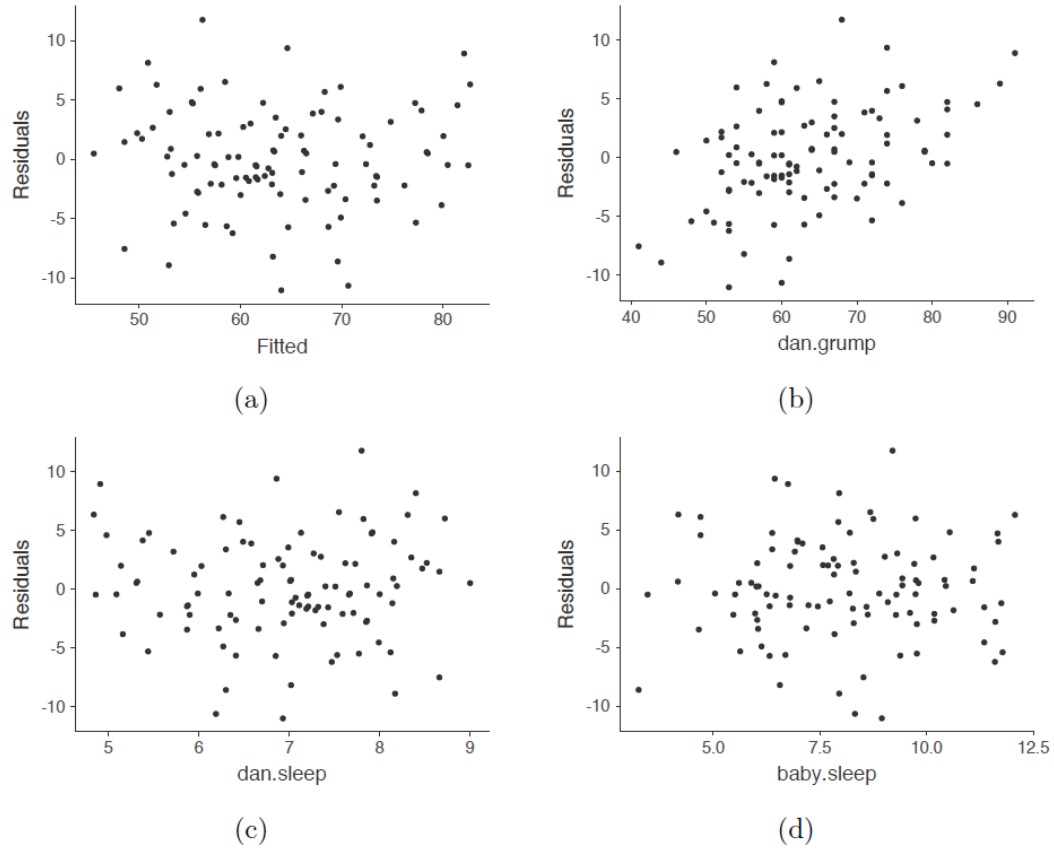


Figure 12-22 : Graphique des résidus réalisés avec Jamovi

Si nous étions inquiets, la solution à ce problème (et à bien d'autres encore) consiste dans bien des cas de transformer une ou plusieurs variables. Nous avons discuté des bases de la transformation des variables dans les sections 6.3 et 6.4, mais je tiens à souligner une autre possibilité que je n'ai pas expliquée en détail plus tôt : la transformation Box-Cox. La fonction Box-Cox est assez simple et elle est très largement utilisée

$$f(x,\lambda) = \frac{x^{\lambda} - 1}{\lambda}$$

pour toutes les valeurs de λ sauf $\lambda = 0$. Quand $\lambda = 0$ nous prenons juste le logarithme naturel (i.e. ln(x)). Vous pouvez le calculer à l'aide de la fonction BOXCOX de l'écran « Compute » des feuilles de données dans Jamovi.

Vérification de la colinéarité

Le dernier type de diagnostic de régression que je vais aborder dans ce chapitre est l'utilisation des **facteurs d'inflation de variance** (VIF), qui sont utiles pour déterminer si les prédicteurs de votre modèle de régression sont trop fortement corrélés les uns aux autres. Il y a un facteur d'inflation de variance associé à chaque prédicteur X_k du modèle.

La formule pour le *k-ème* VIF est :

$$\text{VIF}_k = \frac{1}{1 - R^2_{(-k)}}$$

Où $R^2_{(-k)}$ renvoie à la valeur *R-carrée* que vous obtiendriez si vous exécutiez une régression en utilisant X_k comme variable de résultat, et toutes les autres variables X comme prédicteurs. L'idée ici est que $R^2_{(-k)}$ est une très bonne mesure de la mesure où X_k est corrélé avec toutes les autres variables du modèle.

La racine carrée de VIF est aisément interprétable. Elle est une mesure du fait que l'intervalle de confiance pour le coefficient b_k correspondant est plus grand que ce à quoi vous vous attendriez si les variables prédictrices étaient toutes indépendantes et non corrélées les unes aux autres. Si vous n'avez que deux prédicteurs, les valeurs VIF seront toujours les mêmes, comme nous pouvons le voir en cochant « Collinearity » dans les options « Regression » - « Assumptions » dans Jamovi. Pour dan.sleep et baby.sleep, la VIF est de 1,65. Et puisque la racine carrée de 1,65 est de 1,28, nous constatons que la corrélation entre nos deux prédicteurs ne pose pas de problème.

Pour donner une idée de ce que nous pourrions nous trouver avec un modèle qui pose des problèmes de colinéarité plus importants, supposons que j'exécute un modèle de régression beaucoup moins intéressant, dans lequel j'essaie de prédire le jour où les données sont recueillies, en fonction de toutes les autres variables de l'ensemble des données. Pour comprendre pourquoi ce serait un peu problématique, jetons un coup d'oeil à la matrice de corrélation des quatre variables :

```
             dan.sleep  baby.sleep   dan.grump         day
dan.sleep   1.00000000  0.62794934 -0.90338404 -0.09840768
baby.sleep  0.62794934  1.00000000 -0.56596373 -0.01043394
dan.grump  -0.90338404 -0.56596373  1.00000000  0.07647926
day        -0.09840768 -0.01043394  0.07647926  1.00000000
```

Nous avons des corrélations assez importantes entre certaines de nos variables prédictrices ! Lorsque nous exécutons le modèle de régression et que nous examinons les valeurs de VIF, nous constatons que la colinéarité cause beaucoup d'incertitude pour les coefficients. Si vous exécutez la régression, comme dans la Figure 12-23 alors vous pouvez voir d'après les valeurs de VIF que c'est là une très bonne colinéarité.

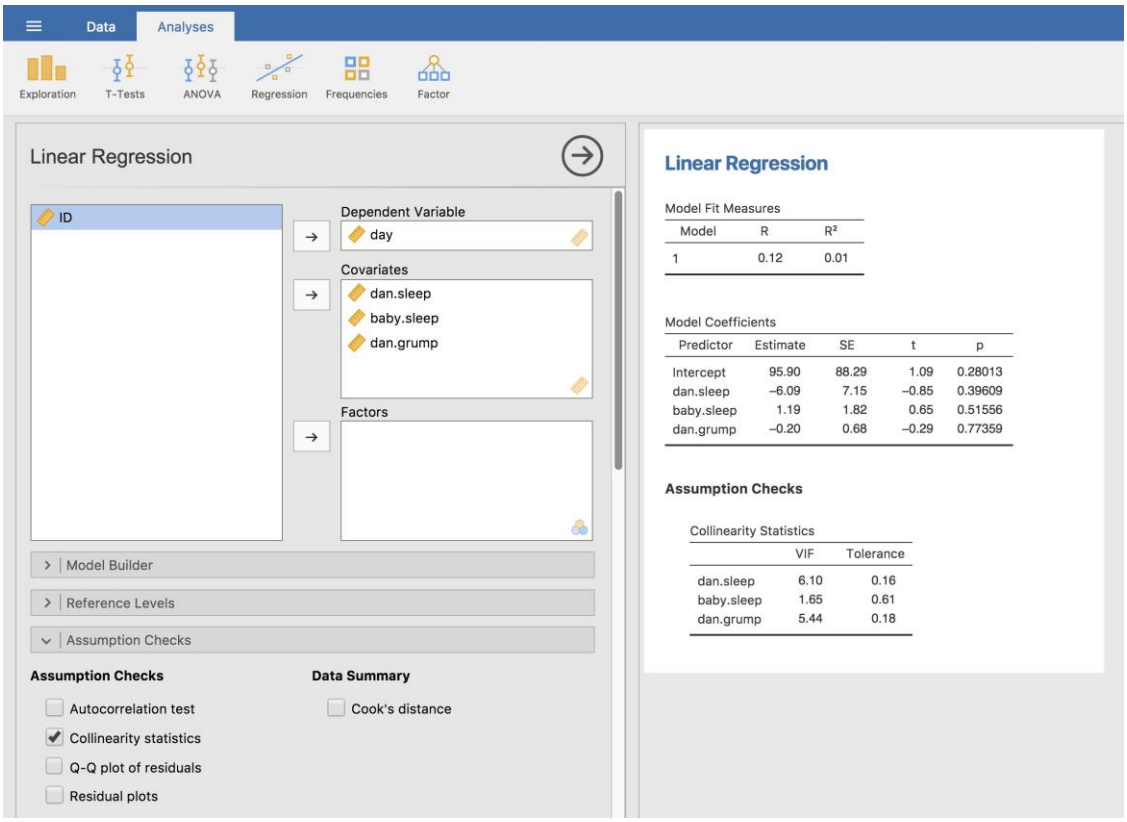


Figure 12-23 : Statistiques de colinéarité pour la régression multiple produites avec Jamovi

Choix du modèle

Un problème assez important subsiste, celui de la « sélection des modèles ». Autrement dit, si nous avons un ensemble de données qui contient plusieurs variables, lesquelles devrions-nous inclure comme prédicteurs et lesquelles ne devrions-nous pas inclure ? Dit autrement, nous avons un problème de **sélection de variables**. En général, la sélection d'un modèle est une affaire complexe, mais elle est quelque peu simplifiée si nous nous limitons au problème du choix d'un sous-ensemble des variables qui devraient être incluses dans le modèle. Néanmoins, je ne vais pas essayer de couvrir ce sujet, même réduit, très en détails. Au lieu de cela, je vais parler de deux grands principes auxquels vous devez réfléchir, puis discuter d'un outil concret que Jamovi fournit pour vous aider à sélectionner un sous-ensemble de variables à inclure dans votre modèle. Premièrement, les deux principes :

- C'est bien d'avoir une base substantielle pour vos choix. En d'autres termes, dans bien des situations, le chercheur a de bonnes raisons de choisir un petit nombre de modèles de régression possibles qui présentent un intérêt théorique. Ces modèles auront une interprétation ayant du sens dans le contexte de votre domaine. Ne négligez jamais l'importance de cette question. Les statistiques servent le processus scientifique, et non l'inverse.

- Dans la mesure où vos choix reposent sur l'inférence statistique, il y a un compromis entre la simplicité et la qualité de l'ajustement. Plus vous ajoutez de variables

prédictrices au modèle, plus il devient complexe. Chaque prédicteur ajoute un nouveau paramètre libre (c.-à-d. un nouveau coefficient de régression) et chaque nouveau paramètre augmente la capacité du modèle à « absorber » les variations aléatoires. Ainsi, la qualité de l'ajustement (p. ex., R2) continue d'augmenter, parfois de façon triviale ou par hasard, à mesure que vous ajoutez d'autres variables prédicteurs quoi qu'il arrive. Si vous voulez que votre modèle soit capable de bien généraliser à de nouvelles observations, vous devez éviter d'ajouter trop de variables.

Ce dernier principe est souvent appelé **le rasoir d'Ockham** et est souvent résumé par le dicton succinct suivant : *ne pas multiplier les entités au-delà de la nécessité*. Dans ce contexte, cela signifie qu'il ne faut pas balancer un tas de prédicteurs largement non pertinents juste pour augmenter votre R^2. Oui, l'original était mieux.

Quoi qu'il en soit, ce dont nous avons besoin, c'est d'un critère mathématique concret qui appliquera le principe qualitatif du rasoir d'Ockham dans le contexte du choix d'un modèle de régression. Il s'avère qu'il y a plusieurs possibilités. Celui dont je vais parler est le **critère d'information Akaike** (AIC ; Akaike, Akaike (1974)) simplement parce qu'il est disponible en option dans Jamovi.

Dans le contexte d'un modèle de régression linéaire (et en ignorant les termes qui ne dépendent d'aucune façon du modèle !), l'AIC pour un modèle qui a K variables prédicteurs plus une intersection est

$$AIC = \frac{\text{SS}_{\text{res}}}{\hat{\sigma}^2} + 2K$$

Plus la valeur AIC est faible, meilleures sont les performances du modèle. Si nous ignorons les détails de bas niveau, ce que fait l'AIC est assez évident. À gauche, nous avons un terme qui augmente à mesure que les prédictions du modèle empirent ; à droite, nous avons un terme qui augmente à mesure que la complexité du modèle augmente. Le meilleur modèle est celui qui correspond bien aux données (faibles résidus, côté gauche) en utilisant le moins de prédicteurs possible (faible K, côté droit). En bref, il s'agit d'une simple mise en oeuvre du rasoir d'Ockham.

L'AIC peut être ajouté à la table de sortie « Model Fit Measures » avec la case à cocher « AIC », et une façon assez fastidieuse d'évaluer différents modèles consiste à voir si la valeur « AIC » est inférieure si l'on supprime un ou plusieurs des prédicteurs du modèle de régression. C'est la seule façon actuellement mise en oeuvre dans Jamovi, mais il existe des alternatives dans d'autres programmes plus puissants, tels que *R*. Ces méthodes alternatives peuvent automatiser le processus de suppression sélective (ou d'ajout) de variables prédictrices pour trouver le meilleur AIC. Bien que ces méthodes ne soient pas implémentées dans Jamovi, je vais les mentionner brièvement ci-dessous pour que vous les connaissiez.

Élimination rétrograde

Dans l'élimination rétrograde, vous commencez par le modèle de régression complet avec tous les prédicteurs possibles. Ensuite, à chaque « étape », nous essayons toutes les façons

possibles d'éliminer l'une des variables, et celle qui est la meilleure (c.-à-d. ayant la valeur AIC la plus faible) est acceptée. Ceci devient notre nouveau modèle de régression, et nous essayons alors toutes les suppressions possibles dans le nouveau modèle, en choisissant à nouveau l'option avec l'AIC le plus bas. Ce processus se poursuit jusqu'à ce que nous obtenions un modèle dont la valeur AIC est inférieure à celle de tous les autres modèles possibles que vous pourriez produire en supprimant un de ses prédicteurs.

Sélection en avant

A la place, vous pouvez également essayer la sélectionner en avant. Cette fois-ci, nous partons du modèle le plus petit possible comme point de départ, et nous ne considérons que les ajouts possibles au modèle. Cependant, il y a une complication. Vous devez également préciser quel est le plus grand modèle possible que vous êtes prêt à accepter. Bien que la sélection rétroactive et en avant puisse mener à la même conclusion, ce n'est pas toujours le cas.

Une mise en garde

Les méthodes automatisées de sélection de variables sont séduisantes, surtout lorsqu'elles sont regroupées dans des fonctions (assez) simples de programmes statistiques puissants. Elles fournissent un élément d'objectivité à votre sélection de modèle, et c'est plutôt sympa. Malheureusement, ils servent parfois d'excuse à l'insouciance. Il n'est plus nécessaire de réfléchir soigneusement aux prédicteurs à ajouter au modèle et aux fondements théoriques de leur inclusion. Tout est résolu par la magie de l'AIC. Et si nous commençons à lancer des principes comme le rasoir d'Ockham, on dirait que tout est emballé dans un joli petit paquet que personne ne peut contester.

Ou, peut-être pas. Tout d'abord, il y a très peu d'accord sur ce qui est considéré comme un critère approprié de sélection de modèle. Lorsqu'on m'a enseigné l'élimination rétrograde durant le premier cycle, nous avons utilisé des tests *F* pour le faire, parce que c'était la méthode par défaut utilisée par le logiciel. J'ai décrit l'utilisation de l'AIC, et puisqu'il s'agit d'un texte d'introduction, c'est la seule méthode que j'ai décrite, mais l'AIC n'est guère la Parole des Dieux de la Statistique. Il s'agit d'une approximation, dérivée de certaines hypothèses, et il est garanti qu'elle ne fonctionne que pour de grands échantillons lorsque ces hypothèses sont respectées. Modifiez ces hypothèses et vous obtenez un critère différent, comme le BIC par exemple (également disponible en Jamovi). Adoptez à nouveau une approche différente et vous obtenez le critère NML. Décidez que vous êtes un Bayésien et vous obtenez une sélection de modèle basée sur les odds ratio à portériori. Il y a également un tas d'outils spécifiques à la régression que je n'ai pas mentionnés. Et ainsi de suite. Toutes ces différentes méthodes ont leurs forces et leurs faiblesses, et certaines sont plus faciles à calculer que d'autres (l'AIC est probablement la plus facile du lot, ce qui pourrait expliquer sa popularité). Presque toutes produisent les mêmes réponses lorsque la réponse est « évidente », mais il y a un certain nombre de désaccords lorsque le problème de sélection du modèle devient difficile.

Qu'est-ce que cela signifie en pratique ? Eh bien, je pourrais passer plusieurs années à vous enseigner la théorie de la sélection des modèles, en vous apprenant tous les tenants et

aboutissants de celle-ci afin que vous puissiez enfin décider ce qui selon vous est la bonne méthode. En tant que personne qui a vraiment fait cela, je ne le recommanderais pas. Vous sortirez probablement encore plus confus que lorsque vous avez commencé. Une meilleure stratégie consiste à faire preuve d'un peu de bon sens. Si vous regardez les résultats d'une procédure de sélection automatisée en avant ou rétrograde, et que le modèle qui a du sens est proche d'avoir le plus petit AIC, mais est battu de justesse par un modèle qui n'a aucun sens, alors faites confiance à votre instinct. Le choix d'un modèle statistique est un outil inexact et, comme je l'ai dit au début, *l'interprétabilité est importante*.

Comparaison de deux modèles de régression

Une solution de rechange à l'utilisation de procédures automatisées de sélection de modèles consiste pour le chercheur à sélectionner explicitement deux ou plusieurs modèles de régression à comparer entre eux. Vous pouvez le faire de différentes façons, en fonction de la question de recherche à laquelle vous essayez de répondre. Supposons que nous voulions savoir si la durée de sommeil de mon fils a un rapport avec ma mauvaise humeur, au-delà de ce à quoi nous pourrions nous attendre pour la durée de mon sommeil. Nous voulons également nous assurer que le jour où la mesure est prise n'influence pas la relation. C'est-à-dire que nous nous intéressons à la relation entre baby.sleep et dan.grump, et de ce point de vue, dan.sleep et day sont des variables parasites ou **covariables** que nous voulons contrôler. Dans cette situation, nous aimerions savoir si dan.grump ~ dan.sleep + day + baby.sleep (que j'appellerai modèle 2, ou M2) est un meilleur modèle de régression pour ces données que dan.grump ~ dan.sleep + day (que je vais appeler modèle 1, ou M1). Il y a deux façons de comparer ces deux modèles, l'une basée sur un critère de sélection de modèle comme l'AIC, et l'autre basée sur un test d'hypothèse explicite. Je vais d'abord vous montrer l'approche fondée sur l'AIC parce qu'elle est plus simple et qu'elle découle naturellement de la discussion de la dernière section. La première chose que je dois faire est d'exécuter les deux régressions, de noter l'AIC pour chacune d'elles, puis de sélectionner le modèle avec la valeur AIC la plus petite, car on estime qu'il s'agit du meilleur modèle pour ces données. En fait, ne le faites pas tout de suite. Lisez la suite parce qu'il y a un moyen facile dans Jamovi d'obtenir les valeurs AIC pour différents modèles inclus dans un seul tableau.[107]

Une approche quelque peu différente du problème découle du cadre de vérification des hypothèses. Supposons que vous ayez deux modèles de régression, dont l'un (modèle 1) contient un *sous-ensemble* des prédicteurs de l'autre (modèle 2). C'est-à-dire que le modèle 2 contient tous les prédicteurs inclus dans le modèle 1, plus un ou plusieurs prédicteurs supplémentaires. Lorsque cela se produit, nous disons que le modèle 1 est **imbriqué** dans le modèle 2, ou peut-être que le modèle 1 est un **sous-modèle** du modèle 2. Indépendamment de la terminologie, cela signifie que nous pouvons considérer le modèle 1

[107] Tant que j'y suis, je dois souligner que les données empiriques suggèrent que le BIC est un meilleur critère que l'AIC. Dans la plupart des études de simulation que j'ai vues, le BIC sélectionne beaucoup mieux le bon modèle.

comme une hypothèse nulle et le modèle 2 comme une hypothèse alternative. Et en fait, nous pouvons construire un test *F* pour cela d'une manière assez simple.

Nous pouvons ajuster les deux modèles aux données et obtenir une somme résiduelle de carrés pour les deux modèles. Je les désignerai respectivement par $\mathrm{SS}_{\mathrm{res}}^{(1)}$ et $\mathrm{SS}_{\mathrm{res}}^{(2)}$. L'exposant ici indique simplement de quel modèle il s'agit. Alors notre statistique *F* est la suivante

$$F = \frac{\mathrm{SS}_{\mathrm{res}}^{(1)} - \mathrm{SS}_{\mathrm{res}}^{(1)}/k}{\left(\mathrm{SS}_{\mathrm{res}}^{(2)}\right)/(N-p-1)}$$

où *N* est le nombre d'observations, *p* est le nombre de prédicteurs dans le modèle complet (à l'exclusion de l'intersection) et *k* est la différence dans le nombre de paramètres entre les deux modèles[108].Les degrés de liberté sont ici *k* et N-p-1. Notez qu'il est souvent plus pratique de considérer la différence entre ces deux valeurs SS comme une somme de carrés à part entière. C'est-à-dire

$$\mathrm{SS}_{\Delta} = \mathrm{SS}_{\mathrm{res}}^{(1)} - \mathrm{SS}_{\mathrm{res}}^{(2)}$$

C'est utile parce que nous pouvons exprimer SS_{Δ} comme une évaluation de la mesure dans laquelle les deux modèles font des prédictions différentes au sujet de la variable des résultats.

$$\mathrm{SS}_{\Delta} = \sum_{i} \left(\hat{y}_{\mathrm{res}}^{(2)} - \hat{y}_{\mathrm{res}}^{(1)}\right)^2$$

Spécifiquement, quand $\hat{y}_{\mathrm{res}}^{(1)}$ est la valeur ajustée pour *yi* selon le modèle M_1 et $\hat{y}_{\mathrm{res}}^{(2)}$ est la valeur ajustée pour y_i selon le modèle M_2.

---

[108] Il convient de noter au passage que cette même statistique *F* peut être utilisée pour tester un éventail beaucoup plus large d'hypothèses que celles que je mentionne ici. Très brièvement, notons que le modèle imbriqué M1 correspond au modèle complet M2 lorsque l'on contraint certains des coefficients de régression à zéro. Il est parfois utile de construire des sous-modèles en plaçant d'autres types de contraintes sur les coefficients de régression. Par exemple, il se peut que deux coefficients différents doivent s'additionner à zéro, ou quelque chose de similaire. Vous pouvez aussi construire des tests d'hypothèse pour ce genre de contraintes, mais c'est un peu plus compliqué et la distribution d'échantillonnage pour *F* devenie une distribution connue sous le nom de distribution *F* non centrale, ce qui va bien au-delà de la portée de ce livre ! Tout ce que je veux faire, c'est vous alerter sur cette possibilité.

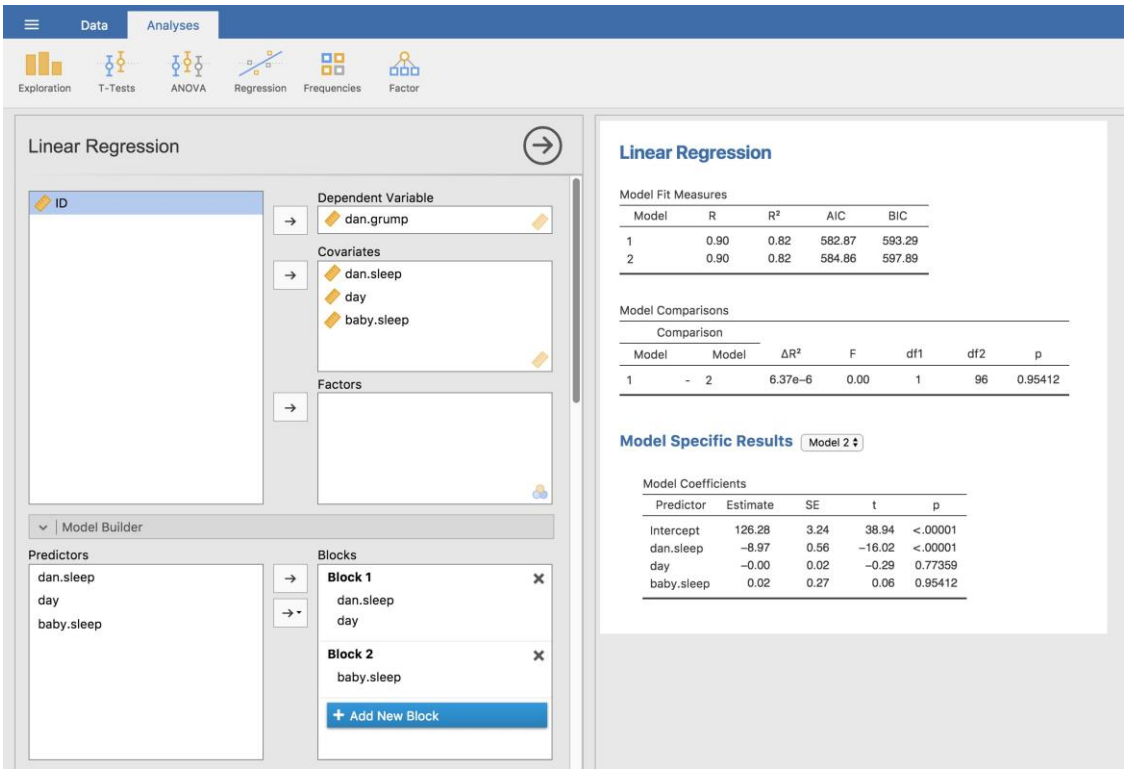


Figure 12-24 : Comparaison de modèles dans Jamovi avec l'option « Model Builder »

C'est donc le test d'hypothèse que nous utilisons pour comparer deux modèles de régression. Maintenant, comment on fait ça à Jamovi ? La solution consiste à utiliser l'option « Model Builder » et de spécifier les prédicteurs du modèle 1 dan.sleep et day dans « Block 1 » puis d'ajouter le prédicteur supplémentaire du modèle 2 (baby.sleep) dans « Block 2 », comme dans la Figure 12-24. Cela montre, dans le tableau « Model Comparisons », que pour les comparaisons entre le modèle 1 et le modèle 2, $F_{(1,96)}=0,00$, $p=,954$. Puisque nous avons $p>.05$, nous retenons l'hypothèse nulle (M_1). Cette approche de régression, dans laquelle nous additionnons toutes nos covariables dans un modèle nul, puis nous *ajoutons* les variables d'intérêt dans un modèle alternatif, puis nous comparons les deux modèles dans un cadre de vérification des hypothèses, est souvent appelée **régression hiérarchique**.

Nous pouvons également utiliser cette option « Model Comparison » pour afficher un tableau qui montre l'AIC et le BIC pour chaque modèle, ce qui facilite la comparaison et l'identification du modèle ayant la valeur la plus faible, comme dans la Figure 12-24.

Résumé

- Vous voulez savoir à quel point la relation est forte entre deux variables ? Calculer une corrélation (section 12.1).
- Dessiner des diagrammes de dispersion (Section 12.2).
- Les idées de base de la régression linéaire et la façon dont les modèles de régression sont estimés (sections 12.3 et 12.4).

- Régression linéaire multiple (section 12.5).
- Mesure du rendement global d'un modèle de régression à l'aide de R2 (section 12.6).
- Tests d'hypothèse pour les modèles de régression (Section 12.7)
- Calcul des intervalles de confiance pour les coefficients de régression et les coefficients standardisés (section 12.8).
- Les hypothèses de régression (section 12.9) et comment les vérifier (section 12.10).
- Sélection d'un modèle de régression (Section 12.11).

Comparaison de plusieurs moyennes (ANOVA à un facteur)

Ce chapitre présente l'un des outils statistiques les plus utilisés en psychologique, connu sous le nom « d'analyse de variance », mais généralement appelé ANOVA. La technique de base a été développée par Sir Ronald Fisher au début du XXe siècle et c'est à lui que nous devons cette terminologie plutôt malheureuse. Le terme ANOVA est un peu trompeur, à deux égards. Premièrement, bien que le nom de la technique fasse référence aux variances, ANOVA s'intéresse à l'étude des différences de moyennes. Deuxièmement, il y a plusieurs choses différentes que l'on appelle toutes des analyses de variance, dont certaines n'ont qu'un lien très ténu les unes avec les autres. Plus loin dans le livre, nous rencontrerons une gamme de différentes méthodes ANOVA qui s'appliquent dans des situations très différentes, mais pour les besoins de ce chapitre, nous ne considérerons que la forme la plus simple d'ANOVA, dans laquelle nous avons plusieurs groupes différents d'observations, et nous nous intéressons à à la différence entre ces groupes pour des variables résultats (variable dépendante) qui nous intéressent. C'est la question à laquelle répond une **ANOVA à un facteur**.

La structure de ce chapitre est la suivante : dans la Section 13.1, je vais introduire un ensemble de données fictives que nous utiliserons comme exemple courant dans tout le chapitre. Après avoir présenté les données, je décrirai la mécanique du fonctionnement réel d'une ANOVA à un facteur (Section 13.2), puis je me concentrerai sur la façon dont vous pouvez en exécuter une avec Jamovi (Section 13.3). Ces deux sections constituent le cœur du chapitre. Le reste du chapitre traite d'une série de sujets importants qui surviennent inévitablement lors de l'exécution d'une analyse de variance, à savoir comment calculer les valeurs de l'effet (section 13.4), les tests et corrections post hoc pour les comparaisons multiples (section 13.5) et les hypothèses sur lesquelles repose l'analyse de variance (section 13.6). Nous parlerons également de la façon de vérifier ces hypothèses et de ce que vous pouvez faire en cas de non-respect de ces hypothèses (sections 13.6.1 à 13.7). Ensuite, nous traiterons de l'ANOVA pour mesures répétées dans les sections 13.8 et 13.9. A la fin du chapitre, nous parlerons un peu de la relation entre l'analyse de variance et les autres outils statistiques (section 13.10).

Un ensemble de données illustratif

Supposons que vous avez participé à un essai clinique dans lequel vous testez un nouvel antidépresseur appelé *Joyzepam*. Afin d'établir un test équitable de l'efficacité du médicament, l'étude comprend trois médicaments distincts à administrer. L'un est un placebo et l'autre est un antidépresseur / médicament anti-anxiété appelé *Anxifree*. Un groupe de 18 participants souffrant de dépression modérée à sévère est recruté pour votre test initial. Comme les médicaments sont parfois administrés conjointement avec une thérapie psychologique, votre étude comprend 9 personnes qui suivent une thérapie cognitivo-comportementale (TCC) et 9 dont ce n'est pas le cas. Les participants sont assignés au hasard (selon la procédure du double aveugle, bien sûr) à un traitement, de sorte qu'il y a 3 personnes en TCC et 3 personnes sans traitement pour chacun des 3 médicaments. Un psychologue évalue l'humeur de chaque personne après trois mois de traitement avec chaque drogue, et l'*amélioration* globale de l'humeur de chaque personne est évaluée sur une échelle allant de -5 à +5. Avec ce plan d'étude, chargeons maintenant le fichier de données dans clinicaltrial.csv. Nous pouvons voir que cet ensemble de données contient les trois variables drug (médicament), therapy (thérapie) and mood.gain (amélioration de l'humeur).

Dans le cadre du présent chapitre, ce qui nous intéresse vraiment, c'est l'effet des médicaments sur l'amélioration de l'humeur. La première chose à faire est de calculer des statistiques descriptives et de faire des graphiques. Au chapitre 4, nous vous avons montré comment faire, et certaines des statistiques descriptives que nous pouvons calculer dans Jamovi sont présentées à la Figure 13-1.

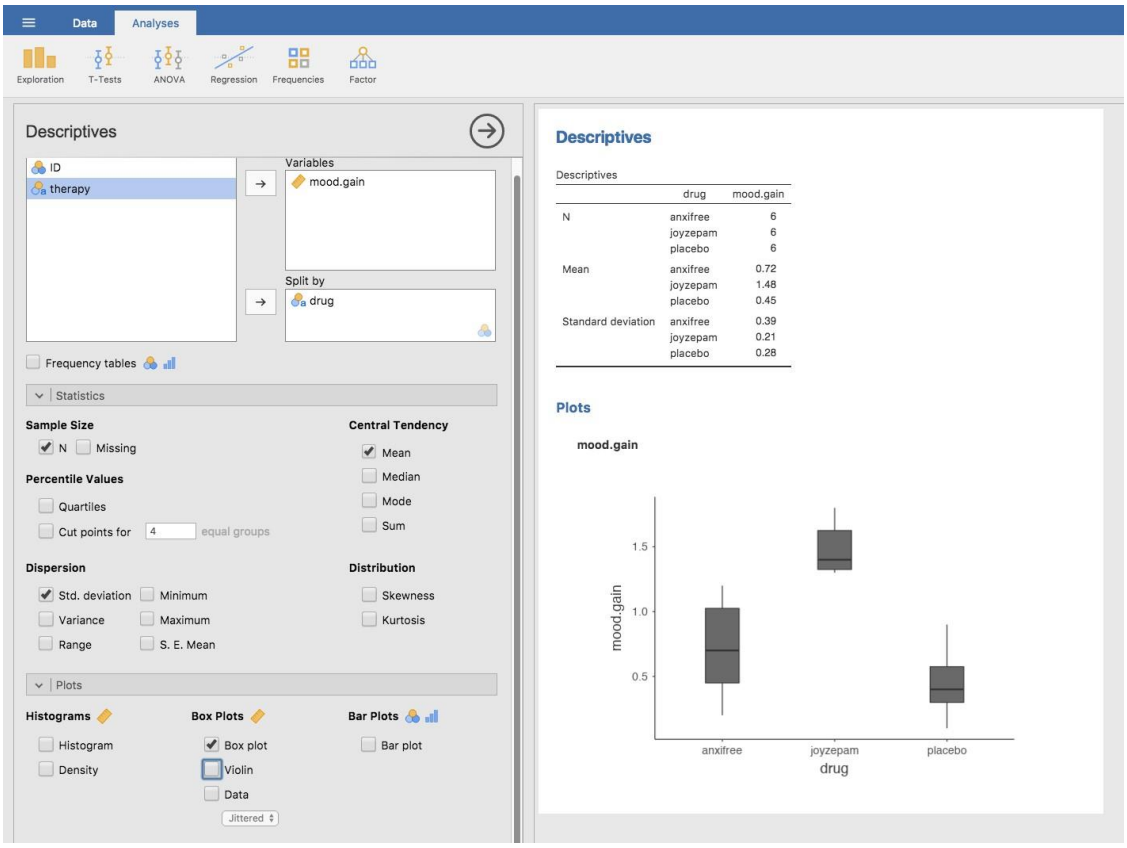


Figure 13-1 : Statistiques descriptive sur l'amélioration de l'humeur et diagrammes en boîtes par médicament administré

Comme le montre clairement la figure, l'humeur des participants du groupe Joyzepam s'est améliorée davantage que celle des participants du groupe Anxifree ou du groupe placebo (témoin). Le groupe Anxifree présente une amélioration de l'humeur plus importante que le groupe témoin, mais la différence n'est pas aussi grande. La question à laquelle nous voulons répondre est la suivante : ces différences sont-elles « réelles » ou sont-elles dues au hasard ?

Comment fonctionne ANOVA

Afin de répondre à la question posée par les données de nos essais cliniques, nous allons procéder à une analyse de variance à un facteur. Je vais commencer par vous montrer comment le faire à la main, en construisant l'outil statistique à partir de zéro et en vous montrant comment vous pourriez le faire si vous n'aviez accès à aucune des fonctions ANOVA intégrées dans Jamovi. Et j'espère que vous le lirez attentivement, que vous essaierez de le faire une ou deux fois pour vous assurer de bien comprendre le fonctionnement d'ANOVA, et qu'une fois que vous aurez compris le concept, ne le referez plus jamais de cette façon.

Le plan expérimental que j'ai décrit dans la section précédente suggère fortement de nous intéresser à la comparaison du changement d'humeur moyen pour les trois différents médicaments. En ce sens, il s'agit d'une analyse similaire au test *t* (chapitre 11) mais

impliquant plus de deux groupes. Si nous posons μ_P pour la moyenne de la population pour le changement d'humeur induit par le placebo, et que μ_A et μ_J représentent les moyennes correspondantes à nos deux médicaments, Anxifree et Joyzepam, alors l'hypothèse nulle (quelque peu pessimiste) que nous voulons tester est que les trois moyennes de populations sont identiques. Autrement dit, *aucun* des deux médicaments n'est plus efficace qu'un placebo. Nous pouvons écrire cette hypothèse nulle comme :

H_0 : il est vrai que $\mu_P = \mu_A = \mu_J$

Par conséquent, notre hypothèse alternative est qu'au moins un des trois traitements différents est différent des autres. C'est un peu délicat d'écrire cela mathématiquement, parce que (comme nous le verrons plus loin) il y a plusieurs façons différentes dont l'hypothèse nulle peut être fausse. Donc pour l'instant, nous allons juste écrire l'hypothèse alternative comme ceci :

H_1 : il n'est pas vrai que $\mu_P = \mu_A = \mu_J$

Cette hypothèse nulle est beaucoup plus difficile à vérifier que toutes celles que nous avons vues auparavant. Comment allons-nous faire ? Une supposition raisonnable serait de « faire une analyse de *variance* », puisque c'est le titre du chapitre, mais il n'est pas particulièrement clair pourquoi une « analyse des *variances* » nous aidera à apprendre quoi que ce soit d'utile sur les *moyennes*. En fait, c'est l'une des plus grandes difficultés conceptuelles que les gens éprouvent lorsqu'ils rencontrent ANOVA pour la première fois. Pour voir comment cela fonctionne, je trouve très utile de commencer par parler des variances. En fait, je vais commencer par jouer à des jeux mathématiques avec la formule qui décrit la variance. C'est-à-dire, nous allons commencer par jouer avec les variances et il s'avérera que cela nous donne un outil utile pour étudier les moyennes.

Deux formules pour la variance de *Y*

Tout d'abord, commençons par introduire quelques notations. Nous utiliserons *G* pour faire référence au nombre total de groupes. Pour notre ensemble de données, il y a trois médicaments, il y a donc trois groupes *G=3*. Ensuite, nous utiliserons *N* pour faire référence à la taille totale de l'échantillon ; il y a un total de *N* = 18 personnes dans notre ensemble de données. De même, utilisons N_k pour indiquer le nombre de personnes dans le *k-ème* groupe. Dans notre essai clinique fictif, la taille de l'échantillon est de $N_k=6$ pour chacun des trois groupes.[109] Enfin, nous utiliserons *Y* pour désigner la variable de résultat. Dans notre cas, *Y* fait référence aux changements d'humeur. Plus précisément, nous utiliserons Y_{ik} pour faire référence au changement d'humeur vécu par le *i-ème* membre du *k-ème* groupe. De même, nous utiliserons $\overline{Y}$ comme changement d'humeur moyen, pour l'ensemble des 18

---

[109] Lorsque tous les groupes ont le même nombre d'observations, le plan expérimental est dit « équilibré ». L'équilibre n'est pas si important pour l'ANOVA à un facteur, qui est le sujet de ce chapitre. Cela devient plus important lorsque vous commencez à faire des analyses de variance plus complexes.

personnes de l'expérience, et $\overline{Y}_k$ *comme* référence au changement d'humeur moyen vécu par les 6 personnes du groupe *k*.

Maintenant que nous avons réglé notre notation, nous pouvons commencer à écrire des formules. Pour commencer, rappelons la formule de variance que nous avons utilisée à la section 4.2, à l'époque où nous ne faisions que des statistiques descriptives. La variance d'échantillon de *Y* est définie comme suit

$$\text{Var}(Y) = \frac{1}{N}\sum_{k=1}^{G}\sum_{i=1}^{N_k}\left(Y_{\text{ik}} - \overline{Y}\right)^2$$

Cette formule semble à peu près identique à celle de la variance de la section 4.2. La seule différence, c'est que cette fois-ci, j'ai deux résumés : Je fais la somme des groupes (c.-à-d. les valeurs pour *k*) et des personnes au sein des groupes (c.-à-d. les valeurs pour *i*). C'est un détail purement esthétique. Si j'avais plutôt utilisé la notation Y_p pour faire référence à la valeur de la variable de résultat pour la personne *p* dans l'échantillon, alors je n'aurais qu'une seule somme. La seule raison pour laquelle nous avons une double sommation ici est que j'ai classé les gens dans les groupes, puis attribué des numéros à des personnes au sein de groupes.

Un exemple concret pourrait être utile ici. Considérons ce tableau, dans lequel nous avons un total de *N=5* personnes triées en groupes *G=2*. Arbitrairement, disons que les gens « cool » sont le groupe 1 et les gens « pas cool » sont le groupe 2, il s'avère que nous avons trois personnes cool (N_1=3) et deux personnes pas cool (N_2=2).

name	person p	group	group num. k	index in group i	grumpiness Y_{ik} or Y_p
Ann	1	cool	1	1	20
Ben	2	cool	1	2	55
Cat	3	cool	1	3	21
Dan	4	uncool	2	1	91
Egg	5	uncool	2	2	22

Notez que j'ai construit deux systèmes d'étiquetage différents ici. Nous avons une variable « personne » *p*, il serait donc tout à fait sensé de parler de Y_p comme de la grinchiosité de la *p-ième* personne de l'échantillon. Par exemple, le tableau montre que Dan est le quatrième, donc nous dirions *p=4*. Donc, quand on parle du caractère grincheux *Y* de « Dan », qui qu'il soit, on pourrait parler de sa grinchiosité en disant que Y_p=91, pour personne *p=4*, bien sûr. Cependant, ce n'est pas la seule façon de parler de Dan. Comme alternative, nous pourrions noter que Dan appartient au groupe « pas cool » (*k=2*), et est en fait la première personne listée dans le groupe pas cool (*i=1*). Il est donc tout aussi valable de faire référence à la mauvaise humeur de Dan en disant que Y_{ik}=91, où *k=2* et *i=1*.

En d'autres termes, chaque personne *p* correspond à une combinaison *ik* unique, et donc la formule que j'ai donnée ci-dessus est en fait identique à notre formule originale pour la variance, qui serait

$$\text{var}(Y) = \frac{1}{N}\sum_{p=1}^{N}\left(Y_p - \overline{Y}\right)^2$$

Dans les deux formules, tout ce que nous faisons est de faire la somme de toutes les observations de l'échantillon. La plupart du temps, nous utiliserions simplement la notation Y_p la plus simple ; l'équation utilisant Y_p est clairement la plus simple des deux. Cependant, lorsque vous faites une analyse de variance, il est important de savoir quels participants appartiennent à quels groupes, et nous devons utiliser la notation avec Y_{ik} pour ce faire.

Des variances aux sommes de carrés

Bien, maintenant que nous avons une bonne idée de la façon dont la variance est calculée, définissons ce qu'on appelle la **somme totale des carrés**, qui est appelée SStot. C'est très simple. Au lieu de faire la moyenne quadratique des écarts , ce que nous faisons lorsque nous calculons la variance, nous nous contentons de les additionner.

La formule de la somme totale des carrés est donc presque identique à la formule de la variance

$$\text{SS}_{\text{tot}} = \sum_{k=1}^{G}\sum_{i=1}^{N_k}\left(Y_{\text{ik}} - \overline{Y}\right)^2$$

Lorsque nous parlons d'analyser les variances dans le contexte d'ANOVA, ce que nous faisons en réalité, c'est de travailler avec les sommes totales des carrés plutôt que la variance en tant que telle. La propriété remarquable de la somme totale des carrés, c'est que nous pouvons la diviser en deux types différents de variation.

Tout d'abord, nous pouvons parler de la **somme des carrés intra groupe**, dans laquelle nous regardons à quel point chaque personne est différente de la moyenne de son propre groupe.

$$\text{SS}_w = \sum_{k=1}^{G}\sum_{i=1}^{N_k}\left(Y_{\text{ik}} - \overline{Y}_k\right)^2$$

où $\overline{Y}_k$ est un groupe moyen. Dans notre exemple, $\overline{Y}_k$ serait le changement d'humeur moyen subi par les personnes qui reçoivent le *k-ème* médicament. Ainsi, au lieu de comparer les individus à la moyenne de *toutes* les personnes participant à l'expérience, nous les comparons seulement à ceux qui font partie du même groupe. Par conséquent, on s'attendrait à ce que la valeur de la SSw soit inférieure à la somme totale des carrés, parce qu'elle ne tient aucunement compte des différences entre les groupes, c.-à-d., si les médicaments ont des effets différents sur l'humeur des gens.

Ensuite, nous pouvons définir une troisième notion de variation qui saisit seulement les différences entre les groupes. Pour ce faire, nous examinons les différences entre la moyenne du groupe $\overline{Y}_k$ et la moyenne générale.

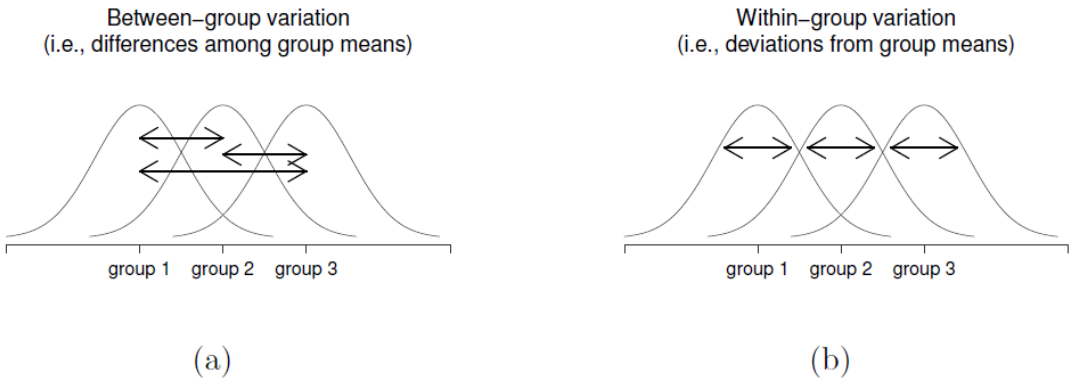


Figure 13-2 : Illustration graphique de la variation « inter groupes » (figure a) et de la variation « intra groupes » (figure b). Sur la gauche, les flèches indiquent les différences entre les moyennes des groupes. Sur la droite, les flèches mettent en évidence la variabilité au sein de chaque groupe.

Afin de quantifier l'ampleur de cette variation, nous calculons la **somme des carrés inter groupes.**

$$\begin{aligned} \mathrm{SS}_b &= \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(\overline{Y}_k - \overline{Y}\right)^2 \\ &= \sum_{k=1}^{G} N_k \left(\overline{Y}_k - \overline{Y}\right)^2 \end{aligned}$$

Il n'est pas trop difficile de montrer que la variation totale entre les personnes dans l'expérience SStot est en fait la somme des différences inter groupes SSb et la variation intra groupes SSw. Nous avons donc,

$$\mathrm{SS}_w = \mathrm{SS}_b + \mathrm{SS}_{tot}$$

Bien, alors qu'est-ce qu'on a trouvé ? Nous avons découvert que la variabilité totale associée à la variable de résultat (SS_{tot}) peut être mathématiquement découpée en la somme de « la variation due aux différences dans les moyennes de l'échantillon pour les différents groupes » (SS_b) plus « le reste de la variation » (SS_w)[110]. En quoi cela m'aide-t-il à savoir que les groupes ont des moyennes de population différentes ? Mais attendez une seconde !

---

[110] SS_w correspond aussi dans l'ANOVA pour groupes indépendants à la variance de l'erreur ou SS_{error}

Maintenant que j'y pense, c'est *exactement* ce qu'on cherchait. Si l'hypothèse nulle est vraie, on s'attendrait à ce que tous les moyennes de l'échantillon soient assez semblables les uns des autres, non ? Et cela impliquerait qu'on s'attendrait à ce que SS_b soit vraiment petit, ou du moins à ce qu'il soit beaucoup plus petit que « la variation associée à tout le reste », SS_w. Je détecte un test d'hypothèse en cours de construction.

De la somme des carrés au *test F*

Comme nous l'avons vu dans la dernière section, l'idée *qualitative* derrière ANOVA est de comparer les deux sommes des carrés SS_b et SS_w entre elles. Si la variation inter groupes SSb est importante par rapport à la variation intra groupe SSw, nous avons des raisons de penser que les moyennes de population pour les différents groupes ne sont pas identiques entre elles. Pour convertir cela en un test d'hypothèse réalisable, il faut un peu de « bricoler ». Je vais d'abord vous montrer ce que nous faisons pour calculer notre statistique de test, le **rapport F**, puis vous donner une idée de la *raison* pour laquelle nous procédons de cette façon.

Afin de convertir nos valeurs SS en un *rapport F*, la première chose que nous devons calculer est le **degré de liberté** associé aux valeurs SS_b et SS_w. Comme d'habitude, les degrés de liberté correspondent au nombre de « données uniques » qui contribuent à un calcul particulier, moins le nombre de « contraintes » qu'ils doivent satisfaire. Pour la variabilité intra-groupe, nous calculons la variation des observations individuelles (N éléments de données) autour de la moyenne du groupe (contraintes G). En revanche, pour la variabilité inter groupes, nous nous intéressons à la variation de la moyenne du groupe (élements de données G) autour de la moyenne générale (1 contrainte). Par conséquent, les degrés de liberté sont ici :

$$ \text{df}_{b}= G-1\\ \text{df}_{w}= N-G $$

Bien, ça semble assez simple. Ensuite, nous convertissons nos sommes des carrés en une valeur les « carrés moyens », en divisant par les degrés de liberté :

$$ \text{MS}_{b} = \frac{\text{SS}_{b}}{\text{ddl}_{b}}\\ \text{MS}_{w} = \frac{\text{SS}_{w}}{\text{ddl}_{w}}\\ $$

Enfin, nous calculons le *rapport F* en divisant le carré moyen intergroupes par le carré moyen intra groupes :

$$F = \frac{\text{MS}_b}{\text{MS}_w}$$

À un niveau très général, l'intuition derrière la statistique F est simple. Des valeurs plus grandes de F signifient que la variation inter groupes est importante par rapport à la variation intra groupes. Par conséquent, plus la valeur de F est élevée, plus nous avons de preuves contre l'hypothèse nulle. Mais quelle doit être la taille de F pour que H_0 soit réellement *rejetée* ? Pour répondre à cela, vous avez besoin d'une compréhension un peu plus profonde de ce qu'est l'ANOVA et de ce que sont réellement les valeurs des carrés moyens.

La section suivante l'aborde un peu plus en détail, mais pour les lecteurs qui ne sont pas intéressés par les détails de ce que le test mesure réellement, je vais aller droit au but. Afin de compléter notre test d'hypothèse, nous devons connaître la distribution d'échantillonnage pour *F* si l'hypothèse nulle est vraie. Il n'est pas surprenant que la distribution d'échantillonnage pour la *statistique F* sous l'hypothèse nulle soit une *distribution F*. Si vous vous souvenez de notre discussion sur la distribution *F* au chapitre 7, la distribution *F* a deux paramètres, correspondant aux deux degrés de liberté impliqués. Le premier df1 est le degré de liberté dfb (dfb) inter groupes, et le second df2 est le degré de liberté ddlw (dfw) intra groupes.

Tableau 13-1 : Toutes les quantités clés d'une ANOVA organisée dans un tableau ANOVA "standard". Les formules pour toutes les quantités (à l'exception de la valeur p qui a une formule très moche et qui serait cauchemardesquement difficile à calculer sans ordinateur) sont présentées.

	df	sum of squares	mean squares	F-statistic	p-value
between groups	$\text{df}_b = G - 1$	$\text{SS}_b = \sum_{k=1}^{G} N_k(\bar{Y}_k - \bar{Y})^2$	$\text{MS}_b = \frac{\text{SS}_b}{\text{df}_b}$	$F = \frac{\text{MS}_b}{\text{MS}_w}$	[complicated]
within groups	$\text{df}_w = N - G$	$\text{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$	$\text{MS}_w = \frac{\text{SS}_w}{\text{df}_w}$	-	-

Le modèle pour les données et la signification de *F*

Au niveau fondamental, ANOVA est une compétition entre deux modèles statistiques différents, H_0 et H_1. Lorsque j'ai décrit les hypothèses nulle et alternative au début de la section, j'étais un peu imprécis quant à la nature réelle de ces modèles. Je vais y remédier maintenant, même si vous ne m'aimerez probablement pas pour cela. Si vous vous souvenez bien, notre hypothèse nulle était que toutes les moyennes de groupe sont identiques les unes aux autres. Si c'est le cas, alors une façon naturelle de penser à la variable de résultat Y_{ik} consiste à décrire les scores individuels en termes d'une seule moyenne de population μ, plus l'écart par rapport à cette moyenne de population. Cet écart est habituellement appelé ϵ_{ik} et est traditionnellement appelé l'**erreur** ou le **résidu** associé à cette observation. Mais attention cependant. Tout comme nous l'avons vu avec le mot « significatif », le mot « erreur » a une signification technique en statistique qui n'est pas tout à fait la même que sa définition française courante. Dans le langage courant, « erreur » implique une erreur quelconque, mais ce n'est pas le cas dans les statistiques (ou du moins, pas nécessairement). Dans cet esprit, le mot « résiduel » est un meilleur terme que le mot « erreur ». En statistique, les deux mots signifient « variabilité résiduelle », c'est-à-dire « choses » que le modèle ne peut expliquer.

Quoi qu'il en soit, voici à quoi ressemble l'hypothèse nulle quand on l'écrit comme modèle statistique

$$Y_{ik} = \mu + \epsilon_{ik}$$

où nous faisons l'*hypothèse* (discutée plus loin) que les valeurs résiduelles ϵ_{ik} sont normalement distribuées, avec une moyenne 0 et un écart type σ qui est le même pour tous les groupes. Pour utiliser la notation que nous avons introduite au chapitre 7, nous écrivons cette hypothèse comme suit

$$\epsilon_{\text{ik}} \sim Normal(0, \sigma^2)$$

Et l'hypothèse alternative, H_1 ? La seule différence entre l'hypothèse nulle et l'hypothèse alternative est que nous permettons à chaque groupe d'avoir une moyenne de population différente. Donc, si nous posons μ_k la moyenne de population pour le *k-ème* groupe dans notre expérience, alors le modèle statistique correspondant à H_1 est le suivant

$$Y_{\text{ik}} = \mu_k + \epsilon_{\text{ik}}$$

où, encore une fois, nous supposons que les termes d'erreur sont normalement distribués avec une moyenne de 0 et un écart-type σ. Autrement dit, l'hypothèse alternative suppose également que

$$\epsilon \sim Normal(0, \sigma^2)$$

Bien, maintenant que nous avons décrit plus en détail les modèles statistiques qui sous-tendent H_0 et H_1, il est maintenant assez simple de dire ce que les carrés moyens mesurent, et ce que cela signifie pour l'interprétation de *F*. Je ne vous ennuierai pas avec la preuve, mais il s'avère que le carré moyen intragroupe, MS_w, peut être considéré comme un estimateur (au sens technique, chapitre 8) de la variance des erreurs σ^2. La moyenne quadratique entre groupes MS_b est également un estimateur, mais ce qu'elle estime est la variance d'erreur *plus* une quantité qui dépend des différences réelles entre les moyennes du groupe. Si nous appelons cette quantité *Q*, alors nous pouvons voir que la *statistique F* est fondamentalement[111]

$$F = \frac{\hat{Q} + \hat{\sigma}^2}{\hat{\sigma}^2}$$

où la valeur vraie *Q=0* si l'hypothèse nulle est vraie, et *Q>0* si l'hypothèse alternative est vraie (par exemple, Hays 1994, ch. 10. Par conséquent, au minimum, *la valeur de F doit être supérieure à 1* pour avoir une chance de rejeter l'hypothèse nulle. Notez que cela *ne signifie pas* qu'il est impossible d'obtenir une valeur *F* inférieure à 1, mais que si l'hypothèse nulle est vraie, la distribution d'échantillonnage du rapport *F* a une moyenne de 1[112], nous devons donc voir des valeurs *F* supérieures à 1 afin de rejeter la valeur nulle.

---

[111] Si vous lisez le chapitre 14 et regardez comment « l'effet de traitement » au niveau k d'un facteur est défini en termes de valeurs α_k (voir section 14.1.2), il s'avère que *Q* fait référence à une moyenne pondérée des effets du traitement au carré, $Q = (\sum_{k=1}^{G} N_k \alpha_k^2)/(G-1)$

[112] Ou, si nous voulons être pointilleux sur la précision, $1 + \frac{2}{\text{df}_2 - 2}$

Pour être un peu plus précis au sujet de la distribution d'échantillonnage, notons que si l'hypothèse nulle est vraie, MSb et MSw sont tous deux des estimateurs de la variance des résidus ϵ_{ik}. Si ces résidus sont normalement distribués, vous pourriez soupçonner que l'estimation de la variance de ϵ_{ik} suit une distribution de chi-carré, parce que (comme nous l'avons vu à la section 7.6) c'est ce qui caractérise une distribution du chi-carré. C'est ce qu'on obtient quand on fait la somme de plusieurs choses normalement distribuées. Et puisque la distribution *F* est (encore une fois, par définition) ce que vous obtenez lorsque vous prenez le rapport entre deux valeurs qui suivent une distribution de χ^2, nous avons notre distribution d'échantillonnage. Évidemment, je passe sous silence beaucoup de choses quand je dis cela, mais en termes généraux, c'est vraiment de là que vient notre distribution d'échantillonnage.

Un exemple travaillé

La discussion précédente était assez abstraite et un peu technique, alors je pense qu'à ce stade, il pourrait être utile de voir un exemple concret. Pour cela, revenons aux données sur les essais cliniques que j'ai présentées au début du chapitre. Les statistiques descriptives que nous avons calculées au début nous indiquent que notre groupe représente une amélioration moyenne de l'humeur de 0,45 pour le placebo, 0,72 pour Anxifree et 1,48 pour Joyzepam. Avec cela à l'esprit, faisons un jeu comme si nous étions en 1899[113]et commençons à faire quelques calculs au crayon et sur papier. Je ne le ferai que pour les 5 premières observations car nous ne sommes pas en 1899 et je suis très paresseux. Commençons par calculer SS_w, la somme des carrés au sein du groupe. D'abord, dressons un beau tableau pour nous aider dans nos calculs :

[113] Ou, pour être plus précis, un jeu comme si « nous sommes en 1899 et nous n'avons pas d'amis et rien de mieux à faire avec notre temps que de faire des calculs qui n'auraient pas eu de sens en 1899 car ANOVA n'existait pas avant les années 1920 ».

group k	outcome Y_{ik}
placebo	0.5
placebo	0.3
placebo	0.1
anxifree	0.6
anxifree	0.4

À ce stade, la seule chose que j'ai incluse dans le tableau, ce sont les données brutes elles-mêmes. C'est-à-dire la variable de regroupement (c.-à-d. drug) et la variable de résultat (c.-à-d. outcomes = mood.gain) pour chaque personne. Notez que la variable de résultat ici correspond à la valeur de Y~*ik*~ de notre équation précédente. L'étape suivante du calcul consiste à noter, pour chaque personne de l'étude, la moyenne du groupe correspondant,$\overline{Y}_k$. C'est un peu répétitif mais pas particulièrement difficile puisque nous avons déjà calculé les moyennes de ces groupes lors de la réalisation de nos statistiques descriptives :

group k	outcome Y_{ik}	**group mean** $\bar{Y}_k$
placebo	0.5	**0.45**
placebo	0.3	**0.45**
placebo	0.1	**0.45**
anxifree	0.6	**0.72**
anxifree	0.4	**0.72**

Maintenant que nous les avons notés, nous devons calculer, encore une fois pour chaque personne, l'écart par rapport à la moyenne du groupe correspondant. C'est-à-dire, nous voulons faire la soustraction $Y_{\text{ik}} - \overline{Y}_k$. Après avoir fait ça, il faut tout ranger. Voilà ce qu'on obtient quand on fait ça :

group k	outcome Y_{ik}	group mean $\bar{Y}_k$	**dev. from group mean** $Y_{ik} - \bar{Y}_k$	**squared deviation** $(Y_{ik} - \bar{Y}_k)^2$
placebo	0.5	0.45	**0.05**	**0.0025**
placebo	0.3	0.45	**-0.15**	**0.0225**
placebo	0.1	0.45	**-0.35**	**0.1225**
anxifree	0.6	0.72	**-0.12**	**0.0136**
anxifree	0.4	0.72	**-0.32**	**0.1003**

La dernière étape est tout aussi simple. Pour calculer la somme des carrés à l'intérieur d'un groupe, nous additionnons simplement les écarts au carré de toutes les observations :

$$\begin{aligned} SS_w &= 0{,}0025 + 0{,}0225 + 0{,}1225 + 0{,}0136 + 0{,}1003 \\ &= 0{,}2614 \end{aligned}$$

Bien sûr, si nous voulions vraiment obtenir la *bonne* réponse, nous devrions le faire pour les 18 observations de l'ensemble des données, pas seulement pour les cinq premières. Nous pourrions continuer les calculs au crayon et au papier si nous le voulions, mais c'est assez fastidieux. Sinon, il n'est pas trop difficile de le faire dans un tableur dédié tel qu'OpenOffice ou Excel. Essayez de le faire vous-même. Celui que j'ai fait, dans Excel, est dans le fichier clinicaltrial_anova.xls. Lorsque vous le faites, vous devriez obtenir une somme intra-groupe de carrés de 1,39.

Bien. Maintenant que nous avons calculé la variation à l'intérieur des groupes, SSw, il est temps concentrer vers la somme des carrés entre groupes, SSb. Les calculs pour ce cas sont très similaires. La principale différence est qu'au lieu de calculer les différences entre une observation Y_{ik} et une moyenne de groupe $\overline{Y}_k$ pour toutes les observations, nous calculons les différences entre la moyenne de groupe $\overline{Y}_k$ et la moyenne générale $\overline{Y}$ (dans ce cas 0,88) pour tous les groupes.

group k	group mean $\bar{Y}_k$	grand mean $\bar{Y}$	deviation $\bar{Y}_k - \bar{Y}$	squared deviations $(\bar{Y}_k - \bar{Y})^2$
placebo	0.45	0.88	-0.43	0.19
anxifree	0.72	0.88	-0.16	0.03
joyzepam	1.48	0.88	0.60	0.36

Cependant, pour les calculs entre les groupes, nous devons multiplier chacun de ces écarts au carré par *Nk*, le nombre d'observations dans le groupe. Nous le faisons parce que chaque *observation* dans le groupe (tous les N_k d'entre eux) est associée à une différence entre les groupes. Donc, s'il y a six personnes dans le groupe placebo et que la moyenne du groupe placebo diffère de la moyenne générale de 0,19, alors la variation *totale* inter groupes associée à ces six personnes est *6 x 0,19=1.14*. Nous devons donc étendre notre petit tableau de calculs :

group k	... ...	squared deviations $(\bar{Y}_k - \bar{Y})^2$	sample size N_k	weighted squared dev $N_k(\bar{Y}_k - \bar{Y})^2$
placebo	...	0.19	6	1.14
anxifree	...	0.03	6	0.18
joyzepam	...	0.36	6	2.16

Ainsi, la somme des carrés inter groupes est obtenue en additionnant ces « écarts quadratiques pondérés » sur les trois groupes de l'étude :

$$\begin{aligned} \text{SS}_b &= 1{,}14 + 0{,}18 + 2{,}16 \\ &= 3{,}48 \end{aligned}$$

Comme vous pouvez le voir, les calculs entre les groupes sont beaucoup plus courts[114]. Maintenant que nous avons calculé nos sommes de valeurs des carrés SSb et SSw, le reste de l'ANOVA est assez indolore. L'étape suivante consiste à calculer les degrés de liberté. Puisque nous avons $G = 3$ groupes et $N = 18$ observations au total, nos degrés de liberté peuvent être calculés par simple soustraction :

$$\begin{aligned} \text{ddl}_b &= G - 1 = 2 \\ \text{ddl}_w &= N - G = 15 \end{aligned}$$

Ensuite, puisque nous avons maintenant calculé les valeurs des sommes des carrés et des degrés de liberté, tant pour la variabilité intra groupes que pour la variabilité inter groupes, nous pouvons obtenir les carrés moyens en divisant l'une par l'autre :

$$\text{MS}_b = \frac{\text{SS}_b}{\text{ddl}_b} = \frac{3{,}48}{2} = 1{,}74$$
$$\text{MS}_w = \frac{\text{SS}_w}{\text{ddl}_w} = \frac{1{,}39}{15} = 0{,}09$$

On a presque fini. Les carrés moyens peuvent être utilisés pour calculer la valeur *F*, qui est la statistique du test qui nous intéresse. Pour ce faire, nous divisons la valeur MS inter groupes par la valeur MS à intra groupes.

$$F = \frac{\text{MS}_b}{\text{MS}_w} = \frac{1{,}74}{0{,}09} = 19{,}3$$

Wouahou ! C'est terriblement excitant, non ? Maintenant que nous disposons de nos statistiques de test, la dernière étape consiste à déterminer si le test lui-même nous donne un résultat significatif. Comme nous l'avons vu au chapitre 9 autrefois, ce que nous ferions est d'ouvrir un manuel de statistiques ou de feuilleter l'annexe qui aurait en fait une énorme table de valeurs et nous trouverions la valeur seuil *F* correspondant à une valeur particulière d'alpha (la région de rejet d'hypothèse nulle), par exemple .05, .01 ou .001, pour 2 et 15 degrés de liberté. En procédant de cette façon, nous obtiendrions une valeur

---

[114] Dans le fichier Excel clinicaltrial_anova.xls, la valeur pour SS_b est très légèrement différente, 3,45, de celle indiquée dans le texte ci-dessus (erreurs d'arrondi !).

seuil *F* pour un alpha de .001 de 11,34. Comme cette valeur est inférieure à notre valeur *F* calculée, nous disons que *p<.001*. Mais c'était le bon vieux temps, et de nos jours le logiciel de statistiques sophistiqué calcule la *valeur p* exacte pour vous. En fait, la *valeur p* exacte est 0,000071. Donc, à moins que nous ne soyons *extrêmement* prudents quant à notre taux d'erreur de type I, nous sommes pratiquement certains de pouvoir rejeter l'hypothèse nulle.

C'est presque fini. Une fois nos calculs terminés, il est de tradition d'organiser tous ces chiffres dans un tableau ANOVA comme celui du Tableau 13-1. Pour les données de nos essais cliniques, le tableau ANOVA ressemblerait à ceci :

	df	sum of squares	mean squares	F-statistic	p-value
between groups	2	3.48	1.74	19.3	0.000071
within groups	15	1.39	0.09	-	-

De nos jours, vous n'aurez probablement jamais beaucoup de d'occasion de vouloir construire un de ces tableaux vous-même, mais vous constaterez que presque tous les logiciels statistiques (Jamovi inclus) ont tendance à organiser la sortie d'une ANOVA dans un tableau comme celui-ci, donc c'est une bonne idée de s'habituer à les lire. Cependant, bien que le logiciel produise un tableau d'ANOVA complète, il n'y a presque jamais une bonne raison d'inclure le tableau entier dans votre rédaction. Une façon assez standard de rapporter ce résultat serait d'écrire quelque chose comme ceci :

L'ANOVA à un facteur a montré un effet significatif du médicament sur l'amélioration de l'humeur (*F(2,15)=19,3,p<.001*).

Soupir ! Tant de travail pour une courte phrase.

Exécuter une ANOVA dans Jamovi

Je suis presque sûr que je sais ce que vous pensez après avoir lu la dernière section, *surtout* si vous avez suivi mes conseils et que vous avez fait tout cela vous-même au crayon et sur papier (c.-à-d. dans un tableur). Faire les calculs d'ANOVA soi-même, c'est nul. Il y a beaucoup de calculs à faire en cours de route, et il serait fastidieux de devoir le faire encore et encore chaque fois que vous voulez faire une analyse de variance.

Utiliser Jamovi pour spécifier votre ANOVA

Pour vous faciliter la vie, Jamovi peut faire une ANOVA…hourra ! Sélectionnez à l'analyse « ANOVA » -« ANOVA », et déplacez la variable mood.gain pour qu'elle se trouve dans la case « Dependent Variable», puis déplacez la variable drug pour qu'elle soit dans la case « Fixed Factors ». Ceci devrait donner les résultats comme le montre la Figure 13-3.[115] Notez que j'ai également coché la case η^2, prononcée « eta » au carré, sous l'option « Effect Size » et

[115] Les résultats de Jamovi sont plus précis que ceux du texte ci-dessus, en raison d'erreurs d'arrondi.

ceci est également indiqué dans le tableau des résultats. Nous reviendrons sur les tailles d'effet un peu plus tard.

ANOVA

ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2
drug	3.45	2	1.73	18.61	0.00009	0.71
Residuals	1.39	15	0.09			

Figure 13-3 : Tableau des résultats de Jamovi pour l'analyse de variance de l'humeur selon le médicament administré

Le tableau des résultats de Jamovi vous montre la somme des valeurs des carrés, les degrés de liberté, et d'autres quantités qui ne nous intéressent pas vraiment pour le moment. Notez, cependant, que Jamovi n'utilise pas les appellations inter groupe (between-group) et intra groupe (within-group). Au lieu de cela, il essaie d'attribuer des noms plus significatifs. Dans notre exemple particulier, la variance *entre les groupes* correspond à l'effet que le facteur drug a sur la variable de résultat, et la variance à l'*intérieur des groupes* correspond à la variabilité « résiduelle », c'est-à-dire la variance *résiduelle*. Si nous comparons ces chiffres à ceux que j'ai calculés à la main à la section 13.2.5, vous pouvez voir que ce sont plus ou moins les mêmes, abstraction faite des erreurs d'arrondi. La somme des carrés entre les groupes est SSb = 3,45, la somme des carrés au sein des groupes est SS_w= 1,39, et les degrés de liberté sont respectivement 2 et 15. Nous obtenons aussi la *valeur F* et la *valeur p* et, encore une fois, elles sont plus ou moins les mêmes, avec des erreurs d'arrondis par rapport aux chiffres que nous avons calculés nous-mêmes en faisant cela de manière longue et fastidieuse.

Taille de l'effet

Il y a différentes façons de mesurer la taille de l'effet dans une ANOVA, mais les mesures les plus couramment utilisées sont η^2 (**eta au carré**) et η^2 partiel. Pour une analyse de la variance à un facteur, ils sont identiques l'un à l'autre, alors pour le moment, je vais simplement expliquer η^2. La définition de η^2 est en fait très simple.

$$\eta^2 = \frac{SS_b}{SS_{tot}}$$

Il ne s'agit que de ça. Ainsi, quand je regarde le tableau d'ANOVA de la Figure 13-3, je vois que SS_b = 3,45 et SS_{tot} = 3,45 + 1,39 = 4,84. Nous obtenons alors une valeu η^2 de

$$\eta^2 = \frac{3,45}{4,84} = 0,71$$

L'interprétation de η^2 est tout aussi simple. Il s'agit de la proportion de la variabilité de la variable résultat (mood.gain) qui peut s'expliquer par le prédicteur (drug). Une valeur de η^2 = 0 signifie qu'il n'y a aucune relation entre les deux, alors qu'une valeur de η^2 = 1 signifie que la relation est parfaite. Mieux encore, la valeur de η^2 est très étroitement liée à R^2, comme nous l'avons vu à la section 12.6, et a une interprétation équivalente.

Bien que de nombreux manuels de statistiques suggèrent η^2 comme mesure par défaut de la taille de l'effet dans ANOVA, il y a un article intéressant de Daniel Lakens sur un blog qui suggère que l'eta-carré n'est peut-être pas la meilleure mesure de l'ampleur de l'effet dans le monde réel, car il peut être un estimateur biaisé[116]. Il est utile de noter qu'il existe également une option dans Jamovi pour spécifier omega-squared (ω^2), qui est moins biaisée, comparé à eta-carré.

Comparaisons multiples et tests post hoc

Chaque fois que vous exécutez une analyse de variance avec plus de deux groupes et que vous vous retrouvez avec un effet significatif, la première chose que vous voudrez probablement demander est quels groupes sont réellement différents les uns des autres. Dans notre exemple de médicaments, notre hypothèse nulle était que les trois médicaments (placebo, Anxifree et Joyzepam) ont exactement le même effet sur l'humeur. Mais si vous y réfléchissez bien, l'hypothèse nulle est en fait de prétendre *trois* choses différentes à la fois. Plus précisément, il prétend que :

- Le médicament de votre concurrent (Anxifree) ne vaut pas mieux qu'un placebo (c.-à-d. $\mu_A = \mu_P$)
- Votre médicament (Joyzepam) ne vaut pas mieux qu'un placebo (c.-à-d.$\mu_J = \mu_P$)
- Anxifree et Joyzepam sont également efficaces (c.-à-d. $\mu_J = \mu_A$)

Si l'une de ces trois affirmations est fausse, alors l'hypothèse nulle est également fausse. Donc, maintenant que nous avons rejeté notre hypothèse nulle, nous pensons qu'*au moins* une de ces choses n'est pas vraie. Mais lesquelles ? Ces trois propositions sont toutes trois intéressantes. Puisque vous voulez certainement savoir si votre nouveau médicament Joyzepam est meilleur qu'un placebo, il serait bien de savoir comment il se compare à une alternative commerciale existante (c.-à-d., Anxifree). Il serait même utile de vérifier la performance d'Anxifree par rapport au placebo. Même si Anxifree a déjà été testé à fond contre placebo par d'autres chercheurs, il peut quand même être très utile de vérifier que votre étude produit des résultats similaires à ceux des travaux antérieurs.

Lorsque nous caractérisons l'hypothèse nulle en fonction de ces trois propositions distinctes, il apparaît clairement qu'il y a huit « états du monde « possibles qu'il faut distinguer :

---

116 http://daniellakens.blogspot.com.au/2015/06/why-you-should-utilisation-omega-squared.html

possibility:	is $\mu_P = \mu_A$?	is $\mu_P = \mu_J$?	is $\mu_A = \mu_J$?	which hypothesis?
1	✓	✓	✓	null
2	✓	✓		alternative
3	✓		✓	alternative
4	✓			alternative
5		✓	✓	alternative
6		✓		alternative
7			✓	alternative
8				alternative

En rejetant l'hypothèse nulle, nous avons décidé que nous *ne croyons pas* que le #1 est le véritable état du monde. La prochaine question à se poser est la suivante : laquelle des sept autres possibilités nous semble la bonne ? Face à cette situation, il est généralement utile d'examiner les données. Par exemple, si nous examinons les graphiques de la Figure 13-1, il est tentant de conclure que le Joyzepam est meilleur que le placebo et meilleur qu'Anxifree, mais il n'y a pas de différence réelle entre Anxifree et le placebo. Cependant, si nous voulons obtenir une réponse plus claire à ce sujet, il pourrait être utile d'effectuer quelques tests.

Exécution de tests *t* par paire

Comment pourrions-nous résoudre notre problème ? Étant donné que nous avons trois paires distinctes de moyennes (placebo contre Anxifree, placebo contre Joyzepam et Anxifree contre Joyzepam) à comparer, nous pourrions faire trois tests *t* distincts et voir ce qui se passe. C'est facile à faire en Jamovi. Allez dans les options de l'ANOVA « Post-hoc Tests », déplacez la variable drug dans la case active sur la droite, puis cliquez sur la case « No correction ». On obtiendra ainsi un tableau clair montrant toutes les comparaisons du test *t* par paires entre les trois niveaux de la variable du médicament, comme le montre la Figure 13-4.

Post Hoc Tests

Post Hoc Comparisons - drug

Comparison									
drug		drug	Mean Difference	SE	df	t	p	$p_{bonferroni}$	p_{holm}
anxifree	-	joyzepam	−0.77	0.18	15.00	−4.36	0.00056	0.00168	0.00112
	-	placebo	0.27	0.18	15.00	1.52	0.15021	0.45064	0.15021
joyzepam	-	placebo	1.03	0.18	15.00	5.88	0.00003	0.00009	0.00009

Figure 13-4 : Tests t par pairs non corrigés comme comparaisons post hoc dans Jamovi

Corrections pour tests multiples

Dans la section précédente, j'ai laissé entendre qu'il y a un problème avec le fait de ne faire que des tas et des tas de tests *t*. Ce qui nous préoccupe, c'est que ce que nous faisons en effectuant ces analyses, c'est une « expédition de pêche ». Nous effectuons de nombreux tests sans trop d'orientation théorique, dans l'espoir que certains d'entre eux soient significatifs. Ce type de recherche sans théorie des différences de groupe qui est appelé **analyse post hoc** (« post hoc » en latin pour « à la suite de cela »).[117]

Il n'y a pas de mal à faire des analyses post hoc, mais il faut faire preuve de beaucoup de prudence. Par exemple, l'analyse que j'ai effectuée dans la section précédente devrait être évitée, car chaque *test t individuel* est conçu pour avoir un taux d'erreur de type I de 5 % (c.-à-d. $\alpha = .05$) et j'ai effectué trois de ces tests. Imaginez ce qui se serait passé si mon ANOVA avait impliqué 10 groupes différents, et que j'avais décidé de faire 45 *tests t* « post hoc » pour essayer de savoir lesquels étaient significativement différents les uns des autres, vous vous attendriez à ce que 2 ou 3 d'entre eux soient significatifs *par hasard seulement*. Comme nous l'avons vu au chapitre 9, le principe d'organisation central derrière les tests d'hypothèse nulle est que nous cherchons à contrôler notre taux d'erreur de type I, mais lorsque que j'exécute beaucoup de *tests t* à la fois afin de déterminer la source de mes résultats ANOVA, mon taux d'erreur réel de type I dans toute cette *famille de* tests est devenu complètement hors de contrôle.

La solution habituelle à ce problème est d'introduire un ajustement de la *valeur p*, qui vise à contrôler le taux d'erreur total dans la famille des tests (voir Shaffer, Shaffer (1995)). Un ajustement de cette forme, qui est habituellement (mais pas toujours) appliqué lorsqu'on qu'on fait une analyse post hoc, est souvent appelé une **correction pour des comparaisons multiples**, bien qu'on l'appelle parfois « inférence simultanée ». Quoi qu'il en soit, il existe plusieurs façons différentes de procéder à cet ajustement. J'en aborderai quelques-unes dans cette section et dans la section 14.8, mais vous devez savoir qu'il existe de nombreuses autres méthodes (voir, par exemple, Hsu, Hsu (1996)).

Corrections de Bonferroni

Le plus simple de ces ajustements s'appelle la **correction de Bonferroni** (Dunn 1961), et c'est très simple. Supposons que mon analyse post hoc consiste en *m* tests séparés, et que je veux m'assurer que la probabilité totale de faire des erreurs de type I est tout au plus α[118].

---

[117] Si vous avez une base théorique pour vouloir examiner certaines comparaisons, mais pas d'autres, c'est une autre histoire. Dans ces circonstances, vous n'effectuez pas du tout des analyses « post hoc », vous faites des « comparaisons planifiées ». Je parlerai de cette situation plus loin dans le livre (Section 14.9), mais pour l'instant, je veux que les choses restent simples.

[118] Il convient de noter au passage que toutes les méthodes d'ajustement n'essaient pas de le faire. Ce que j'ai décrit ici est une approche pour contrôler le « taux d'erreur de type I

Dans ce cas, la correction de Bonferroni dit simplement « multiplier toutes vos valeurs p brutes par m ». Si p indique la *valeur originale de p*, et si p'_j est la valeur corrigée, alors la correction de Bonferroni indique cela :

$$p'_j = m \times p$$

Ainsi, si vous utilisez la correction de Bonferroni, vous rejetteriez l'hypothèse nulle si $p' < \alpha$. La logique derrière cette correction est très simple. Nous faisons m tests différents, donc si nous l'organisons de telle sorte que chaque test ait un taux d'erreur de Type I d'au plus α/m, alors le taux d'erreur de Type I *total* de ces tests ne peut pas être supérieur à α. C'est assez simple, à tel point que dans le document original, l'auteur écrit,

> « La méthode donnée ici est si simple et si générale que je suis sûr qu'elle a dû être utilisée avant cela. Cependant, je ne la trouve pas et je ne peux donc que conclure que sa simplicité a peut-être empêché les statisticiens de se rendre compte qu'il s'agit d'une très bonne méthode dans certaines situations » (Dunn, Dunn (1961), p. 52-53).

Pour utiliser la correction de Bonferroni dans Jamovi, il suffit de cliquer sur la case à cocher « Bonferroni » dans les options « Correction », et vous verrez une autre colonne ajoutée au tableau des résultats ANOVA montrant les valeurs de p ajustées avec la correction de Bonferroni (Figure 13-4). Si nous comparons ces troi valeurs p à celles des tests t non corrigés par paires, il est clair que la seule chose que Jamovi a faite est de les multiplier par 3.

Corrections de Holm

Bien que la correction de Bonferroni soit l'ajustement le plus simple, ce n'est généralement pas le meilleur à utiliser. Une méthode qui est souvent utilisée à la place est la **correction de Holm** (Holm 1979). L'idée derrière la correction Holm est de prétendre que vous faites les tests séquentiellement, en commençant par la plus petite valeur de p (brute) et en passant à la plus grande. Pour la *j-ème* plus grande des valeurs de p, l'ajustement est soit

$$p'_j = j \times p_j$$

(c.-à-d. que la valeur de p la plus élevée reste inchangée, la deuxième *valeur de* p* la plus élevée est doublée, la troisième valeur de p la* plus élevée est triplée, et ainsi de suite), ou

$$p'_j = p'_{j+1}$$

le plus *élevé* des deux. Cela peut sembler un peu déroutant, alors passons les choses en revue un peu plus lentement. Voici ce que fait la correction Holm. Tout d'abord, vous triez toutes vos valeurs p dans l'ordre, de la plus petite à la plus grande. Pour la plus petite valeur p, il suffit de la multiplier par m, et c'est terminé. Cependant, pour tous les autres, il s'agit d'un processus en deux étapes. Par exemple, lorsque vous passez à la deuxième plus petite valeur p, vous la multipliez d'abord par *m-1.* Si cela produit un nombre plus grand que la

---

pour les familles ». Cependant, il existe d'autres tests post hoc qui visent à contrôler le « taux de fausses découvertes », ce qui est un peu différent.

valeur *p* ajustée que vous avez obtenue la dernière fois, alors vous le gardez. Mais si elle est plus petite que la dernière, alors vous copiez la dernière valeur de *p*. Pour illustrer comment cela fonctionne, considérons le tableau ci-dessous, qui montre les calculs d'une correction Holm pour une collection de cinq *valeurs p* :

raw p	rank j	$p \times j$	Holm p
.001	5	.005	.005
.005	4	.020	.020
.019	3	.057	.057
.022	2	.044	.057
.103	1	.103	.103

J'espère que cela clarifie les choses.

Bien qu'elle soit un peu plus difficile à calculer, la correction de Holm a de très bonnes propriétés. Il est plus puissant que Bonferroni (c.-à-d. qu'il a un taux d'erreur de type II plus faible) mais, aussi contre-intuitif que cela puisse paraître, il a le *même* taux d'erreur de type I. Par conséquent, en pratique, il n'y a jamais de raison d'utiliser la correction de Bonferroni plus simple puisqu'elle est toujours surpassée par la correction de Holm légèrement plus élaborée. Pour cette raison, la correction de Holm devrait vous permettre d'accéder à la correction des comparaisons multiples. La Figure 13-4 montre également les *valeurs p* corrigées de Holm et, comme vous pouvez le voir, la plus grande valeur* p* (correspondant à la comparaison entre Anxifree et le placebo) n'est pas modifiée. Avec une valeur de .15, c'est exactement la même valeur que celle que nous avons obtenue à l'origine lorsque nous n'avons appliqué aucune correction du tout. Par contre, la plus petite valeur *p* (Joyzepam contre placebo) a été multipliée par trois.

Rédaction du test post hoc

Enfin, après avoir effectué l'analyse post hoc pour déterminer quels groupes sont significativement différents les uns des autres, vous pouvez écrire le résultat comme ceci :

Des tests post hoc (utilisant la correction Holm pour ajuster *p*) ont indiqué que Joyzepam produisait un changement d'humeur significativement plus important que Anxifree (*p =.001*) et le placebo ($p = 9{,}0 \times 10^{-5}$). Nous n'avons trouvé aucune preuve qu'Anxifree ait donné de meilleurs résultats que le placebo (*p=.15*).

Ou, si vous n'aimez pas l'idée de rapporter des valeurs *p* exactes, alors vous changeriez ces chiffres pour *p<.01*, *p<.001* et *p<.05* respectivement. Quoi qu'il en soit, l'essentiel est que

vous indiquiez que vous avez utilisé la correction de Holm pour ajuster les valeurs *p*. Et bien sûr, je suppose qu'ailleurs dans le rapport, vous avez inclus les statistiques descriptives pertinentes (c.-à-d. les moyennes et les écarts-types des groupes), puisque ces valeurs *p* ne sont pas très informatives en soi.

Hypothèses de l'ANOVA à un facteur

Comme tout test statistique, l'analyse de la variance repose sur certaines hypothèses concernant les données, en particulier les résidus. Il y a trois hypothèses clés que vous devez connaître : la *normalité*, l'*homogénéité de la variance* et l'*indépendance*.

Si vous vous souvenez de la section 13.2.4, que j'espère que vous avez au moins survolé même si vous n'avez pas tout lu, j'ai décrit les modèles statistiques qui sous-tendent ANOVA de cette façon :

$$\begin{aligned} H_0: &\quad Y_{\text{ik}} = \mu + \epsilon_{\text{ik}} \\ H_1: &\quad Y_{\text{ik}} = \mu_k + \epsilon_{\text{ik}} \end{aligned}$$

Dans ces équations, μ fait référence à une seule moyenne de la population qui est la même pour tous les groupes, et μ_k est la moyenne de population pour le *k-ème* groupe. Jusqu'à présent, nous nous sommes surtout intéressés à savoir si nos données sont mieux décrites par la moyenne de la population (l'hypothèse nulle) ou par les différentes moyennes propres à chaque groupe (l'hypothèse alternative). C'est logique, bien sûr, car c'est en fait la question de recherche importante ! Toutefois, toutes nos méthodes de vérification se sont implicitement appuyées sur une hypothèse précise au sujet des résidus, ϵ_{ik} à savoir que

$$\epsilon_{\text{ik}} \sim Normal(0, \sigma^2)$$

Aucune des formules ne fonctionne correctement sans ce présupposé. Ou, pour être précis, vous pouvez toujours faire tous les calculs et vous obtiendrez une statistique *F*, mais vous n'avez aucune garantie que cette statistique *F* mesure réellement ce que vous pensez qu'elle mesure, et donc toute conclusion que vous pourriez tirer sur la base du test *F* pourrait être fausse.

Alors, comment vérifier si l'hypothèse sur les résidus est exacte ? Eh bien, comme je l'ai indiqué plus haut, il y a trois hypothèses distinctes cachées dans cee seul présupposé, et nous les examinerons séparément.

- **Homogénéité de la variance**. Notez que nous n'avons qu'une seule valeur pour l'écart-type de la population (c.-à-d. σ), plutôt que de permettre à chaque groupe d'avoir sa propre valeur (c.-à-d. σ_k). C'est ce qu'on appelle l'hypothèse d'homogénéité de la variance (parfois appelée homoscédasticité). L'analyse de variance suppose que l'écart-type de la population est le même pour tous les groupes. Nous en parlerons en détail à la section 13.6.1.

- **Normalité**. On suppose que les résidus sont normalement répartis. Comme nous l'avons vu à la section 11.8, nous pouvons l'évaluer en examinant les graphiques QQ (ou en effectuant un test Shapiro-Wilk). J'en parlerai davantage dans le contexte de l'analyse de variance à la section 13.6.4.

- **Indépendance**. L'hypothèse d'indépendance est un peu plus délicate. Ce que cela signifie essentiellement, c'est que le fait de connaître un résidu ne vous dit rien au sujet d'un autre résidu. Toutes les valeurs ϵ_{ik} sont supposées avoir été générées sans « égard » ou « relation » avec les autres. Il n'y a pas de façon évidente ou simple de tester cela, mais il y a certaines situations qui constituent des violations évidentes de cette règle. Par exemple, si vous avez un modèle à mesures répétées, où chaque participant à votre étude apparaît dans plus d'une condition, alors l'indépendance ne tient pas. Il y a une relation particulière entre certaines observations, à savoir celles qui correspondent à la même personne ! Lorsque cela se produit, vous devez utiliser l'ANOVA pour mesures répétées (voir Section 13.8).

Vérification de l'hypothèse d'homogénéité de la variance

Faire le test préliminaire sur les variances, c'est un peu comme prendre la mer à la rame pour savoir si les conditions sont suffisamment calmes pour qu'un paquebot quitte le port ! /- George Box (G. E. P. Box 1953)

Il y a plus d'une façon de dépecer un chat, comme on dit, et plus d'une façon de vérifier l'hypothèse d'homogénéité des variance (bien que, pour une raison inconnue, personne n'en ait parlé). Le test de **Levene** (Levene 1960) et le **test de Brown-Forsythe** (Brown and Forsythe 1974) sont les tests les plus couramment utilisés dans la littérature pour cela.

Que vous fassiez le test de Levene standard ou le test de Brown-Forsythe, la statistique du test, qui est parfois désignée par *F*, mais parfois aussi par *W*, est calculée exactement de la même manière que la statistique *F* de l'analyse de variance standard, en utilisant simplement un Z_{ik} plutôt qu'un Y_{ik}. En ayant cela à l'esprit, nous pouvons continuer à examiner comment faire le test avec Jamovi.

Le test de Levene est incroyablement simple. Supposons que nous ayons notre variable de résultat Yik. Tout ce que nous faisons est de définir une nouvelle variable, que j'appellerai Z_{ik}, correspondant à l'écart absolu par rapport à la moyenne du groupe

$$Z_{\text{ik}} = \left|Y_{\text{ik}} - \overline{Y}_k\right|$$

Bien, à quoi cela nous sert-il ? Prenons un moment pour réfléchir à ce qu'est réellement Z_{ik} et à ce que nous essayons de tester. La valeur de Z_{ik} est une mesure de la façon dont la i-ème observation dans le k-ème groupe s'écarte de sa moyenne de groupe. Et notre hypothèse nulle est que tous les groupes ont la même variance, c'est-à-dire les mêmes écarts globaux par rapport aux moyennes du groupe ! Ainsi, l'hypothèse nulle dans un test de Levene est que les moyennes de *Z* de la population sont identiques pour tous les groupes. Bien. Nous avons donc besoin maintenant d'un test statistique de l'hypothèse nulle que toutes les moyennes des groupes sont identiques. Où l'avons-nous déjà vu ? Ah oui, c'est l'ANOVA, et donc tout ce que fait le test de Levene, c'est d'exécuter une ANOVA sur la nouvelle variable Z_{ik}.

Et le test Brown-Forsythe ? Est-ce que cela fait quelque chose de particulièrement différent ? Non. Le seul changement par rapport au test de Levene est qu'il construit la variable transformée *Z* d'une manière légèrement différente, en utilisant des écarts par rapport aux médianes du groupe plutôt que des écarts par rapport aux moyennes du groupe. C'est-à-dire que nous avons, pour le test Brown-Forsythe

$$Z_{\text{ik}} = |Y_{\text{ik}} - \text{median}_k(Y)|$$

où médian$_k$(Y)est la médiane pour le groupe k.

Exécuter le test Levene dans Jamovi

Bien, alors comment fait-on le test Levene ? C'est très simple - sous l'option d'ANOVA « Assumption Checks », cliquez simplement sur la case à cocher « Homogeneity tests ». En regardant le résultat à la Figure 13-5, on constate que le test est non significatif ($F(2,15) = 1{,}45, p = .266$), il semble donc que l'hypothèse d'homogénéité de la variance soit bonne. Cependant, les apparences peuvent être trompeuses ! Si la taille de votre échantillon est assez grande, alors le test de Levene pourrait avoir un effet significatif (c.-à-d. $p < .05$) même si l'hypothèse d'homogénéité de la variance n'est pas trangressée ce qui peut nuire à la robustesse de l'ANOVA. C'est ce que George Box faisait valoir dans la citation ci-dessus. De même, si la taille de votre échantillon est assez petite, l'hypothèse d'homogénéité de la variance pourrait ne pas être satisfaite et pourtant un test de Levene pourrait être non significatif (c.-à-d. $p > .05$). Cela signifie que, parallèlement à tout test statistique de cette d'hypothèse vérifiée, vous devez toujours tracer l'écart-type pour chaque groupe ou catégorie de l'analyse… juste pour voir si elles sont assez semblables (c.-à-d. homogénéité de la variance) ou non.

Assumption Checks

Test for Homogeneity of Variances (Levene's)

F	df1	df2	p
1.45	2	15	0.26569

Figure 13-5 : Sortie de test Levene pour ANOVA unidirectionnelle in Jamovi

Vérification de l'hypothèse d'homogénéité des variances

Dans notre exemple, l'hypothèse d'homogénéité des variances s'est révélée assez sûre : le test de Levene s'est avéré non significatif (même si nous devrions aussi examiner le graphique des écarts-types), nous n'avons donc probablement pas à nous inquiéter. Cependant, dans la vraie vie, nous n'avons pas toujours cette chance. Comment sauver notre analyse de variance lorsque l'hypothèse d'homogénéité des variances n'est pas respectée ? Si vous vous souvenez de notre discussion sur les tests *t*, nous avons déjà vu ce problème auparavant. Le test *t* de Student suppose des variances égales, la solution consistait à utiliser le test *t* de Welch, si ce n'était pas le cas. En fait, Welch (1951) a également montré comment nous pouvons résoudre ce problème pour l'ANOVA aussi (le **test univarié de Welch**). Il est implémenté dans Jamovi dans « ANOVA One Way ». Il s'agit d'une approche d'analyse spécifique pour une analyse ANOVA à un facteur, et pour exécuter l'analyse ANOVA à un facteur de Welch pour notre exemple, nous réexécuterions l'analyse comme

précédemment, mais cette fois en utilisant la commande Jamovi « ANOVA One Way » analysis, et cocherions l'option pour le test de Welch (voir Figure 13-6).

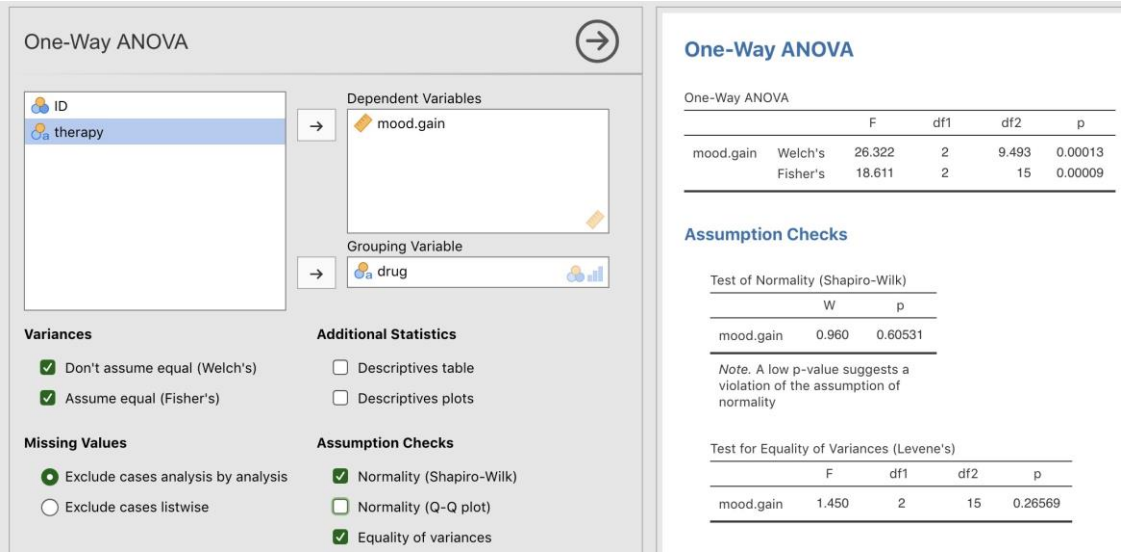


Figure 13-6 : Le test de Welch dans le cadre de l'analyse ANOVA à un facteur avec Jamovi

Pour comprendre ce qui se passe ici, comparons ces chiffres à ce que nous avons obtenu plus tôt dans la section 13.3 lorsque nous avons lancé notre analyse de variance initiale. Pour t'éviter d'avoir à te retourner, voilà ce qu'on a eu la dernière fois : $F(2,15) = 18.611, p = .00009$, également présenté comme le test de Fisher dans l'analyse de variance à un facteur présentée à la Figure 13-6.

Bien, au départ notre ANOVA nous a donné le résultat *F(2,15)=18,6*, alors que le test univarié de Welch nous a donné *F(2, 9,49)=26,32*. En d'autres termes, le test de Welch a réduit les degrés de liberté au sein des groupes de 15 à 9,49 et la valeur *F* est passée de 18,6 à 26,32.

Vérification de l'hypothèse de normalité

Il est relativement simple de tester l'hypothèse de normalité. Nous avons couvert la plupart de ce que vous devez savoir à la section 11.8. La seule chose que nous avons vraiment besoin de faire est de dessiner un tracé QQ et, en plus, s'il est disponible, de faire le test Shapiro-Wilk. Le tracé QQ est illustré à la Figure 13-7 me semble assez normal. Si le test de Shapiro-Wilk n'est pas significatif (c.-à-d.> .05), cela signifie que l'hypothèse de normalité n'est pas violée. Toutefois, comme dans le cas du test de Levene, si la taille de l'échantillon est importante, un test Shapiro-Wilk significatif peut être un faux positif, où l'hypothèse de normalité n'est pas violée de manière importante pour l'analyse. De même, un très petit échantillon peut produire de faux négatifs. C'est pourquoi une inspection visuelle du graphique QQ est importante.

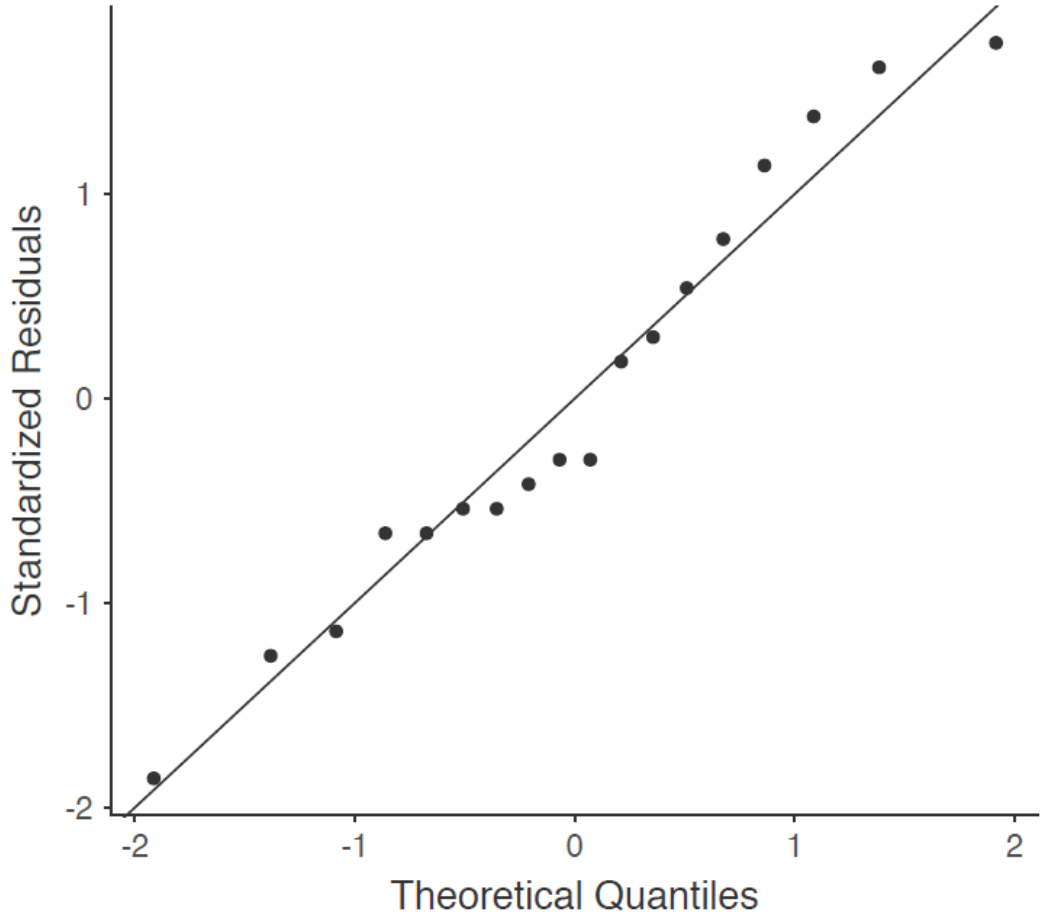


Figure 13-7 : Graphique QQ produit à partir des options de l'ANOVA à un facteur de Jamovi

En plus de l'inspection du graphique QQ pour détecter tout écart par rapport à la normale, le test Shapiro-Wilk pour nos données montre un effet non significatif, avec *p=.6053* (voir Figure 13-6). Cela confirme donc l'analyse du graphique QQ ; les deux contrôles n'indiquent pas que l'hypothèse normalité a été violée.

Suppression de l'hypothèse de normalité

Maintenant que nous avons vu comment vérifier la normalité, nous sommes naturellement amenés à nous demander ce que nous pouvons faire pour remédier aux violations de la normalité. Dans le contexte d'une analyse de variance à un facteur, la solution la plus simple est probablement de passer à un test non paramétrique (c.-à-d. un test qui ne repose sur aucune hypothèse particulière quant au type de distribution en cause). Nous avons déjà vu des tests non paramétriques au chapitre 11. Lorsque vous n'avez que deux groupes, le test de Mann-Whitney ou le test de Wilcoxon fournissent l'alternative non paramétrique dont vous avez besoin. Lorsque vous avez trois groupes ou plus, vous pouvez utiliser le **test de la somme de rang Kruskal-Wallis** (Kruskal and Wallis 1952). C'est le test dont nous parlerons plus tard.

La logique derrière le test Kruskal-Wallis

Le test Kruskal-Wallis est étonnamment similaire à l'ANOVA, à certains égards. Dans l'ANOVA, nous avons commencé par Y_{ik}, la valeur de la variable de résultat pour la ième personne du groupe k. Pour le test Kruskal Wallis, nous classerons toutes ces valeurs Y_{ik} et effectuerons notre analyse sur les données classées.

Posons que R_{ik} désigne le classement donné au ième membre du *k-ième* groupe. Maintenant, calculons $\overline{R}_k$, le rang moyen donné aux observations dans le *k-ième* groupe

$$\overline{R}_k = \frac{1}{N_k} \sum_i R_{\text{ik}}$$

et calculons aussi $\overline{R}$, le rang moyen général

$$\overline{R} = \frac{1}{N_k} \sum_i \sum_k R_{\text{ik}}$$

Maintenant que nous avons fait cela, nous pouvons calculer les écarts quadratiques par rapport au rang moyen général $\overline{R}$. Lorsque nous le faisons pour les scores individuels, c'est-à-dire si nous calculons $\left(R_{\text{ik}} - \overline{R}\right)^2$, nous obtenons une mesure « non paramétrique » de l'écart entre la ik-ième observation et le rang moyen général. Lorsque nous calculons l'écart quadratique de la moyenne du groupe par rapport à la grande moyenne, c'est-à-dire si nous calculons $\left(\overline{R}_k - \overline{R}\right)^2$, nous obtenons une mesure non paramétrique de l'écart entre la moyenne du groupe et la grande moyenne du rang. En gardant cela à l'esprit, nous suivrons la même logique qu'avec ANOVA et définirons nos sommes de carrés de *mesures ordonnées*, un peu comme nous l'avons fait avec ANOVA plus tôt. Tout d'abord, nous avons notre « total des sommes des carrés ordonnés ».

$$\text{RSS}_{\text{tot}} = \sum_k \sum_i \left(R_{\text{ik}} - \overline{R}\right)^2$$

et on peut définir ainsi les « sommes de carrés ordonnés entre groupes » comme suit

$$\begin{aligned} \text{RSS}_b &= \sum_k \sum_i \left(\overline{R}_{\text{k}} - \overline{R}\right)^2 \\ &= \sum_k N_k \left(\overline{R}_{\text{k}} - \overline{R}\right)^2 \end{aligned}$$

Si l'hypothèse nulle est vraie et qu'il n'y a aucune différence réelle entre les groupes, on s'attendrait à ce que les sommes RSS_b entre les groupes soient très petites, beaucoup plus petites que le total des sommes RSS_{tot}. Qualitativement, c'est à peu près la même chose que ce que nous avons trouvé lorsque nous avons construit la *statistique F* de l'ANOVA, mais pour des raisons techniques, la statistique du test Kruskal-Wallis, habituellement appelée K, est construite d'une manière légèrement différente,

$$K = (N-1) \times \frac{\text{RSS}_b}{\text{RSS}_{\text{tot}}}$$

et si l'hypothèse nulle est vraie, alors la distribution d'échantillonnage de *K* est *approximativement* celle de Khi carré avec G-1 degrés de liberté (où *G* est le nombre de groupes). Plus la valeur de *K* est élevée, moins les données sont cohérentes avec l'hypothèse nulle, c'est donc un test unilatéral. Nous rejetons H_0 lorsque *K* est suffisamment grand.

Détails supplémentaires

La description dans la section précédente illustre la logique derrière le test Kruskal-Wallis. Sur le plan conceptuel, c'est la bonne façon de penser au fonctionnement du test. Cependant, d'un point de vue purement mathématique, c'est inutilement compliqué. Je ne vais pas vous montrer la démonstration, mais vous pouvez utiliser un peu de tambouille algébrique[119] pour montrer que l'équation pour *K* peut être réécrite comme suit

$$K = \frac{12}{N(N-1)} \sum_k N_k \overline{R}_k^2 - 3(N+1)$$

C'est cette dernière équation que vous voyez parfois donnée pour *K*. C'est beaucoup plus facile à calculer que la version que j'ai décrite dans la section précédente, seulement c'est totalement dénué de sens pour les humains. Il est probablement préférable de penser à *K* comme je l'ai décrit plus tôt, comme un analogue de l'ANOVA basé sur les rangs. Mais gardez à l'esprit que la statistique du test qui est calculée finit par avoir un aspect assez différent de celui que nous avons utilisé pour notre ANOVA originale.

Mais ce n'est pas tout ! Mon Dieu, pourquoi y a-t-il toujours quelque chose en *plus* ? L'histoire que j'ai racontée jusqu'à présent n'est vraie que lorsqu'il n'y a aucun lien entre les données brutes. C'est-à-dire, s'il n'y a pas deux observations qui ont exactement la même valeur. S'il y a des liens, nous devons introduire un facteur de correction dans ces calculs. À ce stade, je suppose que même le lecteur le plus diligent a cessé de s'en soucier (ou du moins s'est formé l'opinion que le facteur de correction d'égalité est quelque chose qui ne nécessite pas leur attention immédiate). Je vais donc vous dire très rapidement comment il est calculé, et omettre les détails fastidieux des *raisons* pour lesquelles s'est ainsi. Supposons que nous construisons un tableau de fréquence pour les données brutes, et f_j est le nombre d'observations qui ont la *j-ème* valeur unique. Cela peut sembler un peu abstrait, alors voici un exemple concret tiré du tableau de fréquence de mood.gain de l'ensemble de données clinicaltrials.csv :

0.1	0.2	0.3	0.4	0.5	0.6	0.8	0.9	1.1	1.2	1.3	1.4	1.7	1.8
1	1	2	1	1	2	1	1	1	1	2	2	1	1

[119] Un terme technique

En regardant ce tableau, notez que la troisième entrée du tableau de fréquence a une valeur de 2, ce qui correspond à un mood.gain de 0,3, ce tableau nous indique que l'humeur de deux personnes a augmenté de 0,3. Plus précisément, dans la notation mathématique que j'ai présentée ci-dessus, cela nous dit que $f_3=2$. Bien, maintenant que nous le savons, le facteur de correction des liens (TCF) est :

$$TCF = 1 - \frac{\sum_j f_j^3 - f_j}{N^3 - N}$$

La valeur corrigée de la statistique de Kruskal-Wallis est obtenue en divisant la valeur de K par cette quantité. C'est cette version à ex-aequo corrigés que Jamovi calcule. Enfin, nous en avons fini avec la théorie du test Kruskal-Wallis. Je suis sûr que vous êtes tous terriblement soulagés que je vous aie guéri de l'anxiété existentielle qui surgit naturellement lorsque vous réalisez que vous ne savez pas comment calculer le facteur de correction d'égalité pour le test Kruskal-Wallis. N'est-ce pas ?

Comment exécuter le test Kruskal-Wallis avec Jamovi

Malgré l'horreur que nous avons vécue en essayant de comprendre ce que fait réellement le test Kruskal-Wallis, il s'avère que l'exécution du test est plutôt indolore, puisque Jamovi a une analyse dans le cadre de l'ensemble d'analyse ANOVA appelé « Non-Parametric » - « One Way ANOVA (Kruskall Wallis) » La plupart du temps, vous aurez les données du type clinicaltrial.csv, où vous avez vos résultats comme la variable mood.gain et des groupes comme la variable drug. Si c'est le cas, vous pouvez procéder à l'analyse avec Jamovi. Voici ce que nous donne est un Kruskal-Wallis $\chi^2 = 12{,}076, \text{df} = 2, \text{p} - \text{value} = 0{,}00239$, comme dans la figure 13.8

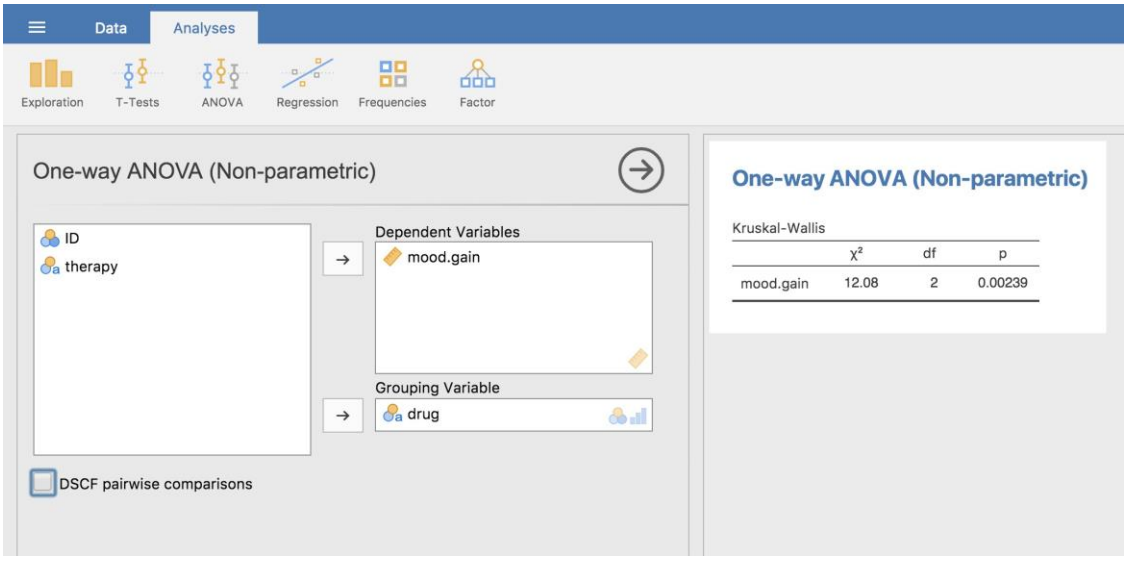


Figure 13-8 : Test non paramétrique d'ANOVA à un facteur Kruskall-Wallis avec Jamovi

ANOVA à un facteur pour mesures appariées

Le test ANOVA à un facteur pour mesures appariée est une méthode statistique de vérification des différences significatives entre trois groupes ou plus où les mêmes participants sont utilisés dans chaque groupe (ou chaque participant est étroitement apparié aux participants d'autres groupes expérimentaux). Pour cette raison, il devrait toujours y avoir un nombre égal de scores (données) dans chaque groupe expérimental. Ce type de conception et d'analyse peut aussi être appelé une « analyse de variance appariée » ou une « analyse de variance intra sujets).

La logique qui sous-tend une analyse de variance pour mesures appariées est très semblable à celle d'une analyse de variance pour groupes indépendants (parfois appelée analyse de variance « inter sujets »). Vous vous souvenez que plus tôt nous avons montré que la variabilité totale de l'ANOVA est divisée en variabilité inter groupes (SS_b) et en variabilité intra groupes (SS_w), et qu'après chacun est divisé par les degrés de liberté respectifs pour donner MS_b et MS_w (voir tableau 13.1) le rapport *F* est calculé de la manière suivante :

$$F = \frac{MS_b}{MS_w}$$

Dans une analyse de variance pour mesures appariées, le rapport F est calculé de la même manière, mais alors que dans une analyse de variance pour groupes indépendants, la variabilité intragroupe (SS_w) sert de dénominateur à la MS_w, dans une analyse de variance pour mesures appariées, la SSw est répartie en deux parties. Comme nous utilisons les mêmes sujets dans chaque groupe, nous pouvons supprimer la variabilité due aux différences individuelles entre les sujets, appelées $SS_{subjects}$, de la variabilité intra groupe. Nous n'entrerons pas trop dans les détails techniques sur la façon dont cela se fait, mais essentiellement chaque sujet devient un niveau d'un facteur appelé sujets. La variabilité de ce facteur intra-sujets est ensuite calculée de la même manière que tout facteur inter-sujets. Ensuite, nous pouvons soustraire $SS_{subjects}$ de SS_w pour obtenir un terme SS_{error} plus petit :

- ANOVA pour groupes indépendants : $SS_{error} = SS_w$
- ANOVA pour mesures appariées : $SS_{error} = SS_w - SS_{subjects}$

Ce changement du terme SSerror conduit souvent à un test statistique plus puissant, mais cela dépend du fait que la réduction du SS_{error} compense largement la réduction des degrés de liberté du valeur d'erreur (les degrés de liberté passant de *(n - k)*[120] à *(n - 1)(k - 1)* (en se rappelant que le plan pour l'ANOVA sur des groupes indépendants contient plus de sujets).

ANOVA pour mesures appariées avec Jamovi

D'abord, il nous faut des données. Geschwind (1972) a suggéré que la nature exacte du déficit du langage d'un patient à la suite d'un AVC peut être utilisée pour diagnostiquer la

---

[120] *(n - k)* : (nombre de sujets - nombre de groupes)

région spécifique du cerveau qui a été endommagée. Un chercheur s'intéresse à l'identification des difficultés de communication spécifiques éprouvées par six patients souffrant de l'aphasie de Broca (un déficit du langage couramment ressenti à la suite d'un AVC).

Tableau 13-2 : Nombre de tentatives réussies pour trois tâches expérimentales

Participant	Discours	Conceptuel	Syntaxe
1	8	7	6
2	7	8	6
3	9	5	3
4	5	4	5
5	6	6	2
6	8	7	4

Les patients devaient effectuer trois tâches de reconnaissance de mots. Lors de la première tâche (production de la parole), les patients devaient répéter des mots simples lus à haute voix par le chercheur. Pour la deuxième tâche (conceptuelle), destinée à tester la compréhension des mots, les patients devaient faire correspondre une série d'images avec leur nom correct. Pour la troisième tâche (syntaxique), conçue pour tester la connaissance de l'ordre correct des mots, on a demandé aux patients de réorganiser les phrases syntaxiquement incorrectes. Chaque patient a accompli les trois tâches. L'ordre dans lequel les patients ont tenté les tâches était contrebalancé entre les participants. Chaque tâche consistait en une série de 10 essais. Le nombre d'essai effectués avec succès par chaque patient est indiqué au Tableau 13-2. Entrez ces données dans Jamovi pour l'analyse (ou prenez un raccourci et chargez le fichier broca.csv).

Pour effectuer une ANOVA à un facteur dans Jamovi, ouvrez la boîte de dialogue ANOVA, comme dans la Figure 13-9, via ANOVA - Repeated Measures ANOVA. Puis :

- Entrez un nom de facteur de mesures appariées. Il s'agit d'une étiquette que vous choisissez pour décrire les conditions répétées pour tous les participants. Par exemple, pour décrire les tâches vocales, conceptuelles et syntaxiques effectuées par tous les participants, une étiquette appropriée serait « Tâche ». Notez que ce nouveau nom de facteur représente la variable indépendante dans l'analyse.
- Ajoutez un troisième niveau dans la zone de texte « Repeated Measures Factors », car il y a trois niveaux représentant les trois tâches : parole, concept et syntaxe. Modifiez les désignations des niveaux en conséquence.
- Ensuite, déplacez chacune des variables de niveau dans la zone de texte « Repeated Measures Cells ».

- Enfin, sous l'option « Assumption Checks », cochez la case « Sphericity checks ».

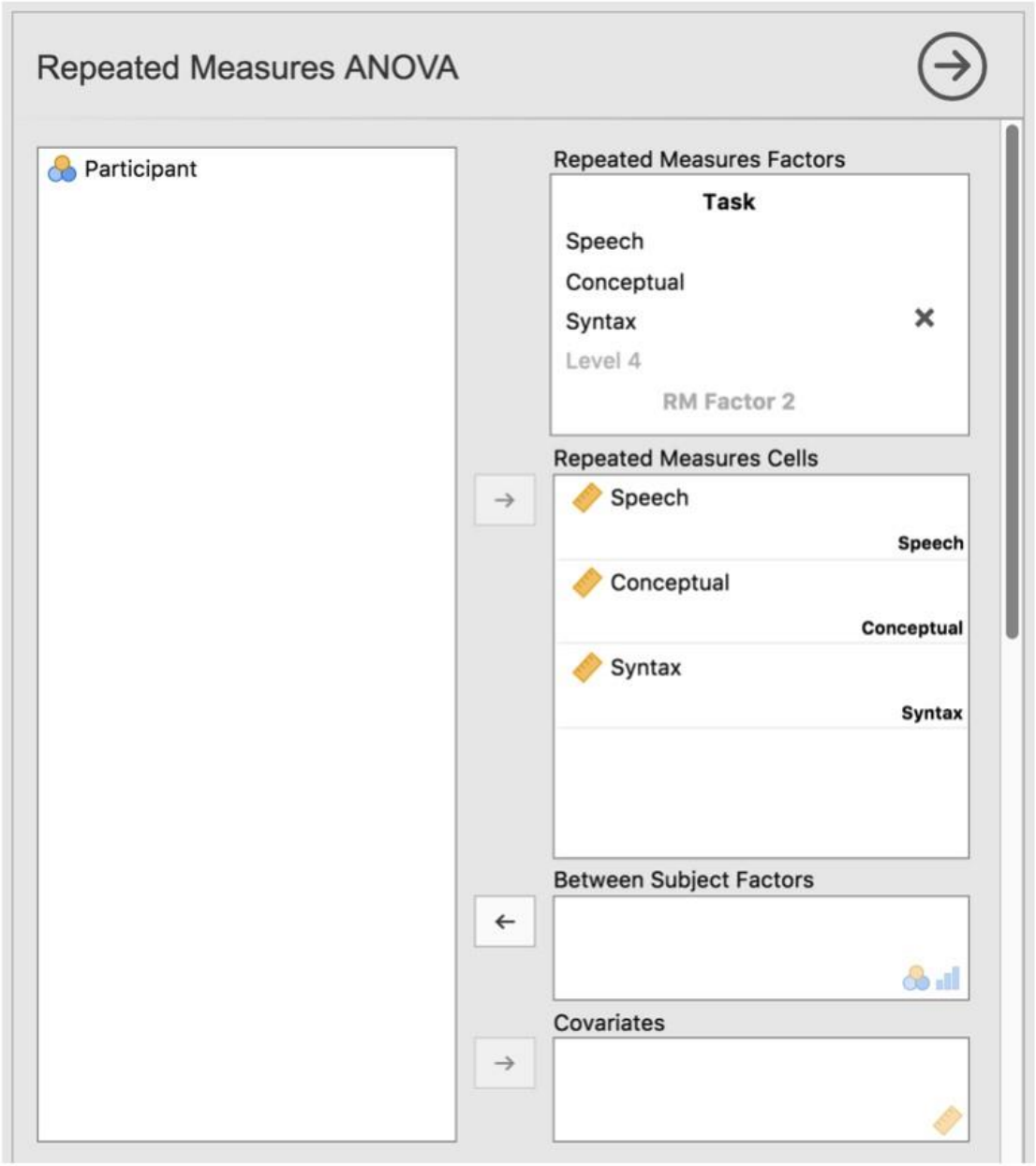


Figure 13-9 : Boîte de dialogue ANOVA pour mesures appariées avec Jamovi

La sortie Jamovi d'une ANOVA à un facteur pour mesures appariées est conformes aux Figure 13-1 à Figure 13-13. Le premier résultat que nous devrions examiner est le Test de Mauchly de la sphéricité, qui teste l'hypothèse selon laquelle les variances des différences entre les conditions sont égales (ce qui signifie que l'écart entre les scores des différences entre les conditions de l'étude est approximativement le même). Dans la Figure 13-10, le niveau de signification du test de Mauchly est $p = .720$. Si le test de Mauchly n'est pas significatif (c.-à-d. >.05, comme c'est le cas dans la présente analyse), il est raisonnable de

conclure que les variances des différences ne sont pas significativement différentes (c.-à-d. qu'elles sont à peu près égales ou que l'on peut supposer la sphéricité).

Tests of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Task	0.85	0.72009	0.87	1.00

Figure 13-10 : résultats de l'ANOVA à un facteur pour mesures appariées : Test de sphéricité de Mauchly

Si, par contre, le test de Mauchly est significatif (*p.< .05*), nous concluons qu'il y a des différences significatives entre la variance des différences et que l'exigence de sphéricité n'a pas été satisfaite. Dans ce cas, nous devons appliquer une correction à la valeur *F* obtenue dans l'analyse ANOVA à un facteur.

Si la valeur Greenhouse-Geisser dans le tableau « Tests de sphéricité » est *> .75*, vous devrez utiliser la correction de Huynh-Feldt. Si la valeur Greenhouse-Geisser est *< .75*, vous devrez utiliser la correction Greenhouse-Geisser. Ces deux valeurs *F* corrigées peuvent être spécifiées dans les cases à cocher « Sphericity Corrections » sous les options « Assumption Checks », les valeurs *F* corrigées sont ensuite affichées dans le tableau des résultats, comme dans la Figure 13-11.

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Task	None	24.78	2	12.39	6.93	0.01296
	Greenhouse-Geisser	24.78	1.74	14.26	6.93	0.01802
	Huynh-Feldt	24.78	2.00	12.39	6.93	0.01296
Residual	None	17.89	10	1.79		
	Greenhouse-Geisser	17.89	8.68	2.06		
	Huynh-Feldt	17.89	10.00	1.79		

Note. Type 3 Sums of Squares

Figure 13-11 : Mesure répétée unidirectionnelle de la sortie ANOVA : Tests des effets à l'intérieur des sujets

Dans notre analyse, nous avons vu que la significativité du Test de Mauchly de la sphéricité était *p=.720* (c.-à-d. *p>.05*). Cela signifie donc que nous pouvons supposer que l'exigence de sphéricité a été satisfaite et qu'aucune correction de la valeur *F* n'est nécessaire. Par conséquent, nous pouvons cocher l'option de correction de sphéricité « None » pour la mesure répétée « Tâche » : *F= 6,93, df=2, p= .013*, et nous pouvons conclure que le nombre

de tests effectués avec succès pour chaque tâche linguistique variait considérablement selon que la tâche était basée sur la parole, la compréhension ou la syntaxe (*F(2,10)=6.93,p=.013*).

Des tests post-hoc peuvent également être spécifiés dans Jamovi pour les mesures appariées ANOVA de la même manière que pour l'ANOVA sur des groupes indépendants. Les résultats sont présentés à la Figure 13-12. Ceux-ci indiquent qu'il existe une différence significative entre la parole et la syntaxe, mais pas entre les autres niveaux.

Post Hoc Comparisons - Task

Comparison							
Task		Task	Mean Difference	SE	df	t	ptukey
Speech	-	Conceptual	1.00	0.77	10.00	1.29	0.42929
	-	Syntax	2.83	0.77	10.00	3.67	0.01097
Conceptual	-	Syntax	1.83	0.77	10.00	2.37	0.09064

Figure 13-12 : Tests post-hoc pour une ANOVA pour des mesures appariées avec Jamovi

Les statistiques descriptives (moyennes marginales) peuvent être examinées pour aider à interpréter les résultats, produits dans la sortie Jamovi comme dans la Figure 13-13. La comparaison du nombre moyen d'essais menés avec succès par les participants montre que les Aphasiques de Broca ont d'assez bons résultats dans les domaines suivants la production de la parole (*moyenne= 7,17*) et la compréhension du langage (*moyenne= 6,17*). Cependant, leur performance était considérablement plus mauvaise sur la tâche syntaxique (*moyenne = 4,33*), avec une différence significative aux tests post-hoc entre la parole et la performance de la tâche syntaxique.

Estimated Marginal Means - Task

			95% Confidence Interval	
Task	Mean	SE	Lower	Upper
Speech	7.17	0.62	5.82	8.51
Conceptual	6.17	0.62	4.82	7.51
Syntax	4.33	0.62	2.99	5.68

Figure 13-13 : Statistiques descriptives pour l'ANOVA à un facteur pour mesure appariées :

Le test non paramétrique de Friedman

Le test de Friedman est une version non paramétrique d'une analyse de variance pour mesures appariées et peut remplacer le test de Kruskall-Wallis lorsqu'il s'agit de déterminer les différences entre trois groupes ou plus où les mêmes participants font partie de chaque groupe, ou lorsque chaque participant est étroitement apparié avec des participants d'autres conditions. Si la variable dépendante est ordinale ou si l'hypothèse de normalité n'est pas respectée, on peut utiliser le test de Friedman.

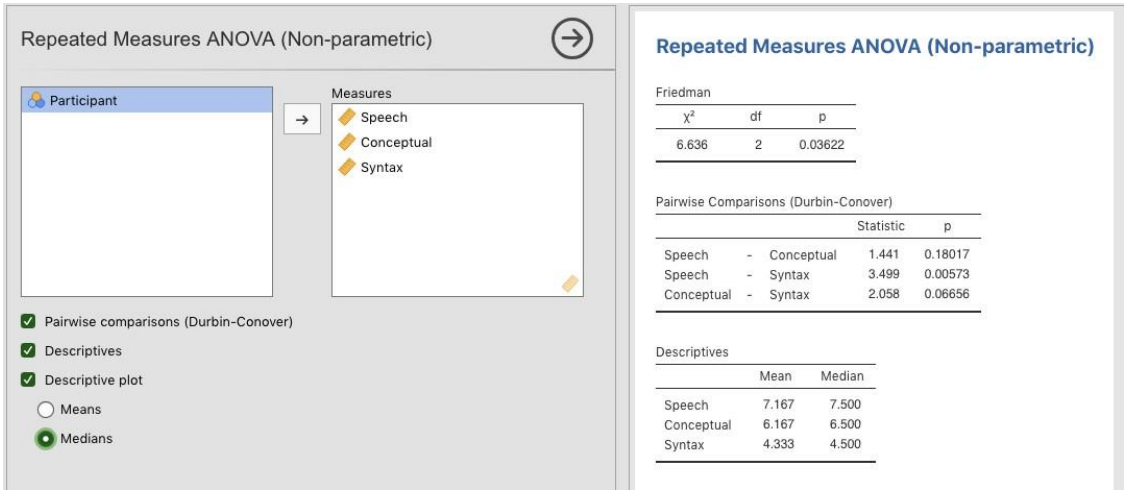


Figure 13-14 : La boîte de dialogue « Repeated Measures ANOVA (Non-parametric) » dans Jamovi

Comme pour le test Kruskall-Wallis, les mathématiques sous-jacentes sont compliquées et ne seront pas présentées ici. Pour les besoins de ce livre, il suffit de noter que Jamovi calcule la version corrigée du test de Friedman dont la Figure 13-14 est un exemple utilisant les données sur l'aphasie de Broca que nous avons déjà examinées en utilisant une analyse de variance à un facteur paramétrique.

Il est assez simple de faire un test de Friedman avec Jamovi. Sélectionnez simplement « Analyses - ANOVA - Repeated Measures ANOVA (Non-parametric) », comme dans la Figure 13-14. Ensuite, sélectionnez et transférez les variables de mesures appariées que vous souhaitez comparer (Parole, Conceptuel, Syntaxe) dans la zone de texte « Measures : ». Pour produire des statistiques descriptives (moyennes et médianes) pour les trois variables de mesures appariés, cliquez sur le bouton « Descriptives ».

Les résultats de Jamovi montrent des statistiques descriptives, la valeur du chi carré, les degrés de liberté et la valeur *p* (Figure 13-14). Puisque la valeur *p* est inférieure au niveau conventionnellement utilisé pour déterminer la signification ($p < .05$), nous pouvons conclure que les Aphasiques de Broca donnent de bons résultats dans les tâches de production de la parole (*médiane=7,5*) et de compréhension du langage (*médiane=6,5*). Cependant, leur performance était considérablement plus mauvaise sur la tâche syntaxique (*médiane=4.5*), avec une différence significative aux tests post-hoc entre la parole et la performance de la tâche syntaxique.

Sur la relation entre ANOVA et le *test* t de Student

Il y a une dernière chose que je tiens à souligner avant de terminer. C'est quelque chose que beaucoup de gens trouvent un peu surprenant, mais cela vaut la peine d'en savoir plus. Une ANOVA sur deux groupes est identique au *t de Student*. Ils ne sont pas seulement semblables, en fait ils sont équivalents en tous points de vue. Je n'essaierai pas de prouver que c'est toujours vrai, mais je vais vous fournir une seule démonstration concrète. Supposons qu'au lieu d'exécuter une analyse de variance sur notre modèle mood.gain ~ drug, nous utilisons plutôt la therapy comme prédicteur. Si nous exécutons cette ANOVA, nous obtenons une statistique *F* de *F(1,16)=1,71*, et une valeur *p= 0,21*. Comme nous n'avons que deux groupes, je n'ai pas besoin de recourir à une ANOVA, j'aurais pu simplement décider de faire un *t de Student*. Voyons ce qui se passe quand je fais ça : J'obtiens une statistique *t* de *t(16)=-1,3068* et une valeur *p=0.21*. Curieusement, les *valeurs p* sont identiques. Une fois de plus, on obtient une valeur de $p = .21$. Mais qu'en est-il de la statistique du test ? Après avoir effectué un test *t* au lieu d'une ANOVA, nous obtenons une réponse quelque peu différente, à savoir *t(16)= -1,3068*. Cependant, il existe une relation assez simple avec F. Si nous élevons au carré la statistique *t*, nous obtenons la statistique *F* précédente : $-1,3068^2 = 1,7077$

Résumé

Il y a beaucoup de choses dans ce chapitre, mais il en manque encore beaucoup. De toute évidence, je n'ai pas encore discuté d'un analogue du *test t pour des* échantillons appariés pour plus de deux groupes. Il y a une façon de le faire, connue sous le nom de *ANOVA pour mesures répétées*, qui apparaîtra plus tard. Je n'ai pas non plus discuté de la façon d'exécuter une analyse de variance lorsque vous vous intéressez à plus d'une variable de regroupement, mais cela sera discuté en détail au chapitre 14. Pour ce que nous avons présenté, les sujets clés étaient :

- La logique de base du fonctionnement de l'ANOVA (Section 13.2) et la façon de l'exécuter avec Jamovi (Section 13.3).
- Comment calculer la taille de l'effet d'une analyse de variance (section 13.4)
- Analyse post hoc et corrections pour les tests multiples (section 13.5).
- Les hypothèses sous-jacents à l'ANOVA (section 13.6).
- Comment vérifier l'hypothèse d'homogénéité de la variance (section 13.6.1).
- Comment vérifier l'hypothèse de normalité (Section 13.6.4) et que faire si elle est violée (Section 13.7).
- ANOVA pour mesures appariées (section 13.8) et l'équivalent non paramétrique, le test de Friedman (section 13.9).

Comme pour tous les chapitres de ce livre, il y a un certain nombre de sources différentes sur lesquelles je me suis appuyé, mais le texte qui m'a le plus influencé est Sahai et Ageel

(2000). Ce n'est pas un bon livre pour les débutants, mais c'est un excellent livre pour les lecteurs plus avancés qui sont intéressés à comprendre les fondements mathématiques de l'ANOVA.

ANOVA Factorielle

Au cours des derniers chapitres, nous avons fait pas mal de choses. Nous avons examiné les tests statistiques que vous pouvez utiliser lorsque vous avez une variable prédictive nominale à deux groupes (c.-à-d. le *test t*, chapitre 11) ou à trois groupes ou plus (p. ex. ANOVA à un facteur, chapitre 13). Le chapitre sur la régression (chapitre 12) a introduit une nouvelle idée puissante, à savoir la construction de modèles statistiques avec *de multiples* variables prédicteurs continues utilisées pour expliquer une seule variable résultat. Par exemple, un modèle de régression pourrait être utilisé pour prédire le nombre d'erreurs qu'un élève commet dans un test de compréhension de la lecture en fonction du nombre d'heures qu'il a étudiées pour le test et de son résultat à un test de QI normalisé.

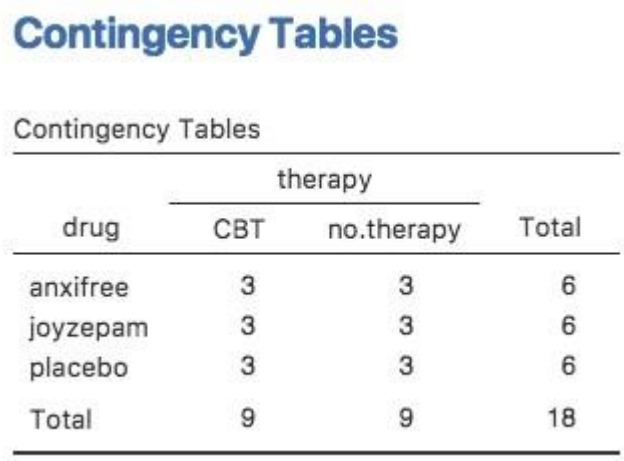
Le but de ce chapitre est d'étendre l'idée d'utiliser plusieurs variables prédictrices dans le cadre de l'analyse de variance. Supposons, par exemple, que nous voulions utiliser le test de compréhension de la lecture pour mesurer le rendement des élèves dans trois écoles différentes, et que nous soupçonnons que les filles et les garçons se développent à des rythmes différents (et que l'on s'attendrait donc à ce qu'ils aient en moyenne des performances différentes). Chaque élève est classé de deux façons différentes : en fonction de son sexe et en fonction de son école. Ce que nous aimerions faire, c'est d'analyser les résultats de compréhension de la lecture en fonction de ces *deux* variables de regroupement. L'outil pour ce faire est appelé **ANOVA factorielle**. Toutefois, comme nous avons deux variables de regroupement, nous l'appelons parfois analyse de variance bifactorielle, contrairement aux analyses de variance à un facteur que nous avons effectuées au chapitre 13.

ANOVA Factorielle 1 : des plans équilibrés, pas d'interactions

Lorsque nous avons discuté de l'analyse de la variance au chapitre 13, nous avons supposé un plan expérimental assez simple. Chaque personne fait partie d'un groupe parmi d'autres et nous voulons savoir si ces groupes ont des scores moyens différents pour une variable résultats. Dans la présente section, je traiterai d'une catégorie plus large de plans expérimentaux appelés **plans factoriels**, dans lesquels nous avons plus d'une variable de groupement. J'ai donné ci-dessus un exemple de la façon dont ce genre de plan pourrait être réalisé. Nous avons vu un autre cas de figure au chapitre 13, dans lequel nous examinions l'effet de différents médicaments sur le l'amélioration de l'humeur ressenti par chaque personne. Dans ce chapitre, nous avons trouvé un effet important du médicament, mais à la fin du chapitre, nous avons également effectué une analyse pour voir s'il y avait un effet de la thérapie. Nous n'en avons pas trouvé, mais il y a quelque chose d'un peu inquiétant à essayer d'effectuer deux analyses *distinctes pour* prédire la même variable résultat. Peut-être qu'il y *a* un effet de la thérapie sur le gain d'humeur, mais nous n'avons pas pu le trouver parce qu'il était « caché » par l'effet du médicament? En d'autres termes, nous voulons effectuer une analyse *unique* qui inclut à la *fois le* médicament et la thérapie comme

prédicteurs. Pour cette analyse, chaque personne est classée selon le médicament qu'on lui a administré (un facteur à 3 niveaux) et la thérapie qu'elle a reçue (un facteur à 2 niveaux). C'est ce que nous appelons un plan factoriel 3 x 2.

Si l'on croise les données sur les médicaments par traitement, en utilisant l'analyse « Frequencies » - « Contingency Tables » de Jamovi (voir la section 6.1), on obtient le tableau présenté à la Figure 14-1.



Contingency Tables

Contingency Tables

	therapy		
drug	CBT	no.therapy	Total
anxifree	3	3	6
joyzepam	3	3	6
placebo	3	3	6
Total	9	9	18

Figure 14-1 : Tableau de contingence Jamovi du drug par therapy

Comme vous pouvez le constater, non seulement nous avons des participants correspondant à toutes les combinaisons possibles des deux facteurs, ce qui indique que notre plan est **complètement croisé**[121], mais il s'avère qu'il y a un nombre égal de personnes dans chaque groupe. En d'autres termes, nous avons un plan **équilibré**. Dans cette section, je parlerai de la façon d'analyser les données à partir de plans équilibrés, puisque c'est le cas le plus simple. L'histoire des plans déséquilibrées est assez fastidieuse, nous allons donc la mettre de côté pour l'instant.

Quelles hypothèses vérifions-nous ?

Comme l'analyse de variance à un facteur, l'analyse de variance factorielle est un outil permettant de tester certains types d'hypothèses sur les moyennes de population. Un bon point de départ serait donc d'être explicite sur ce que sont réellement nos hypothèses. Cependant, avant même d'en arriver là, il est vraiment utile d'avoir une notation claire et simple pour décrire les moyennes de la population. Étant donné que les observations sont classées selon deux facteurs différents, il existe un grand nombre de moyennes auxquels on peut s'intéresser. Pour voir cela, commençons par penser à tous les échantillons que l'on

[121] Ndt. Ici les auteurs considèrent la relation entre les facteurs et non celle avec les sujets. Les sujets sont bien emboités dans le croisement des deux facteurs.

peut calculer pour ce type de plan. Tout d'abord, il y a l'idée évidente que nous pourrions nous intéresser à cette liste de moyennes de groupe :

drug	therapy	mood.gain
placebo	no.therapy	0.300000
anxifree	no.therapy	0.400000
joyzepam	no.therapy	1.466667
placebo	CBT	0.600000
anxifree	CBT	1.033333
joyzepam	CBT	1.500000

Ce tableau présente une liste des moyennes de groupe pour toutes les combinaisons possibles des deux facteurs (p. ex. les personnes qui ont reçu le placebo et aucune thérapie, les personnes qui ont reçu le placebo en recevant la TCC (=CBT), etc.) Il est utile d'organiser tous ces chiffres, ainsi que les moyennes marginales et générales, dans un tableau unique qui ressemble à celui-ci :

	no therapy	CBT	total
placebo	0.30	0.60	0.45
anxifree	0.40	1.03	0.72
joyzepam	1.47	1.50	1.48
total	0.72	1.04	0.88

Chacune de ces moyennes correspond bien sûr à un échantillon de statistiques. C'est une quantité qui se rapporte aux observations précises que nous avons faites dans notre étude. Ce que nous voulons déduire, ce sont les paramètres correspondants de la population. C'est-à-dire, les vraies moyennes telles qu'elles existent au sein d'une population plus large. Ces moyennes de population peuvent aussi être organisées dans un tableau similaire, mais nous aurons besoin d'une petite notation mathématique pour le faire. Comme d'habitude, j'utiliserai le symbole μ pour désigner une moyenne de population. Cependant, parce qu'il y a beaucoup de moyennes différentes, j'utiliserai des indices pour les distinguer.

Voici comment fonctionne la notation. Notre tableau est défini en fonction de deux facteurs. Chaque ligne correspond à un niveau différent de facteur A (dans ce cas-ci, drug) et chaque colonne correspond à un niveau différent de facteur B (dans ce cas, therapy). Si nous

indiquons par R, le nombre de lignes dans le tableau, et *C,* le nombre de colonnes, nous pouvons nous référer à cela comme une ANOVA factorielle *RxC.* Dans ce cas, $R = 3$ et $C = 2$. Nous utiliserons des lettres minuscules pour faire référence à des lignes et colonnes spécifiques, de sorte que μrc se réfère à la moyenne de population associée au *r-ième* niveau du facteur A (c'est-à-dire le numéro de ligne *r*) et au *c-ième* niveau du facteur B (numéro de colonne *c*).[122] Ainsi, les moyennes de la population sont maintenant écrites comme ceci :

	no therapy	CBT	total
placebo	μ_{11}	μ_{12}	
anxifree	μ_{21}	μ_{22}	
joyzepam	μ_{31}	μ_{32}	
total			

Bien, qu'en est-il des cases restantes ? Par exemple, comment décrire le gain d'humeur moyen dans l'ensemble de la population (hypothétique) qui pourraient recevoir du Joyzepam dans le cadre d'une expérience comme celle-ci, qu'elles aient été ou non en thérapie ? Nous utilisons la notation « point » pour l'exprimer. Dans le cas de Joyzepam, notez qu'il s'agit de la moyenne associée à la troisième ligne du tableau. C'est-à-dire que nous calculons la moyenne sur deux moyennes cellules (c.-à-d. $\mu 31$ et $\mu 32$). Le résultat de ce calcul de la moyenne est appelé **moyenne marginale** et, dans ce cas, il s'agit de $\mu 3.$. La moyenne marginale de la TCC correspond à la moyenne de la population associée à la deuxième colonne du tableau ; nous utilisons donc la notation $\mu.2$ pour la désigner. La moyenne générale est désignée par $\mu..$ parce qu'il s'agit de la moyenne obtenue en moyennant (en marginalisant[123]) sur les deux. Ainsi, notre tableau complet des moyennes de population peut être écrit de la façon suivante :

---

[122] Ce qu'il y a de bien avec la notation par indice, c'est qu'elle généralise bien. Si notre expérience avait impliqué un troisième facteur, nous pourrions simplement ajouter un troisième indice. En principe, la notation s'étend à autant de facteurs que vous voudrez bien inclure, mais dans ce livre, nous considérerons rarement des analyses impliquant plus de deux facteurs, et jamais plus de trois.

[123] Techniquement, la marginalisation n'est pas tout à fait identique à une moyenne régulière. Il s'agit d'une moyenne pondérée qui tient compte de la fréquence des différents événements sur lesquels vous faites la moyenne. Cependant, dans un plan équilibré, toutes les fréquences de nos cellules sont égales par définition, de sorte que les deux sont équivalentes. Nous discuterons des plans déséquilibrés plus tard, et lorsque nous le ferons, vous verrez que tous nos calculs deviennent un véritable casse-tête. Mais ignorons cela pour l'instant.

	no therapy	CBT	total
placebo	μ_{11}	μ_{12}	$\mu_{1.}$
anxifree	μ_{21}	μ_{22}	$\mu_{2.}$
joyzepam	μ_{31}	μ_{32}	$\mu_{3.}$
total	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

Maintenant que nous avons cette notation, il est facile de formuler et d'exprimer certaines hypothèses. Supposons que le but est de découvrir deux choses. Premièrement, le choix du médicament a-t-il un effet sur l'humeur ? Deuxièmement, la TCC a-t-elle un effet sur l'humeur ? Ce ne sont pas les seules hypothèses que nous pourrions formuler, bien sûr, et nous verrons un exemple très important d'un autre type d'hypothèse à la section 14.2, mais ce sont les deux hypothèses les plus simples à vérifier, et nous allons donc commencer par là. Considérez le premier test. Si le médicament n'a pas d'effet, on s'attendrait à ce que tous les moyennes de la rangée soient identiques, n'est-ce pas ? Voilà donc notre hypothèse nulle. D'un autre côté, si le médicament a de l'importance, il faut s'attendre à ce que les moyens de cette rangée soient différents. Formellement, nous écrivons nos hypothèses nulles et alternatives en termes d'*égalité des moyennes marginales* :

Hypothèse nulle, H_0: Les moyennes en ligne sont les mêmes, c.-à-d. $\mu_1 = \mu_2 = \mu_3$ Hypothèse alternative, H_1 : La moyenne d'au moins une ligne est différente.

Il convient de noter qu'il s'agit *exactement des* mêmes hypothèses statistiques que celles que nous avons formulées lorsque nous avons effectué une analyse de variance à un facteur sur ces données au chapitre 13. A l'époque, j'utilisais la notation μ_P pour faire référence au gain d'humeur moyen pour le groupe placebo, μ_A et μ_J correspondant à la moyenne du groupe pour les deux médicaments, et l'hypothèse nulle était $\mu_P = \mu_A = \mu_J$. Nous parlons donc en fait de la même hypothèse, c'est juste que l'analyse de variance plus compliquée exige une notation plus prudente en raison de la présence de multiples variables de groupement, c'est pourquoi nous parlons maintenant de cette hypothèse comme de$\mu_{.1} = \mu_{.2} = \mu_{.3}$. Cependant, comme nous le verrons plus loin, bien que l'hypothèse soit identique, le test de cette hypothèse est subtilement différent du fait que nous reconnaissons maintenant l'existence de la deuxième variable de groupement.

En parlant de l'autre variable de groupement, vous ne serez pas surpris de découvrir que notre deuxième test d'hypothèse est formulé de la même façon. Cependant, puisqu'il s'agit de la thérapie psychologique plutôt que du médicament, notre hypothèse nulle correspond maintenant à l'égalité des moyennes en colonne :

Hypothèse nulle, H_0: Les moyennes en colonne sont les mêmes, c.-à-d. $\mu_{.1} = \mu_{.2}$
Hypothèse alternative, H_1: Les moyennes des colonnes sont différentes, c-à-d. $\mu_{.1} \neq \mu_{.2}$

Réalisation de l'analyse dans Jamovi

Les hypothèses nulles et alternatives que j'ai décrites dans la dernière section devraient vous sembler terriblement familières. Il s'agit essentiellement des mêmes hypothèses que celles que nous avons testées dans nos analyses de variance à un facteur plus simples au chapitre 13. Vous vous attendez donc probablement à ce que les tests d'hypothèse utilisés dans l'analyse de variance factorielle soient essentiellement les mêmes que le test *F* du chapitre 13. Vous vous attendez à voir des références à des sommes de carrés (SS), des carrés moyens (MS), des degrés de liberté (df), et finalement une statistique *F* que nous pouvons convertir en une valeur *p*, n'est-ce pas ? Eh bien, vous avez tout à fait raison. A tel point que je vais m'écarter de mon approche habituelle. Tout au long de ce livre, j'ai généralement pris l'approche de décrire la logique (et dans une certaine mesure les aspects mathématiques) qui sous-tend une analyse particulière d'abord et seulement ensuite introduire l'analyse dans Jamovi. Cette fois, je vais le faire dans l'autre sens et vous montrer comment le faire en Jamovi d'abord. La raison en est que je veux souligner les similitudes entre le simple outil ANOVA à un facteur dont nous avons parlé au chapitre 13, et l'approche plus complexe que nous allons utiliser dans ce chapitre.

Si les données que vous essayez d'analyser correspondent à un plan factoriel équilibré, l'analyse de la variance est facile. Pour voir à quel point c'est facile, commençons par reproduire l'analyse originale du chapitre 13.Au cas où vous l'auriez oublié, pour cette analyse, nous n'utilisions qu'un seul facteur (c.-à-d. drug) pour prédire notre variable de résultat (c.-à-d. mood.gain), et nous avons obtenu les résultats présentés à la Figure 14-2.

ANOVA

	Sum of Squares	df	Mean Square	F	p
drug	3.45	2	1.73	18.61	0.00009
Residuals	1.39	15	0.09		

Figure 14-2 : ANOVA à un facteur de mood.gain par drug avec Jamovi

Maintenant, supposons que je sois aussi curieux de savoir si la therapy a une relation avec le mood.gain. A la lumière de ce que nous avons vu dans notre discussion sur la régression multiple au chapitre 12, vous ne serez probablement pas surpris qu'il nous suffise d'ajouter la therapy comme deuxième « Fixed Factor » dans l'analyse, voir la Figure 14-3.

Cette sortie est assez simple à lire aussi. La première ligne du tableau indique la somme des carrés (SS) entre les groupes associés au facteur drug, ainsi que la valeur correspondante aux *df* inter groupes. Il calcule également un carré moyen (MS), une statistique *F* et une valeur *p*.

ANOVA

	Sum of Squares	df	Mean Square	F	p	η²	η²p	ω²
drug	3.45	2	1.73	26.15	0.00002	0.71	0.79	0.68
therapy	0.47	1	0.47	7.08	0.01866	0.10	0.34	0.08
Residuals	0.92	14	0.07					

Figure 14-3 Anova à deux facteurs dans Jamovi pour mood.gain par drug et therapy

Nous avons également une ligne correspondant au facteur therapy et une ligne correspondant aux résidus (c.-à-d. la variation intragroupe).

Non seulement toutes les quantités individuelles sont assez familières, mais les relations entre ces différentes quantités sont restées inchangées, tout comme nous l'avons vu avec l'ANOVA à un facteur. Notez que le carré moyen est calculée en divisant SS par le *df* correspondant. C'est-à-dire qu'il est toujours vrai que qu'il s'agisse de drug, de therapy ou des résidus.

$$MS = \frac{\text{SS}}{\text{df}}$$

Pour le voir, ne nous inquiétons pas de la façon dont les sommes des carrés sont calculées. Au lieu de cela, prenons pour acquis que Jamovi a calculé correctement les valeurs SS, et essayons de vérifier que tous les autres nombres ont un sens. Tout d'abord, notons que pour le facteur drug, si on divise 3,45 par 2 et on obtient un carré moyen de 1,73. Pour le facteur therapy, il n'y a qu'un seul degré de liberté, donc nos calculs sont encore plus simples : diviser 0,47 (la valeur SS) par 1 nous donne un résultat de 0,47 (la valeur de MS).

En ce qui concerne les statistiques *F* et les valeurs *p*, notons que nous en avons deux de chaque, l'une correspondant au facteur drug et l'autre au facteur therapy. Peu importe de laquelle il s'agit, la statistique *F* est calculée en divisant le carrée moyen associé au facteur par le valeur carré moyen associé aux résidus. Si nous utilisons « A » comme notation abrégée pour désigner le premier facteur (facteur A ; dans ce cas, drug) et « R » comme notation abrégée pour désigner les résidus, alors la statistique *F* associée au facteur A est appelée F_A, et est calculée comme suit :

$$F_A = \frac{\text{MS}_A}{\text{MS}_R}$$

et une formule équivalente existe pour le facteur B (c.-à-d. therapy). Notez que cette utilisation de « R » pour parler des résidus est un peu gênante, puisque nous avons également utilisé la lettre R pour faire référence au nombre de lignes dans le tableau, mais j'utiliserai « R » seulement pour désigner les résidus dans le contexte de SSR et MSR, donc j'espère que cela ne sera pas trop confus. Quoi qu'il en soit, pour appliquer cette formule au facteur drug, on prend le carré moyen de 1,73 et on le divise par le carré moyen résiduel de 0,07, ce qui nous donne une statistique *F* de 26,15. Le calcul correspondant pour la variable therapy serait de diviser 0,47 par 0,07, ce qui donne 7,08 pour la statistique *F*. Il n'est pas surprenant, bien sûr, que ces valeurs soient les mêmes que celles que Jamovi a rapportées dans le tableau ANOVA ci-dessus.

Le tableau ANOVA contient également le calcul des valeurs *p.* Encore une fois, il n'y a rien de nouveau ici. Pour chacun de nos deux facteurs, nous essayons de tester l'hypothèse nulle qu'il n'y a pas de relation entre le facteur et la variable résultat (je serai un peu plus précis à ce sujet plus loin). Pour ce faire, nous avons (apparemment) suivi une stratégie similaire à ce que nous avons fait dans le cadre d'ANOVA et nous avons calculé une statistique *F* pour chacune de ces hypothèses. Pour convertir ces valeurs en *p*, il suffit de noter que la distribution d'échantillonnage pour la statistique *F* sous l'hypothèse nulle (que le facteur en question n'est pas pertinent) est une distribution *F*. Notez également que les deux degrés de liberté sont ceux correspondant au facteur et aux résidus. Pour le facteur drug, il s'agit d'une distribution *F* avec 2 et 14 degrés de liberté (je reviendrai sur les degrés de liberté plus en détail plus loin). En revanche, pour le facteur therapy, la distribution d'échantillonnage est *F* avec 1 et 14 degrés de liberté.

À ce stade, j'espère que vous pouvez voir que le tableau ANOVA pour cette analyse factorielle plus complexe devrait être lu de la même façon que le tableau ANOVA pour l'analyse à un facteur plus simple. Bref, il nous dit que l'analyse de variance factorielle pour notre plan 3x2 a permis de trouver un effet significatif du médicament ($F(2,14) = 26,15$, $p < .001$) ainsi qu'un effet significatif du traitement ($F(1,14) = 7,08$, $p = .02$). Ou, pour utiliser la terminologie techniquement plus correcte, nous dirions qu'il y a deux **effets principaux** du médicament et de la thérapie. Pour l'instant, il semble probablement un peu redondant de parler d'effets « principaux », mais cela a du sens. Plus tard, nous parlerons de la possibilité « d'interactions » entre les deux facteurs, et nous ferons donc généralement une distinction entre les effets principaux et les effets d'interaction.

Comment la somme des carrés est-elle calculée ?

Dans la section précédente, j'avais deux objectifs. Tout d'abord, pour vous montrer que la méthode Jamovi nécessaire pour faire l'ANOVA factorielle est à peu près la même que celle que nous avons utilisée pour une ANOVA à un facteur. La seule différence est l'ajout d'un deuxième facteur. Deuxièmement, je voulais vous montrer à quoi ressemble le tableau ANOVA dans ce cas, afin que vous puissiez voir d'emblée que la logique et la structure de base de l'ANOVA factorielle sont les mêmes que celles qui sous-tendent l'ANOVA à un facteur. Essayez de vous accrocher à cette idée. C'est tout à fait vrai, dans la mesure où l'ANOVA factorielle est construite plus ou moins de la même manière que le modèle ANOVA à un facteur plus simple. C'est juste que ce sentiment de familiarité commence à s'évaporer une fois que vous commencez à creuser les détails. Traditionnellement, cette sensation réconfortante est remplacée par un besoin irrépressible de maltraiter les auteurs des manuels de statistiques.

Bien, commençons par examiner certains de ces détails. L'explication que j'ai donnée dans la dernière section illustre le fait que les tests d'hypothèse pour les principaux effets (du médicament et de la thérapie dans ce cas) sont des *tests F*, mais ce qu'il ne fait pas, c'est vous montrer comment la somme des valeurs des carrés (SS) est calculée. Il ne vous dit pas non plus explicitement comment calculer les degrés de liberté (valeurs *df*) bien que ce soit une chose simple en comparaison. Supposons pour l'instant que nous n'ayons que deux variables prédictrices, le facteur A et le facteur B. Si nous utilisons *Y* pour nous désigner la variable de résultat, nous utiliserions Y_{rci} pour nous parler du résultat associé au *i-ième*

membre du groupe *rc* (c.-à-d. niveau/ligne *r* pour le facteur A et niveau/colonne *c* pour le facteur B). Ainsi, si l'on utilise $\overline{Y}$ pour se référer à une moyenne d'échantillon, on peut utiliser la même notation que précédemment pour se référer aux moyennes de groupe, aux moyennes marginales et aux grandes moyennes. C'est-à-dire que $\overline{Y}_{rc}$ est la moyenne de l'échantillon associée au *r-ième* niveau du facteur A et le *c-ième* niveau du facteur B, $\overline{Y}_{r.}$ serait la moyenne marginale du *r-ième* niveau du facteur A, $\overline{Y}_{.c}$ serait la moyenne marginale du *c-ième* niveau du facteur B, et $\overline{Y}_{..}$ est la moyenne générale. En d'autres termes, les moyennes de notre échantillon peuvent être organisées dans le même tableau que les moyennes de population. Pour les données de nos essais cliniques, ce tableau ressemble à ceci :

	no therapy	CBT	total
placebo	$\bar{Y}_{11}$	$\bar{Y}_{12}$	$\bar{Y}_{1.}$
anxifree	$\bar{Y}_{21}$	$\bar{Y}_{22}$	$\bar{Y}_{2.}$
joyzepam	$\bar{Y}_{31}$	$\bar{Y}_{32}$	$\bar{Y}_{3.}$
total	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{..}$

Et si nous regardons les moyennes de l'échantillon que j'ai montré plus tôt, nous avons $\overline{Y}_{11} = 0{,}30$,$\overline{Y}_{12} = 0{,}60$ etc. Dans notre exemple d'essai clinique, le facteur drug a 3 niveaux et le facteur therapy a 2 niveaux, et ce que nous essayons d'exécuter est une ANOVA factorielle 3 x 2. Cependant, pour être un peu plus général, disons que le Facteur A (le facteur de ligne) a *R* niveaux et que le Facteur B (le facteur de colonne) a *C* niveaux , et donc ce que nous faisons ici est une ANOVA factorielle *R x C*.

Maintenant que nous avons rectifié notre notation, nous pouvons calculer la somme des carrés pour chacun des deux facteurs d'une manière relativement familière. Pour le facteur A, la somme des carrés entre les groupes est calculée en évaluant dans quelle mesure les moyennes marginales (ligne) $\overline{Y}_{1.}$, $\overline{Y}_{2.}$ etc., sont différentes de la moyenne générale$\overline{Y}_{..}$. Nous procédons de la même manière que pour l'analyse de variance à sens unique : nous calculons la somme de la différence au carré entre les valeurs de $\overline{Y}_{i.}$ et de $\overline{Y}_{..}$. Plus précisément, s'il y a *N* personnes dans chaque groupe, alors nous calculons ceci

$$\mathrm{SS}_A = (N \times C) \sum_{r=1}^{R} \left(\overline{Y}_{\mathrm{r.}} - \overline{Y}_{..}\right)^2$$

Comme pour l'ANOVA à un facteur, la partie la plus intéressante[124] de cette formule est $\left(\overline{Y}_{r.} - \overline{Y}_{..}\right)^2$, qui correspond à l'écart quadratique associé au niveau *r*. Tout ce que fait cette formule est de calculer cet écart au carré pour tous les niveaux *R* du facteur, de les additionner, puis de multiplier le résultat par *N x C*. La raison de cette dernière partie est qu'il y a plusieurs cellules dans notre plan qui ont le niveau *r* du facteur A. En fait, il y en a *C*, une pour chaque niveau possible du facteur B ! Ainsi, dans notre exemple, il y a deux cellules différentes dans le plan correspondant au médicament anxifree : une pour les personnes no therapy et une pour le groupe CBT. De plus, à l'intérieur de chacune de ces cellules, il y a *N* observations. Ainsi, si nous voulons convertir notre valeur SS en une quantité qui détermine la somme des carrés entre les groupes pour « chaque observation », nous devons multiplier par *N x C*. La formule pour le facteur B est bien sûr la même, mais avec des indices remplacés.

$$\mathrm{SS}_B = (N \times R) \sum_{c=1}^{C} \left(\overline{Y}_{.\mathrm{c}} - \overline{Y}_{..}\right)^2$$

Maintenant que nous disposons de ces formules, nous pouvons les comparer à la sortie Jamovi du fichier section précédente. Une fois de plus, un tableur est utile pour ce genre de calculs, alors n'hésitez pas à vous lancer. Vous pouvez également consulter la version que j'ai faite dans Excel dans le fichier clinicaltrial_factorialanova.xls.

Tout d'abord, calculons la somme des carrés associés à l'effet principal de la variable drug. Il y a un total de *N* = 3 personnes dans chaque groupe et *C* = 2 types de thérapie différents. Ou, pour le dire autrement, il y a 3 x 2 = 6 personnes qui ont reçu un médicament en particulier. Lorsque nous faisons ces calculs dans un tableur, nous obtenons une valeur de 3,45 pour la somme des carrés associés à l'effet principal de drug. Il n'est donc pas surprenant que ce chiffre soit le même que celui que vous obtenez lorsque vous recherchez la valeur SS pour le facteur drug dans le tableau ANOVA que j'ai présenté plus tôt, à la Figure 14-3.

Nous pouvons répéter le même type de calcul pour l'effet de la thérapie. Encore une fois il y a *N* = 3 personnes dans chaque groupe, mais puisqu'il y a *R* = 3 médicaments différents, cette fois-ci on note qu'il y a 3 X 3 = 9 personnes qui ont reçu la CBT et 9 autres personnes qui ont reçu le placebo. Ainsi, notre calcul dans ce cas nous donne une valeur de 0,47 pour la somme des carrés associés à l'effet principal de therapy. Encore une fois, nous ne sommes pas surpris de constater que nos calculs sont identiques à ceux de l'analyse de variance dans la Figure 14-3.

C'est donc ainsi que vous calculez les valeurs SS pour les deux effets principaux. Ces valeurs SS sont analogues à la somme des valeurs des carrés entre les groupes que nous avons

---

[124] Traduction langage courant: « la moins ennuyeuse ».

calculées lors de l'analyse de variance à sens unique au chapitre 13. Cependant, ce n'est plus une bonne idée de les considérer comme des valeurs SS inter groupes, simplement parce que nous avons deux variables de groupement différentes et qu'il est facile de se tromper. Cependant, pour construire un test *F*, nous devons également calculer la somme des carrés à l'intérieur d'un groupe. Conformément à la terminologie que nous avons utilisée dans le chapitre sur la régression (chapitre 12) et à la terminologie utilisée par Jamovi lors de création du tableau ANOVA, je commencerai par faire référence à la valeur SS à l'intérieur des groupes comme la somme *résiduelle* des carrés SSR.

La façon la plus simple de comprendre les valeurs résiduelles de la SS dans ce contexte, je pense, est de d'imaginer qu'il s'agit de la variation résiduelle de la variable résultats après avoir pris en compte les différences dans les moyennes marginales (c.-à-d. après avoir enlevé le SSA et le SSB). Ce que je veux dire par là, c'est que nous pouvons commencer par calculer la somme totale des carrés, que j'appellerai SST. La formule est à peu près la même que pour l'ANOVA à un facteur. Nous prenons la différence entre chaque observation *Yrci* et la grande moyenne

$$\overline{Y}_{..}$$

.Elevez au carré les différences et additionnez-les toutes.

$$\mathrm{SS}_T = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{i=1}^{N} (Y_{\mathrm{rci}} - \overline{Y}_{..})^2$$

La « triple sommation » semble ici plus compliquée qu'elle ne l'est. Dans les deux premières sommations, nous additionnons tous les niveaux du facteur A (c.-à-d. toutes les lignes *r* possibles de notre tableau) et tous les niveaux du facteur B (*c.*-à-d. toutes les colonnes *c* possibles). Chaque combinaison *rc* correspond à un seul groupe et chaque groupe contient *N* personnes, nous devons donc faire la somme de toutes ces personnes (c'est-à-dire toutes les valeurs *i*) également. En d'autres termes, tout ce que nous faisons ici est de faire la somme pour toutes les observations de l'ensemble de données (c.-à-d. toutes les combinaisons *rci* possibles).

A ce stade, nous connaissons la variabilité totale de la variable de résultat SST, et nous savons quelle part de cette variabilité peut être attribuée au facteur A (SSA) et quelle part peut être attribuée au facteur B (SSB). La somme résiduelle des carrés est donc définie comme étant la variabilité en *Y* qui ne peut être attribuée à aucun de nos deux facteurs. En d'autres termes

$$\mathrm{SS}_R = \mathrm{SS}_T - (\mathrm{SS}_A + \mathrm{SS}_B)$$

Bien sûr, il existe une formule que vous pouvez utiliser pour calculer directement la SS résiduelle, mais je pense qu'il est plus conceptuel de la considérer comme ceci. L'intérêt d'appeler cela un résidu, c'est qu'il s'agit d'une variation résiduelle, et la formule ci-dessus l'indique clairement. Il convient également de noter que, conformément à la terminologie utilisée dans le chapitre sur la régression, il est courant de parler de SS_A + SS_B comme étant la variance attribuable au « modèle d'ANOVA », noté SS_M, on peut ainsi dire que la somme des carrés totale est égale à la somme des carrés du modèle plus la somme des carrés

résiduelle. Plus loin dans ce chapitre, nous verrons qu'il ne s'agit pas seulement d'une similitude de surface : ANOVA et régression sont au fond la même chose.

Quoi qu'il en soit, il vaut probablement la peine de prendre un moment pour vérifier que nous pouvons calculer le SSR à l'aide de cette formule et vérifier que nous obtenons la même réponse que celle produite par Jamovi dans son tableau d'ANOVA. Les calculs sont assez simples lorsqu'ils sont effectués dans un tableur (voir le fichier clinicaltrial_factorialanova.xls). Nous pouvons calculer la SS totale à l'aide des formules ci-dessus (pour obtenir une SS totale = 4,85) et ensuite la SS résiduelle (= 0,92). Encore une fois, nous obtenons la même réponse.

Quels sont nos degrés de liberté ?

Les degrés de liberté sont calculés de la même manière que pour l'ANOVA à un facteur. Pour un facteur donné, les degrés de liberté sont égaux au nombre de niveaux moins 1 (c.-à-d. R - 1 pour la variable de ligne Facteur A et C - 1 pour la variable de colonne Facteur B). Ainsi, pour le facteur drug on obtient $df = 2$, et pour le facteur thérapeutique on obtient $df = 1$. Plus loin, lorsque nous discuterons de l'interprétation d'ANOVA comme modèle de régression (voir la section 14.6), je donnerai un énoncé plus clair de la façon dont nous en arrivons à ce chiffre. Mais pour l'instant, nous pouvons utiliser la simple définition des degrés de liberté, à savoir que les degrés de liberté sont égaux au nombre de quantités observées, moins le nombre de contraintes. Ainsi, pour le facteur drug, nous observons 3 moyennes de groupe distinctes, mais celles-ci sont limitées par 1 grande moyenne, et donc les degrés de liberté sont de 2. Pour les résidus, la logique est similaire, mais pas tout à fait la même. Le nombre total d'observations dans notre expérience est de 18. Les contraintes correspondent à 1 grande moyenne, les 2 moyennes de groupes supplémentaires que le facteur drug introduit, et 1 moyenne de groupe supplémentaire pour le facteur therapy, donc notre nombre de degrés de liberté est de 14. Nous avons comme formule $N-1-(R-1)-(C-1)$, qui se simplifie en $N-R-C+1$.

ANOVA factorielle par opposition aux ANOVA à un facteur

Maintenant que nous avons vu *comment* fonctionne une ANOVA factorielle, il vaut la peine de prendre un moment pour la comparer aux résultats des analyses à un facteur, car cela nous donnera une très bonne idée de la *raison pour laquelle* l'ANOVA factorielle est intéressante. Au chapitre 13, j'ai effectué une analyse de variance à un facteur pour voir s'il y avait des différences entre les médicaments, et une deuxième analyse de variance à un facteur pour voir s'il y avait des différences entre les traitements. Comme nous l'avons vu à la section 14.1.1, les hypothèses nulles et alternatives testées par les ANOVA à un facteur sont en fait identiques aux hypothèses testées par l'ANOVA factorielle. En regardant encore plus attentivement les tableaux ANOVA, on constate que la somme des carrés associés aux facteurs est identique dans les deux analyses (3,45 pour drug et 0,92 pour therapy), tout comme les degrés de liberté (2 pour drug, 1 pour therapy). Mais ils ne donnent pas les mêmes réponses ! Plus particulièrement, lorsque nous avons utilisé l'analyse de variance à un facteur pour therapy à la section 13.10, nous n'avons pas trouvé d'effet significatif (la *valeur p* était de .21). Cependant, quand on regarde l'effet principal de therapy dans le

contexte de l'ANOVA bifactorielle, on obtient un effet significatif (p=.019). Les deux analyses ne sont manifestement pas les mêmes.

Pourquoi cela se produit-il ? Pour répondre, il faut comprendre comment les *résidus* sont calculés. Rappelons que l'idée derrière un test *F* est de comparer la variabilité qui peut être attribuée à un facteur particulier avec la variabilité qui ne peut être prise en compte (les résidus). Si vous utilisez une ANOVA à un facteur pour therapy, et que vous ignorez donc l'effet de drug, l'ANOVA comptabilisera toute la variabilité induite par drug dans les résidus ! Cela a pour effet de d'introduire plus de bruit dans les données qu'il n'y en a en réalité, et l'effet de therapy qui s'avère à juste titre significatif dans l'ANOVA bifactorielle devient maintenant non significatif. Si nous ignorons quelque chose qui compte vraiment (p. ex., le facteur drug) lorsque nous essayons d'évaluer la contribution d'autre chose (p. ex., le facteur therapy), notre analyse sera faussée. Bien sûr, il est tout à fait normal d'ignorer les variables qui ne sont pas vraiment pertinentes pour le phénomène d'intérêt. Si nous avions enregistré la couleur des murs et que cela se soit avéré être un facteur non important dans une analyse de variance à trois facteurs, il serait tout à fait acceptable de ne pas en tenir compte et de signaler simplement l'analyse de variance à deux facteurs plus simple qui ne comprend pas ce facteur non pertinent. Ce que vous ne devriez pas faire, c'est laisser tomber les variables qui font vraiment une différence !

Quels types de résultats cette analyse saisit-elle ?

Le modèle ANOVA dont nous avons parlé jusqu'à présent couvre une gamme de modèles différents que nous pourrions observer dans nos données. Par exemple, dans une conception ANOVA bifactorielle, il y a quatre possibilités : (a) seul le facteur A compte, (b) seul le facteur B compte, (c) à la fois le facteur A et le facteur B compte, et (d) ni A ni B ne comptent. Un exemple de chacune de ces quatre possibilités est présenté à la Figure 14-4.

ANOVA Factorielle 2 : conceptions équilibrées, interactions permises

Les quatre modèles de données présentés à la Figure 14-4 sont tous très réalistes. Il existe un grand nombre d'ensembles de données qui produisent exactement ces tendances. Cependant, ils ne représentent pas toute l'histoire et le modèle ANOVA dont nous avons parlé jusqu'à présent ne suffit pas à rendre pleinement compte d'un tableau des moyennes de groupe. Pourquoi pas ? Pourquoi pas ? Eh bien, jusqu'à présent, nous avons la possibilité de parler de l'idée que les drogues peuvent influencer l'humeur, et la thérapie peut influencer l'humeur, mais pas la possibilité d'une **interaction entre les** deux. On dit qu'une interaction entre A et B se produit lorsque l'effet du facteur A est *différent*, selon le niveau du facteur B dont il est question. Plusieurs exemples d'un effet d'interaction avec le contexte d'une ANOVA 2^2 sont présentés à la Figure 14-5. Pour donner un exemple plus concret, supposons que le fonctionnement d'Anxifree et Joyzepam est régi par des mécanismes physiologiques très différents. L'une des conséquences de cette situation est que bien que Joyzepam ait plus ou moins le même effet sur l'humeur, que l'on soit en thérapie ou non, Anxifree est en fait beaucoup plus efficace lorsqu'il est administré conjointement avec la TCC.

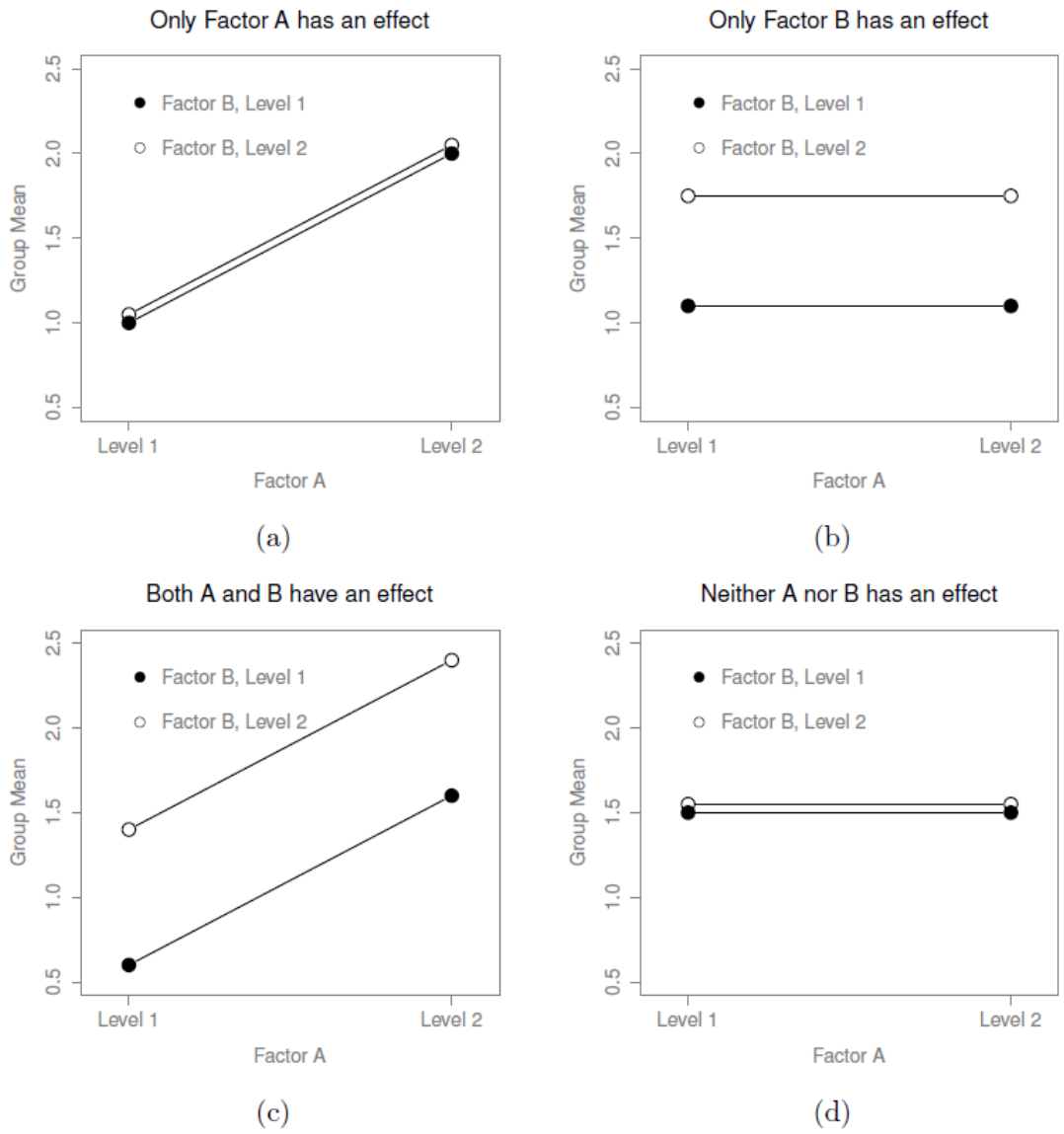


Figure 14-4 : Les quatre résultats différents d'une 2 x 2 ANOVA en l'absence d'interaction. Dans la figure (a), nous voyons un effet principal du facteur A et aucun effet du facteur B. La figure (b) montre un effet principal du facteur B mais aucun effet du facteur A. La figure (c) montre des effets principaux du facteur A et du facteur B. Enfin, la figure (d) ne montre aucun effet des deux facteurs.

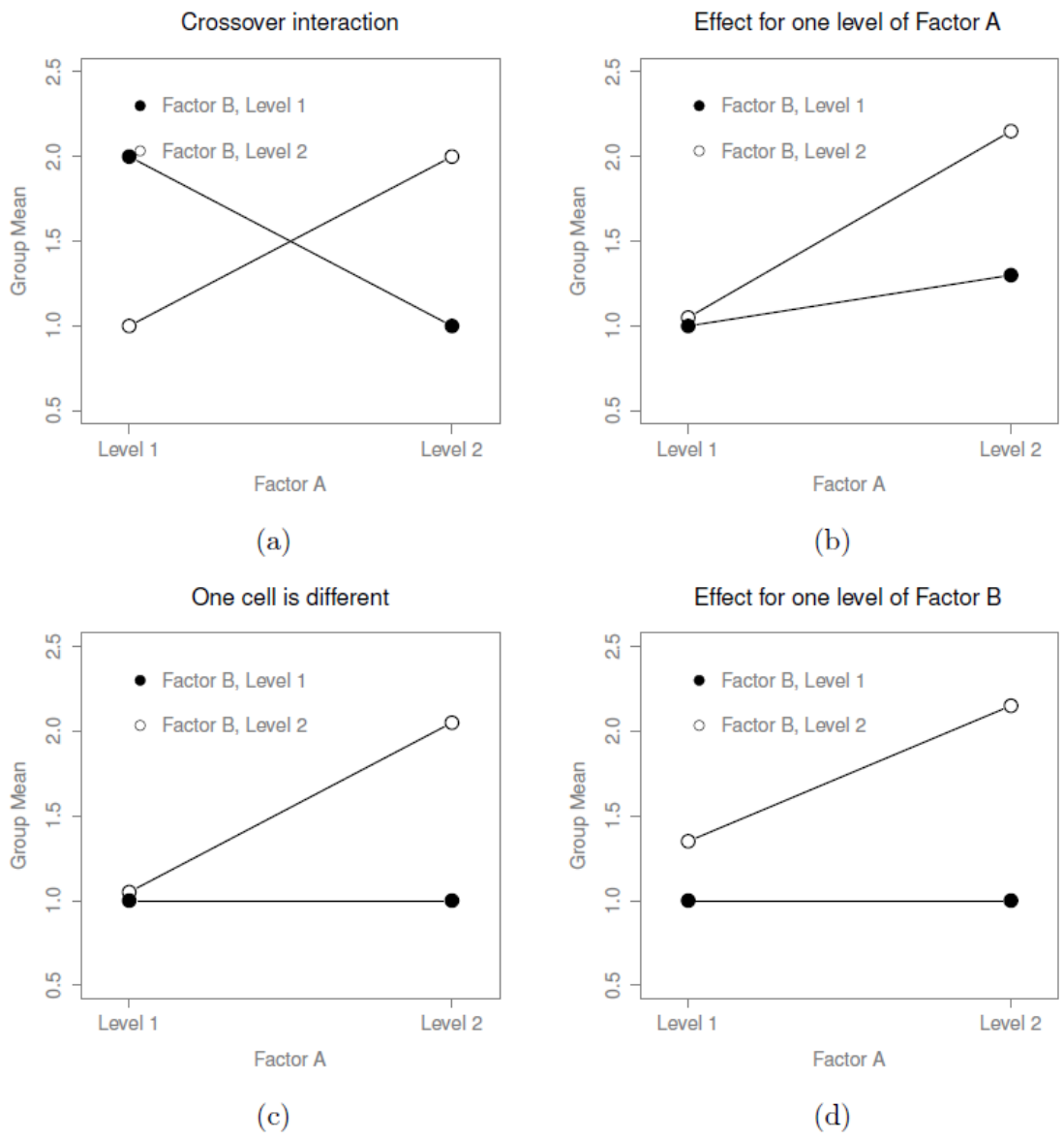


Figure 14-5 : Des interactions qualitativement différentes pour une 2 x 2 ANOVA

L'analyse de variance que nous avons élaborée dans la section précédente ne tient pas compte de cette idée. Pour se faire une idée de la réalité d'une interaction ici, il est utile de tracer les différentes moyennes de groupe. Dans le Jamovi, cela se fait via l'option « Descriptive Plots » de l'ANOVA - il suffit de déplacer le facteur drug dans la case « Horizontal axis », et de déplacer therapy dans la case « Separate Lines ». Ceci devrait ressembler à la Figure 14-6. Notre principale préoccupation concerne le fait que les deux lignes ne sont pas parallèles. L'effet de la CBT (différence entre la ligne pleine et la ligne pointillée) lorsque le médicament est le Joyzepam (côté droit) semble être près de zéro, encore plus petit que l'effet de la CBT lorsqu'un placebo est utilisé (côté gauche). Cependant, lorsqu'Anxifree est administré, l'effet de la CBT est plus important que celui du placebo (milieu). Cet effet est-il réel ou s'agit-il d'une variation aléatoire due au hasard ? Notre analyse de variance originale ne peut pas répondre à cette question, car nous ne

tenons pas compte de l'idée que les interactions existent même ! Dans cette section, nous allons régler ce problème.

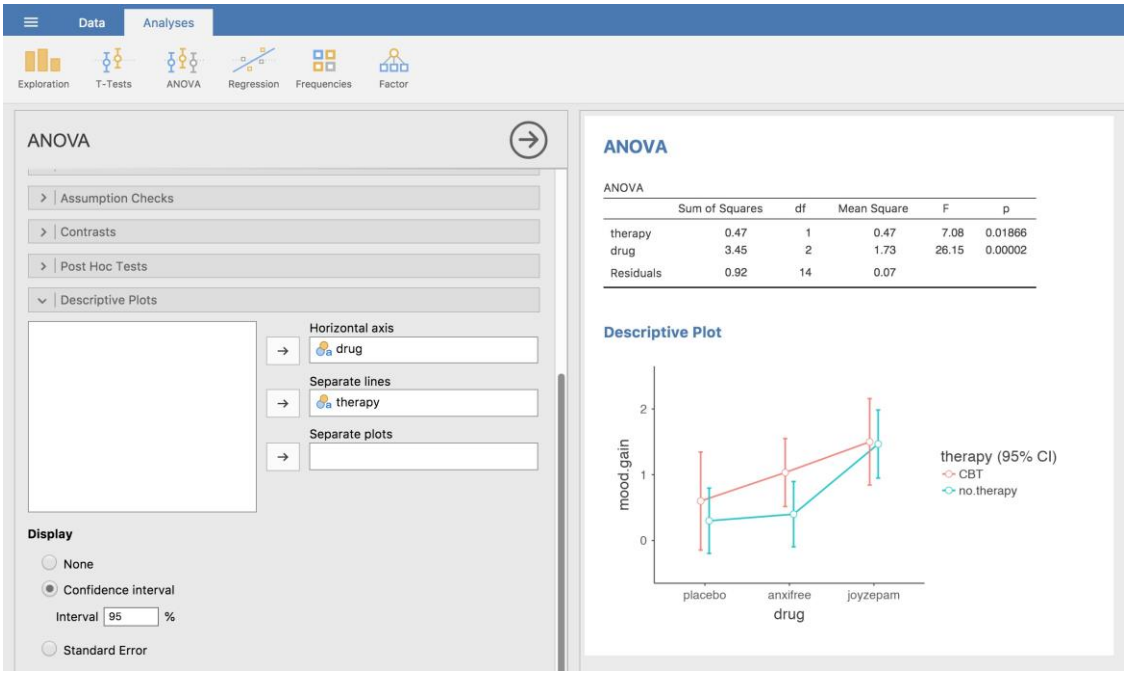


Figure 14-6 : copie d'écran Jamovi montrant comment générer un diagramme d'interaction descriptif dans ANOVA en utilisant les données des essais cliniques

Qu'est-ce qu'un effet d'interaction ?

L'idée clé que nous allons introduire dans cette section est celle d'un effet d'interaction. Dans le modèle ANOVA que nous avons examiné jusqu'à présent, il n'y a que deux *facteurs* en cause dans notre modèle (c.-à-d. drug et la therapy). Mais lorsque nous ajoutons une interaction, nous ajoutons une nouvelle composante au modèle : la combinaison de drug et de therapy. Intuitivement, l'idée derrière un effet d'interaction est assez simple. Cela signifie simplement que l'effet du facteur A est différent selon le niveau du facteur B dont nous parlons. Mais qu'est-ce que cela signifie réellement en termes de données ? Le graphique de la Figure 14-5 illustre plusieurs modèles qui, bien que très différents les uns des autres, seraient tous considérés comme un effet d'interaction. Il n'est donc pas tout à fait simple de traduire cette idée qualitative en une notion mathématique avec laquelle un statisticien peut travailler.

Par conséquent, la façon dont l'idée d'un effet d'interaction est formalisée en termes d'hypothèses nulles et alternatives est un peu difficile, et je suppose que beaucoup de lecteurs de ce livre ne seront probablement pas très intéressés. Néanmoins, je vais essayer de donner l'idée de base ici.

Pour commencer, nous devons être un peu plus explicites sur nos principaux effets. Considérons l'effet principal du facteur A (drug dans notre exemple courant). Nous avons initialement formulé cette hypothèse en fonction de l'hypothèse nulle que les deux moyennes marginales $\mu_{r.}$ sont égales l'une à l'autre. Évidemment, si toutes ces valeurs sont

égales les unes aux autres, alors elles doivent aussi être égales à la grande moyenne $\mu_{..}$ On peut donc définir l'effet du facteur A au niveau *r* comme étant égal à la différence entre la moyenne marginale $\mu_{r.}$ et la moyenne générale $\mu_{..}$. Signalons cet effet par α_r, et notons que

$$\alpha_r = \mu_{r.} - \mu_{..}$$

Maintenant, par définition, la somme de toutes les valeurs de α_r doivent être égales à zéro, pour la même raison que la moyenne des moyennes marginales $\mu_{r.}$ doit être égale à la grande moyenne $\mu_{..}$. De même, nous pouvons définir l'effet du facteur B au niveau *i* comme étant la différence entre la moyenne marginale de la colonne $\mu_{.c}$ et la moyenne générale $\mu_{..}$

$$\beta_c = \mu_{.c} - \mu_{..}$$

et une fois de plus, la somme de ces valeurs de β_c doit être égale à zéro. Les statisticiens aiment parfois parler des principaux effets avec ces valeurs α_r et β_c car cela leur permet d'être précis sur ce que signifie il n'y a aucun effet d'interaction. S'il n'y a aucune interaction, alors ces valeurs α_r et β_c décrivent parfaitement le la moyenne de groupe μ_{rc}. Plus précisément, cela signifie que

$$u_{rc} = u_{..} + \alpha_r + \beta_c$$

En d'autres termes, il n'y a rien de *particulier* pour à propos des moyennes de groupe que vous ne pourriez pas prédire parfaitement en connaissant tous les moyennes marginales. Et c'est notre hypothèse nulle, justement. L'hypothèse alternative est que

$$u_{rc} \neq u_{..} + \alpha_r + \beta_c$$

pour au moins un groupe *rc* dans notre tableau. Cependant, les statisticiens aiment souvent écrire cela un peu différemment. Ils définiront habituellement l'interaction spécifique associée au groupe *rc* comme étant un certain nombre, maladroitement appelé $(\alpha\beta)_{rc}$, puis ils diront que l'hypothèse alternative est que

$$u_{rc} = u_{..} + \alpha_r + \beta_c + (\alpha\beta)_{rc}$$

où $(\alpha\beta)_{rc}$ est différent de zéro pour au moins un groupe. Cette notation est plutôt moche à regarder, mais elle est pratique comme nous le verrons dans la prochaine section lorsque nous discuterons de la façon de calculer la somme des carrés.

Calcul des sommes de carrés pour l'interaction

Comment calculer la somme des carrés des termes d'interaction, $SS_{A:B}$? Eh bien, tout d'abord, il est utile de noter comment la section précédente a défini l'effet d'interaction en fonction de la mesure de la différence entre les moyennes réelles du groupe et ce à quoi on pourrait s'attendre en regardant simplement les moyennes marginales. Bien sûr, toutes ces formules font référence à des paramètres de population plutôt qu'à des statistiques d'échantillonnage, de sorte que nous ne savons pas vraiment ce qu'elles sont. Cependant, nous pouvons les estimer en utilisant des moyennes d'échantillonnage au lieu des moyennes de population. Ainsi, pour le facteur A, une bonne façon d'estimer l'effet principal

au niveau *r* est la différence entre la moyenne marginale $\overline{Y}_{rc}$ *de l'échantillon* et la moyenne générale $\overline{Y}_{..}$. En d'autres termes, nous utiliserions ceci comme notre estimation de l'effet

$$\hat{\alpha}_r = \overline{Y}_{r.} - \overline{Y}_{..}$$

De la même façon, notre estimation de l'effet principal du facteur B au niveau *c* peut être définie comme suit

$$\hat{\beta}_c = \overline{Y}_{.c} - \overline{Y}_{..}$$

Maintenant, si vous revenez aux formules que j'ai utilisées pour décrire les valeurs SS pour les deux effets principaux, vous remarquerez que ces termes d'effets sont exactement les quantités que nous avons élevées au carré et additionnées ! Alors, quel est l'analogie de ceci pour les termes d'interaction ? La réponse à cette question peut être trouvée en réarrangeant d'abord la formule de ma moyenne μ_{rc} pour le groupe sous l'hypothèse alternative, donc

$$\begin{aligned}(\alpha\beta)_{rc} &= \mu_{rc} - \mu_{..} - \alpha_{.r} - \beta_c \\ &= \mu_{rc} - \mu_{..} - (\mu_{r.} - \mu_{..}) - (\mu_{.c} - \mu_{..}) \\ &= \mu_{rc} - \mu_{r.} - \mu_{.c} + \mu_{..}\end{aligned}$$

Donc, encore une fois, si nous substituons nos statistiques d'échantillon à la moyenne de la population, nous obtenons ce qui suit comme estimation de l'effet d'interaction pour le groupe *rc*,

$$\left(\hat{\alpha\beta}\right)_{\mathrm{rc}} = \overline{Y}_{\mathrm{rc}} - \overline{Y}_{\mathrm{r.}} - \overline{Y}_{.\mathrm{c}} + \overline{Y}_{..}$$

Il ne nous reste plus qu'à additionner toutes ces estimations pour tous les niveaux *R* du facteur A et tous les niveaux *C* du facteur B, et nous obtenons la formule suivante pour la somme des carrés associés à l'interaction dans son ensemble

$$\mathrm{SS}_{A:B} = N \sum_{r=1}^{R} \sum_{c=1}^{C} \left(\overline{Y}_{\mathrm{rc}} - \overline{Y}_{\mathrm{r.}} - \overline{Y}_{.\mathrm{c}} + \overline{Y}_{..}\right)^2$$

où nous multiplions par *N* parce qu'il y a *N* observations dans chacun des groupes, et nous voulons que nos valeurs SS reflètent la variation entre les *observations* expliquées par l'interaction, et non la variation entre groupes.

Maintenant que nous avons une formule pour calculer $\mathrm{SS}_{A:B}$, il est important de reconnaître que le terme d'interaction fait partie du modèle (bien sûr), donc la somme totale des carrés associés au modèle, SS_M, est maintenant égale à la somme des trois valeurs SS pertinentes, $\mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{A:B}$. La somme résiduelle des carrés SSR est toujours définie comme la variation restante, à savoir SS_T-SS_M, mais maintenant que nous avons le terme d'interaction cela devient

$$\mathrm{SS}_R = \mathrm{SS}_T - (\mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{A:B})$$

Par conséquent, la somme résiduelle des carrés SSR sera plus petite que dans notre ANOVA originale qui ne comprenait pas les interactions

Degrés de liberté pour l'interaction

Le calcul des degrés de liberté pour l'interaction est, une fois de plus, légèrement plus délicat que le calcul correspondant pour les effets principaux. Pour commencer, pensons au modèle ANOVA dans son ensemble. Une fois que nous incluons les effets d'interaction dans le modèle, nous permettons à chaque groupe d'avoir une moyenne unique, μ_{rc}. Pour une ANOVA factorielle *R x C*, cela signifie qu'il y a *R x C* quantités d'intérêt dans le modèle et qu'il n'y a qu'une seule contrainte : toutes les moyennes du groupe doivent être égales à la moyenne générale. Ainsi, le modèle dans son ensemble doit avoir (R x C)-1 degrés de liberté. Mais l'effet principal du facteur A a *R*-1 degrés de liberté, et l'effet principal du facteur B a *C-1* degrés de liberté. Cela signifie que les degrés de liberté associés à l'interaction sont les suivants

$$\begin{aligned} df_{A:B} &= (R \times C - 1) - (R - 1) - (C - 1) \\ &= RC - C + 1 \\ &= (R - C)(C - 1) \end{aligned}$$

qui n'est que le produit des degrés de liberté associés au facteur de ligne et au facteur de colonne.

Qu'en est-il des degrés de liberté résiduels ? Parce que nous avons ajouté des termes d'interaction qui absorbent certains degrés de liberté, il reste moins de degrés de liberté résiduels. Plus précisément, notez que si le modèle avec interaction a un total de *(R x C)-1*, et qu'il y a *N* observations dans votre ensemble de données qui sont contraintes de satisfaire 1 grande moyenne, vos degrés de liberté résiduels deviennent maintenant *N-(R x C)-1+1*, ou seulement *N-(R x C)*.

Exécuter l'ANOVA dans Jamovi

L'ajout de termes d'interaction au modèle ANOVA dans Jamovi est simple. En fait, c'est plus que simple parce que c'est l'option par défaut pour ANOVA. Cela signifie que lorsque vous spécifiez une ANOVA avec deux facteurs, par exemple drug et therapy, la composante d'interaction - drug*therapy - est automatiquement ajoutée au modèle[125]. Lorsque nous exécutons l'analyse de variance avec le terme d'interaction inclus, nous obtenons les résultats présentés à la Figure 14-7.

[125] Vous l'avez peut-être déjà remarqué en regardant l'analyse des effets principaux dans Jamovi que nous avons décrite plus haut. Pour les besoins des explications dans ce livre, j'ai supprimé la composante d'interaction du modèle précédent pour garder les choses propres et simples.

Il s'avère que, bien que nous ayons un effet principal significatif du médicament ($F(2,12) = 31,7$, p<.001) et le type de thérapie ($F(1,12) = 8,6$, p=.013), il n'y a aucune interaction significative entre les deux ($F(2,12) = 2,5$, p=0.125).

ANOVA

	Sum of Squares	df	Mean Square	F	p	η²	η²p	ω²
drug	3.45	2	1.73	31.71	0.00002	0.71	0.84	0.68
therapy	0.47	1	0.47	8.58	0.01262	0.10	0.42	0.08
drug ✻ therapy	0.27	2	0.14	2.49	0.12460	0.06	0.29	0.03
Residuals	0.65	12	0.05					

Figure 14-7 : Résultats pour le modèle factoriel complet, y compris la composante d'interaction drug*therapy

Interprétation des résultats

Il y a quelques éléments très importants à prendre en considération lors de l'interprétation des résultats de l'analyse de variance factorielle. Tout d'abord, il y a le même problème que nous avions avec l'analyse de variance à un facteur, si vous obtenez un effet principal important d'un facteur (disons) drug, cela ne vous dit rien sur les différences entre médicaments. Pour le savoir, vous devez exécuter des analyses supplémentaires. Nous parlerons de certaines analyses que vous pouvez exécuter dans les sections 14.7 et 14.8. Il en va de même pour les effets d'interaction. Savoir qu'il y a une interaction importante ne vous dit rien sur le type d'interaction qui existe. Encore une fois, vous devrez effectuer des analyses supplémentaires.

Deuxièmement, il y a un problème d'interprétation très particulier qui se pose lorsque vous obtenez un effet d'interaction significatif mais aucun effet principal correspondant. Cela arrive parfois. Par exemple, dans l'interaction croisée illustrée à la Figure 14-5a, c'est exactement ce que vous trouverez. Dans ce cas, ni l'un ni l'autre des principaux effets ne serait significatif, mais l'effet d'interaction le serait. C'est une situation difficile à interpréter, et les gens sont souvent un peu confus. Le conseil général que les statisticiens aiment donner dans cette situation est que vous ne devriez pas accorder beaucoup d'attention aux effets principaux quand une interaction est présente. La raison en est que, bien que les tests des effets principaux soient parfaitement valables d'un point de vue mathématique, lorsqu'il y a un effet d'interaction significatif, les effets principaux testent rarement des hypothèses intéressantes. Rappelons, à la section 14.1.1, que l'hypothèse nulle pour un effet principal est que les *moyennes marginales* sont égales les unes aux autres et qu'une moyenne marginale est formée en faisant la moyenne de plusieurs groupes différents. Mais si vous avez un effet d'interaction significatif, vous *savez* que les groupes qui composent la moyenne marginale ne sont pas homogènes, alors le motif de l'intérêt pour ces moyennes marginales n'est pas vraiment évident.

Je tenais à le préciser. Encore une fois, restons-en à un exemple clinique. Supposons que nous ayons un plan 2x2 comparant deux thérapies différentes pour les phobies (p. ex. désensibilisation systématique vs exposition in vivo) et deux médicaments anti-anxiété différents (p. ex. Anxifree vs Joyzepam). Supposons maintenant que ce que nous avons

découvert, c'est qu'Anxifree n'avait aucun effet lorsque la thérapie est la désensibilisation, et que Joyzepam n'avait aucun effet avec l'exposition in vivo. Mais les deux ont été assez efficaces pour l'autre thérapie. Il s'agit d'une interaction croisée classique, et ce que nous constatons en exécutant l'analyse de variance, c'est qu'il n'y a pas d'effet principal du médicament, mais une interaction significative. Maintenant, qu'est-ce que cela *signifie* de dire qu'il n'y a pas d'effet principal ? Eh bien, cela signifie que si nous faisons la moyenne sur les deux traitements psychologiques, alors l'effet *moyen* d'Anxifree et de Joyzepam est le même. Mais qui cela intéresse-t-il ? Lorsqu'on traite quelqu'un pour des phobies, il n'est jamais possible de traiter une personne en utilisant une « moyenne » d'exposition et de désensibilisation. Ça n'a pas beaucoup de sens. Soit vous avez l'un, soit l'autre. Pour un traitement, un médicament est efficace, et pour l'autre, c'est l'autre médicament qui est efficace. Ce qui importe, c'est l'interaction et l'effet principal n'a pas d'importance.

Ce genre de choses arrive souvent. Les principaux effets sont des tests de moyennes marginales, et lorsqu'une interaction est présente, nous trouvons souvent les moyennes marginales sans grand intérêt parce qu'elles impliquent de faire la moyenne des choses dont l'interaction nous dit de ne pas en faire la moyenne ! Bien sûr, il n'est pas toujours vrai qu'un effet principal n'a pas de sens lorsqu'une interaction est présente. Souvent, on peut obtenir un grand effet principal et une très petite interaction, auquel cas on peut encore dire des choses comme « le médicament A est généralement plus efficace que le médicament B » (parce qu'il y avait un grand effet du médicament), mais il faudrait le modifier un peu en ajoutant que « la différence d'efficacité était différente en fonction des différents traitements psychologiques ». Quoi qu'il en soit, le point principal ici est que chaque fois que vous obtenez une interaction significative, vous devriez vous arrêter et *réfléchir* à ce que l'effet principal signifie réellement dans ce contexte. Ne supposez pas automatiquement que l'effet principal est intéressant.

Taille de l'effet

Le calcul de la valeur de l'effet d'une ANOVA factorielle est assez semblable à celui d'une ANOVA à un facteur (voir section 13.4). Plus précisément, nous pouvons utiliser η^2 (eta-carré) comme un moyen simple de mesurer la taille de l'effet global pour un terme donné. Comme précédemment, η^2 est défini en divisant la somme des carrés associés à ce terme par la somme totale des carrés. Par exemple, pour déterminer l'ampleur de l'effet principal du facteur A, nous utiliserions la formule suivante :

$$\eta^2 = \frac{\text{SS}_A}{\text{SS}_T}$$

Comme précédemment, ceci peut être interprété de la même manière que R2 en régression.[126] Il vous indique la proportion de variance de la variable résultat qui peut être

[126] Ce chapitre semble établir un nouveau record pour le nombre de choses différentes que la lettre R peut représenter. Jusqu'à présent, R fait référence au progiciel, au nombre de lignes de notre tableau de moyennes, aux résidus dans le modèle et maintenant au coefficient de corrélation dans une régression. Désolée. Nous n'avons clairement pas assez

expliquée par l'effet principal du facteur A. Il s'agit donc d'un nombre qui varie de 0 (aucun effet du tout) à 1 (qui explique *toute* la variabilité du résultat). De plus, la somme des valeurs de η^2 pour tous les termes du modèle est égale au R^2 total pour le modèle d'ANOVA. Si, par exemple, le modèle ANOVA est parfaitement adapté (c'est-à-dire qu'il n'y a aucune variabilité à l'intérieur des groupes !), la somme des valeurs η^2 sera égale à 1. Bien sûr, cela arrive rarement, voire jamais, dans la vraie vie.

Cependant, lorsqu'on effectue une analyse de variance factorielle, il existe une deuxième mesure de la taille de l'effet que les gens aiment signaler, connue sous le nom de η^2 partiel. L'idée qui sous-tend le η^2 partiel (noté parfois $_p\,\eta^2$ ou η_p^2) est que, lorsqu'on mesure l'ampleur de l'effet pour un terme particulier (disons, l'effet principal du facteur A), on veut délibérément ignorer les autres effets du modèle (p. ex., l'effet principal du facteur B). C'est-à-dire, vous souhaiteriez faire semblant que l'effet de tous ces autres termes est nul afin de calculer ce que la valeur de η^2 aurait été. C'est en fait assez facile à calculer. Tout ce que vous avez à faire est d'enlever la somme des carrés associés aux autres termes du dénominateur. En d'autres termes, si vous voulez l'effet principal du Facteur A sur η^2, le dénominateur est juste la somme des carrés du Facteur A et des résidus.

$$\text{partial } \eta_A^2 = \frac{\text{SS}_A}{\text{SS}_A + \text{SS}_R}$$

Cela vous donnera toujours un nombre plus grand que η^2, ce que le cynique que je suis soupçonne d'expliquer la popularité de η^2 partiel. Et encore une fois, vous obtenez un nombre entre 0 et 1, où 0 représente aucun effet. Cependant, il est un peu plus difficile d'interpréter ce que signifie une grande valeur partielle de η^2. En particulier, vous ne pouvez pas comparer les valeurs partielles de η^2 d'un terme à l'autre ! Supposons, par exemple, qu'il n'y ait aucune variabilité à l'intérieur des groupes, dans ce cas, SSR = 0. Cela signifie que *chaque* terme a une valeur partielle η^2 de 1. Mais cela ne signifie pas que tous les termes dans votre modèle sont également importants, ou même qu'ils sont aussi grands. Tout ce que cela signifie, c'est que tous les termes de votre modèle ont des valeurs d'effet qui sont importantes par rapport à la *variation résiduelle*. Elle n'est pas comparable d'un terme à l'autre.

Pour voir ce que j'entends par là, il est utile de voir un exemple concret. Examinons d'abord la taille de l'effet de l'analyse de variance originale sans le terme d'interaction, à la Figure 14-3 :

	Eta.sq	Partial.eta.sq
drug	0,71	0,79
therapy	0,10	0,34

---

de lettres dans l'alphabet. Cependant, j'ai essayé d'être assez clair sur ce à quoi R fait référence dans chaque cas.

En regardant d'abord les valeurs de η^2, on constate que drug représente 71 % de la variance (c.-à-d. $\eta^2 = 0{,}71$) pour la variable mood.gain, alors que le facteur therapy ne représente que 10 %. Cela laisse un total de 19 % de la variation non prise en compte (c.-à-d. que les résidus constituent 19 % de la variation du résultat). Dans l'ensemble, cela implique que nous avons un très grand effet de[127] drug et un effet modeste de therapy.

Regardons maintenant les valeurs partielles de η^2, illustrées à la Figure 14-3. Parce que l'effet de therapy n'est pas si important, le contrôle de l'effet ne fait pas beaucoup de différence, donc la valeur partielle η^2 pour la variable drug n'augmente pas beaucoup, et on obtient une valeur de $_p\eta^2 = 0{,}79$. En revanche, parce que l'effet de drug était très important, la prise en compte de l'effet de drug fait une grande différence, et donc lorsque nous calculons la valeur partielle de η^2 pour la variable therapy, vous pouvez voir qu'elle s'élève à $_p\eta^2 = 0{,}34$. La question que nous devons nous poser est la suivante : que *signifient* réellement ces valeurs partielles de η^2 ? La façon dont j'interprète généralement le η^2 partiel pour l'effet principal du facteur A est de l'interpréter comme un énoncé au sujet d'une expérience hypothétique dans laquelle seul le facteur A était modifié. Ainsi, même si, dans cette expérience, nous avons deux facteurs A et B, nous pouvons facilement imaginer une expérience dans laquelle seul le facteur A utilisé, et la statistique partielle η^2 vous indique quelle part de la variance de la variable résultat que vous vous attendriez à voir prise en compte dans cette expérience. Cependant, il faut noter que cette interprétation, comme beaucoup de choses associées aux effets principaux, n'a pas beaucoup de sens lorsqu'il y a un effet d'interaction important et significatif.

En parlant d'effets d'interaction, voici ce que nous obtenons lorsque nous calculons la taille de l'effet pour le modèle qui inclut le terme d'interaction, comme dans la Figure 14-7. Comme vous pouvez le voir, les valeurs de η^2 pour les effets principaux ne changent pas, contrairement aux valeurs partielles de η^2 :

```
                eta.sq partial.eta.sq
drug            0.71        0.84
therapy         0.10        0.42
drug*therapy    0.06        0.29
```

Moyenne estimée du groupe

Dans de nombreuses situations, vous voudrez déclarer des estimations de toutes les moyennes de groupe en fonction des résultats de votre analyse de variance, ainsi que des

---

[127] Impossible à croire, je dirais. L'artificialité de cet ensemble de données commence vraiment à se manifester !

intervalles de confiance qui y sont associés. Pour ce faire, vous pouvez utiliser l'option « Estimated Marginal Means » dans l'analyse ANOVA de Jamovi, comme dans la Figure 14-8. Si l'analyse de variance que vous avez exécutée est un **modèle saturé** (c.-à-d. qu'elle contient tous les effets principaux possibles et tous les effets d'interaction possibles), les estimations des moyennes des groupes sont en fait identiques aux moyennes de l'échantillon, bien que les intervalles de confiance utilisent une estimation globale des erreurs types plutôt que des estimations distinctes pour chaque groupe.

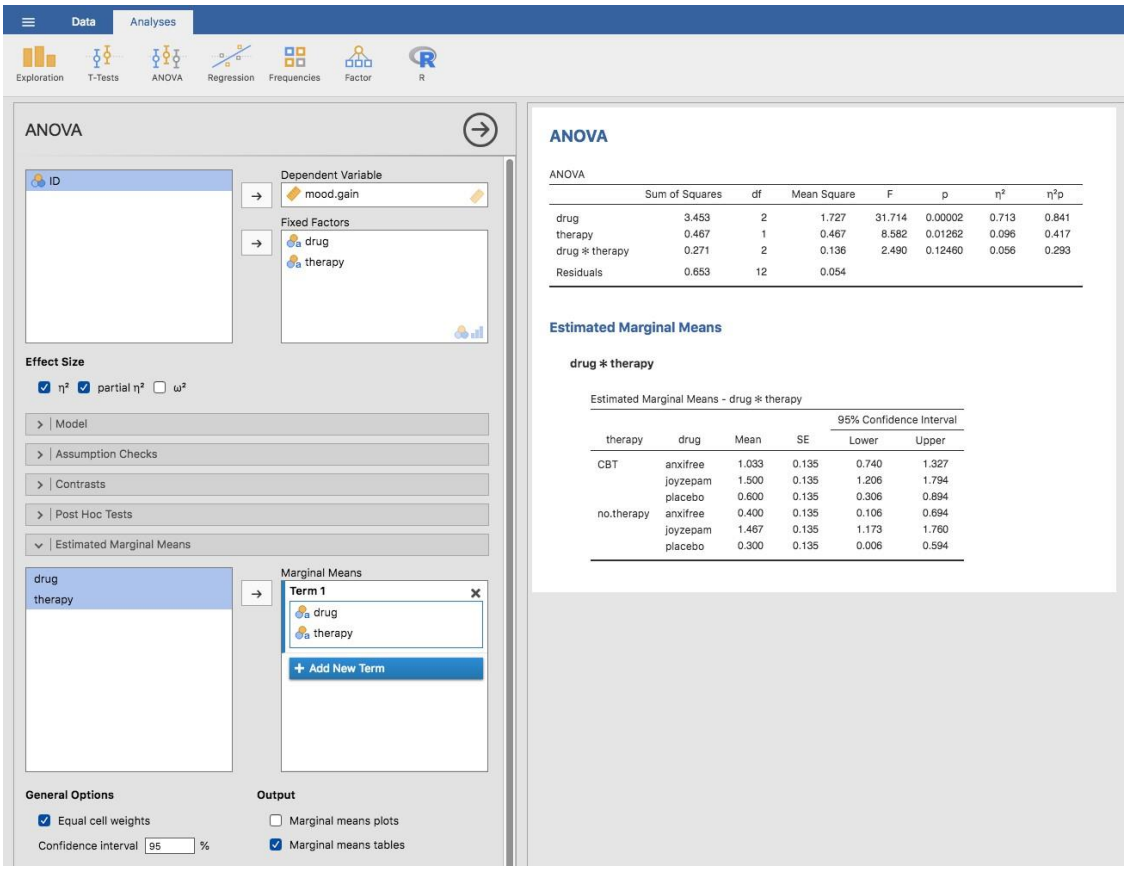


Figure 14-8 : capture d'écran de Jamovi montrant les moyennes marginales du modèle saturé, c'est-à-dire incluant la composante d'interaction, avec l'ensemble des données clinicaltrial

Les résultats montrent que le gain d'humeur moyen estimé pour le groupe placebo sans traitement était de 0,300, avec un intervalle de confiance à 95 % allant de 0,006 à 0,594. Il est à noter qu'il ne s'agit pas des mêmes intervalles de confiance que ceux que vous obtiendriez si vous les calculiez séparément pour chaque groupe, parce que le modèle ANOVA suppose l'homogénéité de la variance et utilise donc une estimation globale de l'écart type.

Lorsque le modèle ne contient pas le terme d'interaction, la moyenne estimée du groupe sera différente de la moyenne de l'échantillon. Au lieu de déclarer la moyenne de l'échantillon, Jamovi calculera la valeur de la moyenne du groupe à partir de la moyenne marginale (c.-à-d. en supposant qu'il n'y a aucune interaction). En utilisant la notation que

nous avons développée précédemment, l'estimation rapportée pourμ_{rc}, la moyenne pour le niveau *r* sur le facteur A (ligne) et le niveau *c* sur le facteur B (colonne) serait $u_{..} + \alpha_r + \beta_c$. S'il n'y a vraiment aucune interaction entre les deux facteurs, il s'agit en fait d'une meilleure estimation de la moyenne de la population que la moyenne brute de l'échantillon. La suppression du terme d'interaction du modèle, via les options « Model » de l'analyse ANOVA de Jamovi, fournit les moyennes marginales pour l'analyse présentée à la Figure 14-9.

ANOVA

ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	η^2p
drug	3.453	2	1.727	26.149	0.00002	0.713	0.789
therapy	0.467	1	0.467	7.076	0.01866	0.096	0.336
Residuals	0.924	14	0.066				

Estimated Marginal Means

drug ✻ therapy

Estimated Marginal Means - drug ✻ therapy

				95% Confidence Interval	
therapy	drug	Mean	SE	Lower	Upper
CBT	anxifree	0.878	0.121	0.618	1.138
	joyzepam	1.644	0.121	1.385	1.904
	placebo	0.611	0.121	0.351	0.871
no.therapy	anxifree	0.556	0.121	0.296	0.815
	joyzepam	1.322	0.121	1.062	1.582
	placebo	0.289	0.121	0.029	0.549

Figure 14-9 : capture d'écran de Jamovi montrant les moyennes marginales du modèle insaturé, c'est-à-dire sans la composante interaction, avec l'ensemble des données clinicaltrial

Vérification des hypothèses

Comme pour l'analyse de variance à un facteur, les hypothèses clés de l'analyse de variance factorielle sont l'homogénéité de la variance (tous les groupes ont le même écart-type), la normalité des résidus et l'indépendance des observations. Les deux premiers sont des choses qu'on peut vérifier. La troisième est quelque chose que vous devez évaluer vous-même en vous demandant s'il y a des relations spéciales entre les différentes observations, par exemple des mesures répétées où la variable indépendante est le temps, de sorte qu'il y a une relation entre les observations au temps un et au temps deux : les observations à

différents moments proviennent des *mêmes* personnes. De plus, si vous n'utilisez pas un modèle saturé (par exemple, si vous avez omis les termes d'interaction), vous supposez également que les termes omis ne sont pas importants. Bien sûr, vous pouvez vérifier cette dernière en exécutant une ANOVA avec les termes omis inclus et voir s'ils sont significatifs, c'est assez donc facile. Qu'en est-il de l'homogénéité de la variance et de la normalité des résidus ? Il s'avère que c'est assez facile à vérifier. Ce n'est pas différent des contrôles que nous avons effectués pour une ANOVA à un facteur.

Homogénéité de la variance

Comme nous l'avons mentionné à la section 13.6.1, il est bon d'inspecter visuellement un graphique des écarts-types comparés entre différents groupes ou catégories, et de voir si le test de Levene est conforme à l'inspection visuelle. La théorie qui sous-tend le test de Levene a été abordée à la section 13.6.1, de sorte que je n'en parlerai plus. Ce test s'attend à ce que vous ayez un modèle saturé (c.-à-d., incluant tous les éléments suivants les termes pertinents), parce que le test porte principalement sur la variance intra-groupe et qu'il n'est pas vraiment logique de calculer cela autrement que par rapport au modèle complet. Le test de Levene peut être spécifié dans le cadre de l'option de l'ANOVA « Assumption Checks » - « Homogeneity Tests » dans Jamovi, avec le résultat indiqué à la Figure 14-10. Le fait que le test de Levene ne soit pas significatif signifie que, à condition qu'il soit cohérent avec une inspection visuelle du graphique des écarts-types, nous pouvons supposer avec certitude que l'hypothèse d'homogénéité de la variance n'est pas violée.

Normalité des résidus

Comme pour l'analyse de variance à sens unique, nous pouvons tester la normalité des résidus d'une manière simple et directe (voir la section 13.6.4). Cependant, c'est généralement une bonne idée d'examiner les résidus graphiquement à l'aide d'un graphe QQ. Voir la Figure 14-10.

Assumption Checks

Test for Homogeneity of Variances (Levene's)

F	df1	df2	p
0.21	5	12	0.95384

Q-Q Plot

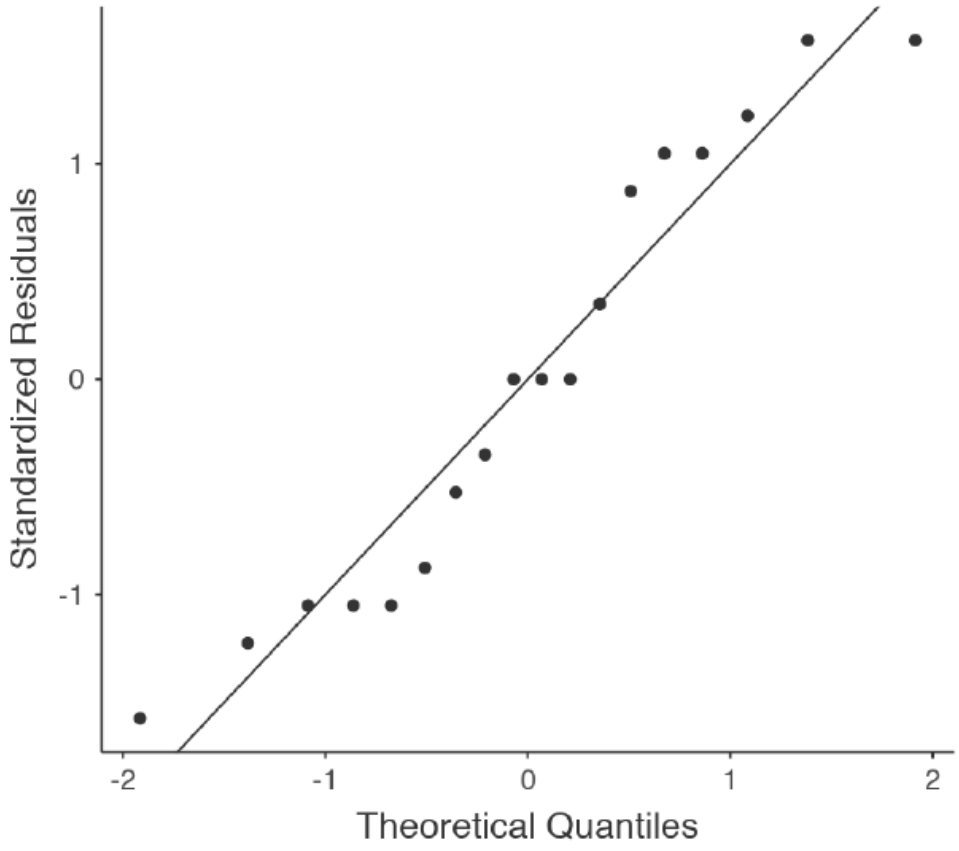


Figure 14-10 : Vérification des présupposés dans un modèle d'ANOVA

Analyse de la covariance (ANCOVA)

Une variation de l'analyse de variance se produit lorsqu'il y a une variable continue supplémentaire qui, à votre avis, pourrait être liée à la variable dépendante. Cette variable supplémentaire peut être ajoutée à l'analyse en tant que covariable, dans l'analyse de covariance bien nommée (ANCOVA).

Dans ANCOVA, les valeurs de la variable dépendante sont « ajustées » pour tenir compte de l'influence de la covariable, puis les moyennes de score « ajustées » sont testées entre groupes de la manière habituelle. Cette technique peut augmenter la précision d'une expérience, et donc fournir un test plus « puissant » de l'égalité des moyennes de groupe pour la variable dépendante. Comment ANCOVA s'y prend-elle ? Bien que la covariable elle-même ne présente généralement aucun intérêt expérimental, l'ajustement pour la covariable peut diminuer l'estimation de l'erreur expérimentale et donc, en réduisant la variance de l'erreur, la précision est accrue. Cela signifie que rejeter l'hypothèse nulle de façon inappropriée (faux négatif ou erreur de type II) est moins probable.

Malgré cet avantage, ANCOVA court le risque d'aplanir les différences réelles entre les groupes, ce qu'il faut éviter. Par exemple, regardez la Figure 14-11, qui montre un graphique de l'aversion pour les statistiques par rapport à l'âge et dans deux groupes distincts - les élèves qui ont une formation ou une préférence en arts ou en sciences. ANCOVA avec l'âge comme covariable pourrait mener à la conclusion que l'anxiété statistique ne diffère pas entre les deux groupes. Cette conclusion serait-elle raisonnable - probablement pas parce que les âges des deux groupes ne se chevauchent pas et que l'analyse de la variance a essentiellement « extrapolé à une région sans données » (Everitt Everitt (1996), p. 68).

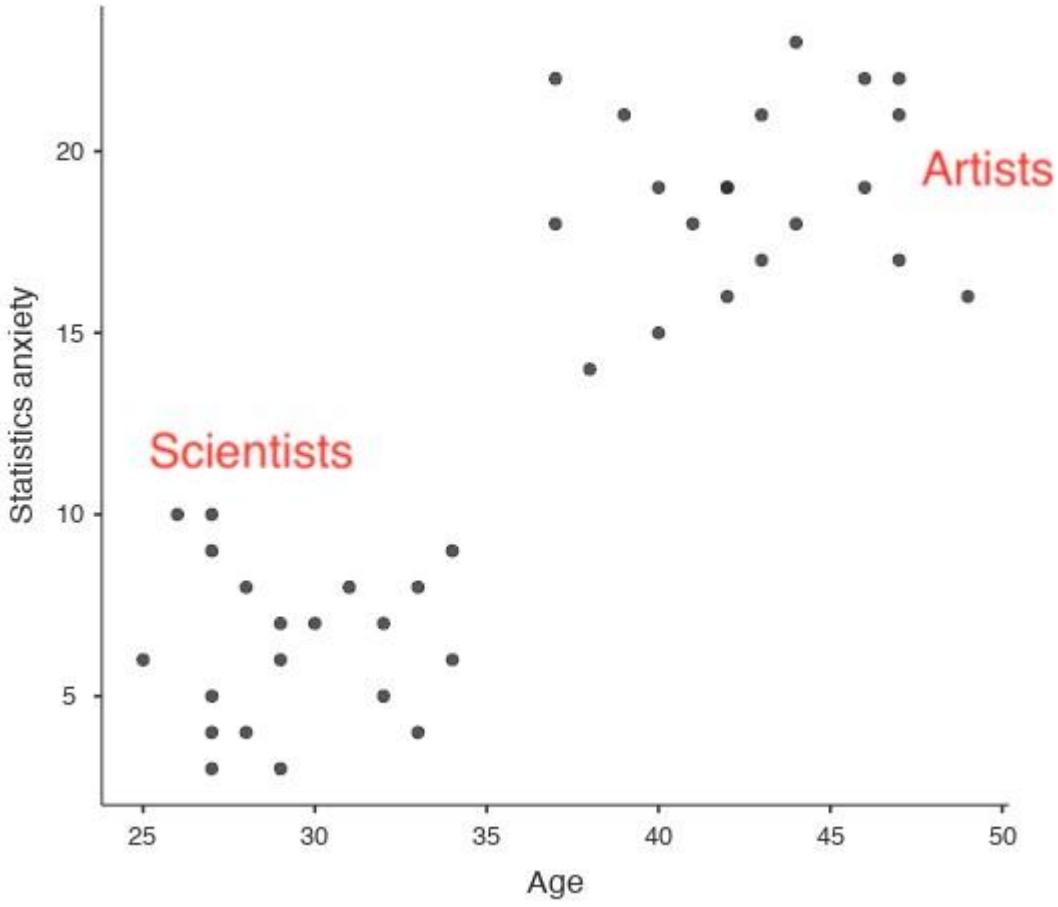


Figure 14-11 : Représentation graphique de l'aversion aux statistiques par rapport à l'âge pour deux groupes distincts

De toute évidence, il faut réfléchir soigneusement à l'analyse de la covariance avec des groupes distincts. Ceci s'applique à la fois aux plans à un facteur et factorielles, car ANCOVA peut être utilisé avec les deux.

Exécuter ANCOVA en Jamovi

Un psychologue de la santé s'est intéressé à l'effet de l'utilisation habituelle du vélo et du stress sur les niveaux de bonheur, avec l'âge comme covariable. Vous pouvez trouver l'ensemble de données dans le fichier ancova.csv. Ouvrez ce fichier dans Jamovi et ensuite, pour entreprendre une ANCOVA, sélectionnez Analyses - ANOVA - ANCOVA pour ouvrir la fenêtre ANCOVA analysis (Figure 14-12). Sélectionnez la variable dépendante « bonheur » et transférez-la dans la zone de texte « Dependant Variable ». Sélectionnez les variables indépendantes « stress » et « commute » et transférez-les dans la zone de texte « Fixed Factors ». Mettez en surbrillance la covariable « âge » et transférez-la dans la zone de texte « Covariables ». Cliquez ensuite sur Moyennes marginales estimées… pour afficher les options des graphiques et des tableaux.

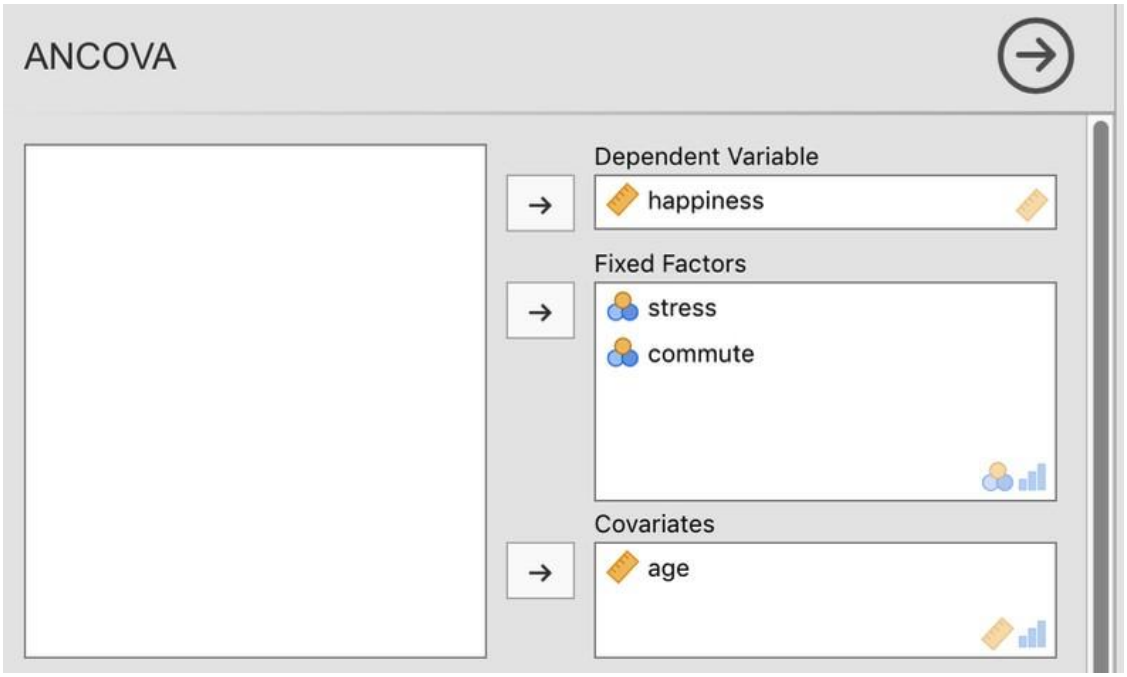


Figure 14-12 : La fenêtre d'analyse ANCOVA de Jamovi

Un tableau ANCOVA montrant les tests sur les effets inter sujet est produit dans la fenêtre de résultats Jamovi (Figure 14-13). La valeur F de la covariable « âge » est significative à p=.023, ce qui suggère que l'âge est un prédicteur important de la variable dépendante, le bonheur. Lorsque nous examinons les scores marginaux moyens estimés (Figure 14-14), des ajustements ont été faits (par rapport à une analyse sans covariable) en raison de l'inclusion de la covariable « âge » dans cet ANCOVA. Un graphique (Figure 14-15) est un bon moyen de visualiser et d'interpréter les effets significatifs.

La valeur F de l'effet principal « contrainte » (52,61) est associée à une probabilité de p<.001. La valeur F de l'effet principal « commute » (42,33) est associée à une probabilité de p<.001. Comme ces deux valeurs sont inférieures à la probabilité habituellement utilisée pour décider si un résultat statistique est significatif (p<.05), nous pouvons conclure qu'il y a eu un effet principal significatif du stress (F(1,15) = 52,61, p<.001) et un effet principal significatif de la méthode du transport quotidien (F(1,15) = 42,33, p<.001). Une interaction significative entre le stress et la mode de déplacement a également été trouvée (F(1,15) = 14,15, p=.002).

ANCOVA

	Sum of Squares	df	Mean Square	F	p	η²	η²p	ω²
stress	2751.52	1	2751.52	52.61	<.00001	0.40	0.78	0.39
commute	2213.93	1	2213.93	42.33	<.00001	0.32	0.74	0.31
age	334.35	1	334.35	6.39	0.02316	0.05	0.30	0.04
stress ✻ commute	740.12	1	740.12	14.15	0.00188	0.11	0.49	0.10
Residuals	784.45	15	52.30					

Figure 14-13 : Résultats de l'ANCOVA dans Jamovi pour la variable bonheur (hapiness) en fonction du stress et du mode de déplacement (commute), avec l'âge comme covariable.

Estimated Marginal Means - stress ✻ commute

				95% Confidence Interval	
commute	stress	Mean	SE	Lower	Upper
drive	high	36.11	3.24	29.21	43.02
	low	51.09	3.26	44.13	58.04
cycle	high	43.58	3.84	35.40	51.76
	low	85.82	3.71	77.90	93.74

Figure 14-14 : Tableau du niveau de bonheur moyen en fonction du stress et de la mode de déplacement (ajusté pour l'âge covarié) avec des intervalles de confiance à 95 %.

Dans la Figure 14-15, nous pouvons voir les scores de bonheur ajustés, marginaux et moyens lorsque l'âge est une covariable dans une ANCOVA. Dans cette analyse, il existe un effet d'interaction significatif, selon lequel les personnes peu stressées qui se rendent au travail à vélo sont plus heureuses que les personnes peu stressées qui y vont en voiture et les personnes très stressées, qu'elles se rendent au travail à vélo ou en voiture. Il y a aussi un effet principal important du stress - les personnes peu stressées sont plus heureuses que celles qui sont très stressées. Et il y a aussi un effet principal important du comportement de déplacement domicile-travail - les gens qui font du vélo sont en moyenne plus heureux que ceux qui se rendent au travail en voiture.

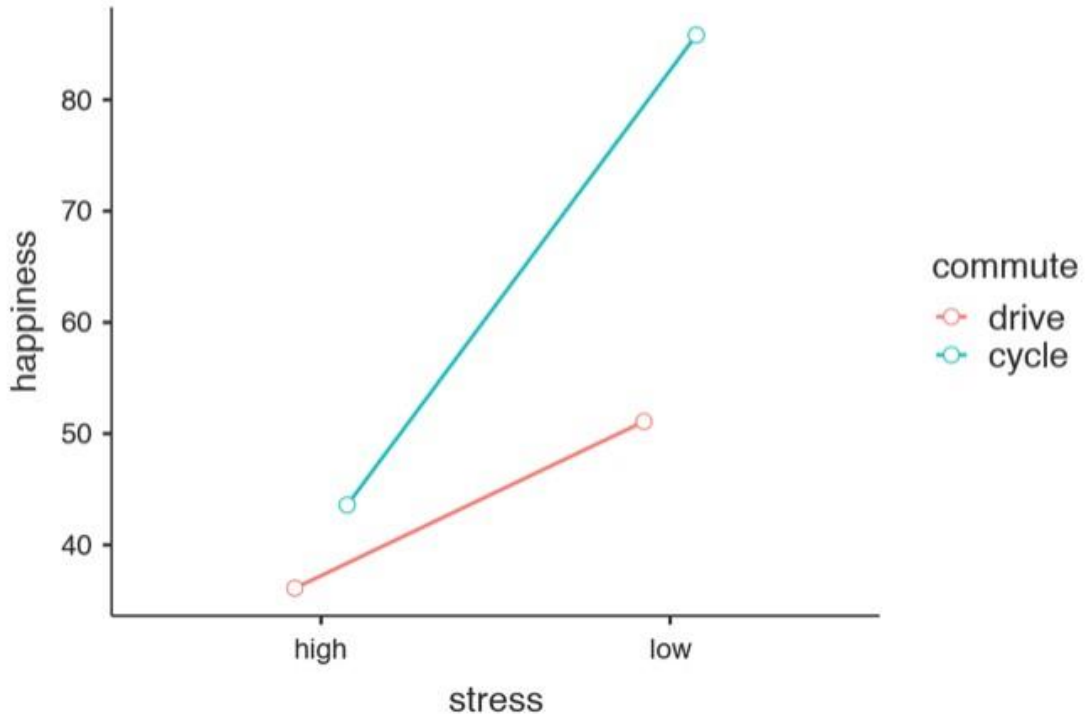


Figure 14-15 : Diagramme du niveau de bonheur moyen en fonction du stress et de la méthode de déplacement

Vous devez être attentif au fait que, si vous songez à inclure une covariable dans votre analyse de variance, il y a une hypothèse supplémentaire : la relation entre la covariable et la variable dépendante doit être semblable pour tous les niveaux de la variable indépendante. Ceci peut être vérifié par l'ajout d'un terme d'interaction entre la covariable et chaque variable indépendante dans les options de Jamovi « Model - Model terms ». Si l'effet d'interaction n'est pas significatif, il peut être supprimé. S'il est significatif, alors une technique statistique différente et plus avancée pourrait être appropriée (ce qui dépasse le cadre de ce livre et vous voudrez peut-être consulter un statisticien amical).

ANOVA comme modèle linéaire

L'une des choses les plus importantes à comprendre au sujet de l'analyse de variance et de la régression est qu'il s'agit essentiellement de la même chose. À première vue, on ne le croirait peut-être pas. Après tout, la façon dont je les ai décrites jusqu'à présent suggère que l'analyse de variance vise principalement à vérifier les différences entre les groupes et que la régression vise principalement à comprendre les corrélations entre les variables. Et, pour autant que je sache, c'est tout à fait vrai. Mais quand on regarde dans le moteur, pour ainsi dire, les mécanismes sous-jacents de l'analyse de variance et de la régression sont terriblement semblables. En fait, si vous y pensez, vous en avez déjà vu la preuve. L'analyse de variance et la régression reposent toutes deux fortement sur des sommes de carrés (SS),

toutes deux font appel à des tests *F*, et ainsi de suite. Rétrospectivement, il est difficile d'échapper au sentiment que les chapitres 12 et 13 étaient un peu répétitifs.

La raison en est que l'analyse de variance et la régression sont deux types de **modèles linéaires**. Dans le cas de la régression, c'est un peu évident. L'équation de régression que nous utilisons pour définir la relation entre les prédicteurs et les résultats est l'équation d'une droite, donc c'est de toute évidence un modèle linéaire, avec l'équation suivante

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \epsilon_p$$

où Y_p est la valeur finale de la *p-ième* observation (*c.-à-d., p-ième* personne), X_{1p} est la valeur du premier prédicteur de la *p-ième* observation, X_{2p} est la valeur du deuxième prédicteur de la *p-ième* observation, les termes b_0, b_1 et b_2 sont nos coefficients de régression, et ϵ_p est le *p-ième* résidu. Si nous ignorons les résidus ϵ_p et que nous nous concentrons sur la ligne de régression elle-même, nous obtenons la formule suivante :

$$\hat{Y}_p = b_0 + b_1 X_{1p} + b_2 X_{2p}$$

où $\hat{Y}_p$ est la valeur de *Y* que la ligne de régression prédit pour la personne *p*, par opposition à la valeur Y_p réellement observée. Ce qui n'est pas immédiatement évident, c'est que nous pouvons aussi écrire ANOVA comme modèle linéaire. C'est en fait assez simple à faire. Commençons par un exemple très simple, en réécrivant une ANOVA factorielle 2 x 2 comme modèle linéaire.

Quelques données

Pour concrétiser les choses, supposons que notre variable de résultat est la grade (note) qu'un élève reçoit dans mon cours, une variable sur une échelle de rapport correspondant à une note de 0% à 100%. Il y a deux variables prédictrices d'intérêt : si l'étudiant s'est présenté aux cours (la variable attend (fréquentation)) et si l'étudiant a lu ou non le manuel (la variable reading (lecture)). Nous dirons que attend=1 si l'élève a assisté au cours, et de 0 s'il n'y a pas assisté. De même, nous dirons que reading = 1 si l'élève a lu le manuel, et reading = 0 s'il ne l'a pas lu.

Bien, pour l'instant c'est assez simple. La prochaine chose que nous devons faire est d'enrober cela d'un peu de maths (désolé !). Pour les besoins de cet exemple, supposons que Y_p indique la note du *cinquième* élève de la classe. Ce n'est pas tout à fait la même notation que celle que nous avons utilisée plus tôt dans ce chapitre. Auparavant, nous avons utilisé la notation Y_{rci} pour désigner la *i-ème* personne du *r-ème* groupe pour le prédicteur 1 (le facteur de ligne) et le *c-ème* groupe pour le prédicteur 2 (le facteur de colonne). Cette notation générale était vraiment pratique pour décrire le calcul des SS, mais c'est une souffrance dans le contexte actuel, alors je vais changer de notation ici. Maintenant, la notation Y_p est visuellement plus simple que Y_{rci}, mais elle a le défaut de ne pas garder la trace des membres du groupe ! C'est-à-dire, si je vous disais que $Y_{0,0,3}$=35, vous sauriez immédiatement qu'il s'agit d'un étudiant (le 3e de ce type, en fait) qui n'a pas assisté aux cours (c.-à-d., attend=0) et n'a pas lu le manuel (c.-à-d., reading=0), et qui a échoué en cours (Grade=35). Mais si je vous dis que Y_p=35, tout ce que vous savez, c'est que le *p-ième*

étudiant n'a pas eu une bonne note. Nous avons perdu des informations clés. Bien sûr, il ne faut pas beaucoup de réflexion pour comprendre comment régler ce problème. Ce que nous allons faire à la place est d'introduire deux nouvelles variables X_{1p} et X_{2p} qui gardent la trace de ces informations. Dans le cas de notre étudiant hypothétique, nous savons que X_{1p}=0 (c.-à-d., attend = 0) et X_{2p}=0 (c.-à-d., reading=0). Les données pourraient donc ressembler à ceci :

personne, p	grade, Y_p	attendance, *X~1p	~* | lecture, X_{2p}
1	90	1	1
2	87	1	1
3	75	0	1
4	60	1	0
5	35	0	0
6	50	0	0
7	65	1	0
8	70	0	1

Il n'y a rien de particulier, bien sûr. C'est exactement le format dans lequel nous nous attendons à voir nos données ! Voir le fichier rtfm.csv. Nous pouvons utiliser l'analyse « Descriptives » de Jamovi pour confirmer que cet ensemble de données correspond à un plan équilibré, avec 2 observations pour chaque combinaison de attend et de read. De la même manière, nous pouvons également calculer la note moyenne pour chaque combinaison. C'est ce que montre la Figure 14-16. En regardant les notes moyennes, on a la forte impression que la lecture du texte et le fait d'assister aux cours sont très importants.

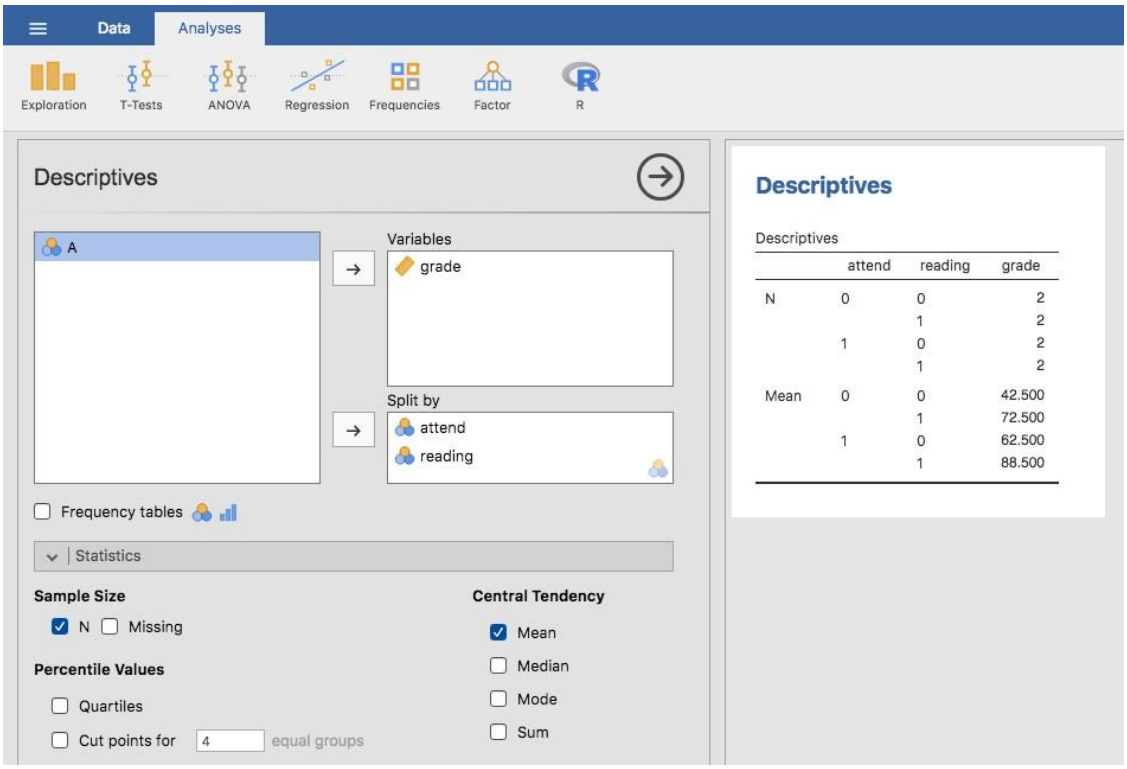


Figure 14-16: Statistiques descriptives dans Jamovi pour l'ensemble de données rtfm.csv

ANOVA avec des facteurs binaires comme modèle de régression

Bien, revenons aux mathématiques. Nous avons maintenant nos données exprimées avec trois variables numériques : la variable continue Y et les deux variables binaires X_1 et X_2. Ce que je veux que vous reconnaissiez, c'est que notre ANOVA factorielle 2x2 est *strictement* équivalente au modèle de régression.

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \epsilon_p$$

Bien sûr, c'est exactement la même équation que celle que j'ai utilisée plus tôt pour décrire un modèle de régression à deux prédicteurs ! La seule différence est que X_1 et X_2 sont maintenant des variables *binaires* (c.-à-d. que les valeurs ne peuvent être que 0 ou 1), alors que dans une analyse de régression, nous prévoyons que X_1 et X_2 seront continues. Il y a deux ou trois façons dont je pourrais essayer de vous en convaincre. Une possibilité serait de faire un long exercice mathématique pour prouver que les deux sont identiques. Cependant, je vais anticiper et deviner que la plupart des lecteurs de ce livre trouveront cela ennuyeux plutôt qu'utile. Au lieu de cela, j'expliquerai les idées de base et je m'appuierai sur Jamovi pour montrer que les analyses ANOVA et les analyses de régression ne sont pas seulement similaires, elles sont identiques. Commençons par faire une analyse de variance. Pour ce faire, nous utiliserons l'ensemble de données rtfm.csv, et regardons la Figure 14-17 qui montre ce qu'on obtient quand on fait l'analyse à Jamovi.

ANOVA

	Sum of Squares	df	Mean Square	F	p	η²	η²p	ω²
attend	648.00	1	648.00	21.60	0.00559	0.27	0.81	0.26
reading	1568.00	1	1568.00	52.27	0.00079	0.66	0.91	0.64
Residuals	150.00	5	30.00					

Figure 14-17 : ANOVA de l'ensemble de données rtfm.csv dans Jamovi, sans le terme d'interaction

En lisant les chiffres clés du tableau de l'ANOVA et les scores moyens que nous avons présentés plus haut, nous pouvons voir que les élèves ont obtenu une meilleure note s'ils ont suivi la classe ($F(1,5)=21,6$, p=.0056) et s'ils lisent le manuel ($F(1,5)=52.3$, p=.0008). Notons ces valeurs *p* et ces statistiques *F.*

Pensons maintenant à la même analyse dans une perspective de régression linéaire. Dans l'ensemble de données rtfm.csv, nous avons encodé attend et la read comme s'il s'agissait de prédicteurs numériques. Dans ce cas, c'est tout à fait acceptable. Il y a vraiment un sens au fait qu'un étudiant qui se présente en classe (c.-à-d. attend = 1) a en fait « plus de présence » qu'un étudiant qui ne le fait pas (c.-à-d. attend = 0). Il n'est donc pas du tout déraisonnable de l'inclure comme prédicteur dans un modèle de régression. C'est un peu inhabituel, car le prédicteur ne prend que deux valeurs possibles, mais il ne viole aucune des hypothèses de la régression linéaire. Et c'est facile à interpréter. Si le coefficient de régression est supérieur à 0, cela signifie que les étudiants qui assistent à des cours ont des notes plus élevées. Si elle est inférieure à zéro, les étudiants qui assistent à des cours magistraux obtiennent des notes moins élevées. Il en va de même pour notre variable de read.

Attendez une seconde. *Pourquoi* est-ce vrai ? C'est quelque chose qui est intuitivement évident pour tous ceux qui ont suivi quelques cours de statistiques et qui sont à l'aise avec les mathématiques, mais ce *n'est pas* clair pour tout le monde au premier abord. Pour comprendre pourquoi c'est vrai, il est utile d'examiner de près quelques élèves en particulier. Commençons par considérer les 6e et 7e élèves de notre ensemble de données (c.-à-d. $p = 6$ et $p = 7$). Ni l'un ni l'autre n'a lu le manuel, de sorte que dans les deux cas, nous pouvons fixer read à 0, ou, pour dire la même chose dans notre notation mathématique, nous observons $X_{2,6}=0$ et $X_{2,7}=0$, mais l'étudiant numéro 7 est venu aux cours (c'est à dire attend = 1, $X_{1,7}=1$) tandis que l'étudiant numéro 6 ne l'est pas (c'est-à-dire attend=0, $*X_{1,6}=0$). Voyons maintenant ce qui se passe lorsque nous insérons ces nombres dans la formule générale de notre ligne de régression. Pour l'élève numéro 6, la régression prédit que

$$\begin{aligned} \hat{Y}_6 &= b_0 + b_1 X_{1,6} + b_2 X_{2,6} \\ &= b_0 + (b_1 \times 0) + (b_2 \times 0) \\ &= b_0 \end{aligned}$$

On s'attend donc à ce que cet élève obtienne une note correspondant à la valeur du terme d'intersection b_0. Et l'élève 7 ? Cette fois, lorsque nous insérons les nombres dans la formule de la ligne de régression, nous obtenons ce qui suit

$$\begin{aligned}\hat{Y}_7 &= b_0 + b_1 X_{1,7} + b_2 X_{2,7} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 0) \\ &= b_0 + b_1\end{aligned}$$

Étant donné que cet élève a fréquenté la classe, la note prévue est égale au terme d'intersection b_0 plus le coefficient associé à la variable attend, b_1. Donc, si b_1 est supérieur à zéro, nous nous attendons à ce que les étudiants qui se présentent aux cours magistraux obtiennent de meilleures notes que ceux qui ne le font pas. Si ce coefficient est négatif, on s'attend à l'inverse : les élèves qui se présentent en classe obtiennent de bien pires résultats. En fait, nous pouvons aller un peu plus loin. Qu'en est-il de l'élève numéro 1, qui s'est présenté en classe ($X_{1,1}$=1) et a lu le manuel ($X_{2,1}$=1) ? Si nous connectons ces chiffres à la régression, nous obtenons

$$\begin{aligned}\hat{Y}_1 &= b_0 + b_1 X_{1,1} + b_2 X_{2,2} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 1) \\ &= b_0 + b_1 + b_2\end{aligned}$$

Donc, si nous supposons que le fait d'aller en classe vous aide à obtenir une bonne note (c.-à-d. b_1>0) et si nous supposons que la lecture du manuel vous aide également à obtenir une bonne note (c.-à-d. $b2$ > 0), nous nous attendons à ce que l'élève 1 obtienne une note supérieure à celle des élèves 6 et 7.

Et à ce stade, vous ne serez pas du tout surpris d'apprendre que le modèle de régression prédit que l'étudiant 3, qui a lu le livre mais n'a pas assisté aux cours, obtiendra une note b_2>b_0. Je ne vous ennuierai pas avec une autre formule de régression. Je vais plutôt vous montrer le tableau suivant des *notes attendues* :

		read textbook?	
		no	yes
attended?	no	b_0	$b_0 + b_2$
	yes	$b_0 + b_1$	$b_0 + b_1 + b_2$

Comme vous pouvez le constater, le terme d'interception b_0 agit comme une sorte de note de base à laquelle on s'attendrait de la part des élèves qui ne prennent pas le temps d'aller en classe ou de lire le manuel scolaire. De même, b_1 représente l'augmentation que l'on s'attend à recevoir si vous venez en classe, et b_2 représente l'augmentation qui vient de la lecture du manuel scolaire. En fait, s'il s'agissait d'une ANOVA, vous pourriez très bien vouloir caractériser b_1 comme l'effet principal de la fréquentation, et b_2 comme l'effet principal de la lecture ! En fait, pour un simple 2 x 2 ANOVA, c'est *exactement* comme ça que ça se passe.

Ok, maintenant que nous commençons vraiment à voir pourquoi ANOVA et la régression sont fondamentalement la même chose, exécutons notre régression en utilisant les données rtfm.csv et l'analyse de régression de Jamovi pour nous convaincre que cela est vraiment vrai. L'exécution de la régression de la manière habituelle donne les résultats présentés à la Figure 14-18.

Model Coefficients

			95% Confidence Interval			
Predictor	**Estimate**	**SE**	**Lower**	**Upper**	**t**	**p**
Intercept	43.50	3.35	34.88	52.12	12.97	0.00005
attend	18.00	3.87	8.04	27.96	4.65	0.00559
reading	28.00	3.87	18.04	37.96	7.23	0.00079

Figure 14-18 : Analyse de régression de l'ensemble de donnéesrtfm.csv dans Jamovi, sans le terme d'interaction

Il y a quelques choses intéressantes à noter ici. Notons d'abord que le terme d'intersection est 43,5, ce qui est proche de la moyenne de 42,5 observée pour les deux élèves qui n'ont pas lu le texte ou qui n'ont pas assisté aux cours. Deuxièmement, nous avons le coefficient de régression b_1=18,0 pour la variable attend, ce qui suggère que les élèves qui ont assisté aux cours ont obtenu 18 % de plus que ceux qui ne l'ont pas fait. Nous nous attendions donc à ce que les élèves qui se présentaient en classe mais qui ne lisaient pas le manuel obtiennent une note de $b_0 + b_1$, ce qui est égal à 43,5 + 18,0 = 61,5. Vous pouvez vérifier par vous-même que la même chose se produit lorsque nous regardons les élèves qui lisent le manuel.

En fait, nous pouvons aller un peu plus loin dans l'établissement de l'équivalence de notre analyse de variance et de notre régression. Examinez les valeurs *p* associées à la variable attend et à la variable read dans la sortie de régression. Elles sont identiques à celles que nous avons rencontrées plus tôt lors de l'exécution de l'ANOVA. Cela peut paraître un peu surprenant, puisque le test utilisé lors de l'exécution de notre modèle de régression calcule une statistique *t* et que l'ANOVA calcule une statistique *F*. Cependant, si vous vous rappelez tout au long du chapitre 7, j'ai mentionné qu'il y a une relation entre la distribution *t* et la *distribution F*. Si vous avez une certaine quantité qui est distribuée selon une *distribution t* avec *k* degrés de liberté et que vous l'élevez au carré, alors cette nouvelle quantité au carré suit une distribution *F* dont les degrés de liberté sont 1 et *k*. Nous pouvons vérifier ceci par rapport aux statistiques *t* dans notre modèle de régression. Pour la variable attend, nous obtenons une valeur *t* de 4,65. Si nous élevons au carré ce nombre, nous obtenons 21,6, ce qui correspond à la statistique *F* correspondante dans notre analyse de variance.

Enfin, une dernière chose que vous devriez savoir. Parce que Jamovi intègre le fait que ANOVA et la régression sont deux exemples de modèles linéaires, il vous permet d'extraire la table ANOVA classique de votre modèle de régression en utilisant la « Linear Regression » - « Model Coefficients » - « Omnibus Test » - « ANOVA Test », et ceci vous donnera le tableau présenté dans la Figure 14-19.

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
attend	648.00	1	648.00	21.60	0.00559
reading	1568.00	1	1568.00	52.27	0.00079
Residuals	150.00	5	30.00		

Note. Type 3 sum of squares

Figure 14-19 : Résultats du test Omnibus ANOVA de l'analyse de régression de Jamovi

Comment coder les facteurs non binaires en tant que contrastes

A ce stade, je vous ai montré comment nous pouvons visualiser un 2x2 ANOVA dans un modèle linéaire. Et il est assez facile de voir comment cela se généralise à une ANOVA 2 x 2 x 2 x 2 ou ANOVA une 2 x 2 x 2 x 2 x 2. C'est la même chose, vraiment. Vous ajoutez simplement une nouvelle variable binaire pour chacun de vos facteurs. Ce qui commence à se compliquer, c'est quand on considère les facteurs qui ont plus de deux niveaux. Prenons, par exemple, l'analyse de variance 3 x 2 que nous avons effectuée plus tôt dans ce chapitre à l'aide des données de clinicaltrial.csv. Comment pouvons-nous convertir le facteur médicament à trois niveaux en une forme numérique qui convient à une régression ?

La réponse à cette question est assez simple, en fait. Tout ce que nous avons à faire est de réaliser qu'un facteur à trois niveaux peut être réécrit comme *deux* variables binaires. Supposons, par exemple, que je crée une nouvelle variable binaire appelée druganxifree. Chaque fois que la variable drug est égale à « anxifree » on met druganxifree = 1. Cette variable établit un **contraste**, dans ce cas-ci entre anxifree et les deux autres médicaments. En soi, bien sûr, le contraste druganxifree n'est pas suffisant pour saisir toute l'information de notre variable drug. Nous avons besoin d'un deuxième contraste, un contraste qui nous permette de distinguer le joyzepam du placebo. Pour ce faire, nous pouvons créer un second contraste binaire, appelé drugjoyzepam, qui est égal à 1 si le médicament est le joyzepam et à 0 s'il ne l'est pas. Ensemble, ces deux contrastes nous permettent d'établir une distinction parfaite entre les trois drogues possibles. Le tableau ci-dessous l'illustre bien :

drug	druganxifree	drugjoyzepam
"placebo"	0	0
"anxifree"	1	0
"joyzepam"	0	1

Si le médicament administré à un patient est un placebo, les deux variables de contraste seront égales à 0 ; si le médicament est Anxifree, la variable druganxifree sera égale à 1, et

drugjoyzepam sera égal à 0 ; l'inverse est vrai pour Joyzepam : drugjoyzepam est égal à 1 et druganxifree est égal à 0.

Créer des variables de contraste n'est pas trop difficile à faire à l'aide de la commande de calcul de nouvelles variables de Jamovi. Par exemple, pour créer la variable druganxifree, écrivez cette expression logique dans la boîte de calcul de la nouvelle variable : IF(drug ==‘anxifree’, 1, 0)‘. De même, pour créer la nouvelle variable drugjoyzepam utiliser cette expression logique : IF(drug ==’joyzepam’, 1, 0). Il en va de même pour la CBT Therapy : IF(therapy ==‘CBT’, 1, 0). Vous pouvez voir ces nouvelles variables, et les expressions logiques correspondantes, dans le fichier de données Jamovi clinicaltrial2.omv.

Nous avons maintenant recodé notre facteur à trois niveaux en termes de deux variables binaires, et nous avons déjà vu que l'ANOVA et la régression se comportent de la même manière pour les variables binaires. Toutefois, d'autres complexités surgissent dans ce cas, dont nous parlerons dans la section suivante.

L'équivalence entre ANOVA et régression pour les facteurs non binaires

Nous avons maintenant deux versions différentes du même ensemble de données. Nos données originales dans lesquelles la variable drug du fichier clinicaltrial.csv est exprimée comme un seul facteur à trois niveaux, et les données étendues clinicaltrial2.omv dans lesquelles elle est développée en deux contrastes binaires. Encore une fois, ce que nous voulons démontrer, c'est que notre ANOVA factorielle originale de 3 x 2 est équivalente à un modèle de régression appliqué aux variables de contraste. Commençons par relancer l'analyse de variance, dont les résultats sont présentés à la Figure 14-20.

ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	η^2p	ω^2
drug	3.45	2	1.73	26.15	0.00002	0.71	0.79	0.68
therapy	0.47	1	0.47	7.08	0.01866	0.10	0.34	0.08
Residuals	0.92	14	0.07					

Figure 14-20 : résultats de l'analyse de variance de Jamovi, sans composante d'interaction

Évidemment, il n'y a pas de surprise ici. C'est exactement la même analyse de variance que celle qu'on a faite tout à l'heure. Ensuite, effectuons une régression en utilisant comme prédicteurs le druganxifree, le drugjoyzepam et la CBTtherapy. Les résultats sont présentés à la Figure 14-21.

Model Coefficients

Predictor	Estimate	SE	t	p
Intercept	0.29	0.12	2.38	0.03178
druganxifree	0.27	0.15	1.80	0.09386
drugjoyzepam	1.03	0.15	6.97	< .00001
CBTtherapy	0.32	0.12	2.66	0.01866

Figure 14-21 : résultats de la régression de Jamovi, avec variables de contraste druganxifree et drugjoyzepam

Ouais. Ce n'est pas le même résultat que la dernière fois. Comme on pouvait s'y attendre, la sortie de régression donne les résultats de chacun des trois prédicteurs séparément, tout comme elle l'a fait chaque fois que nous avons effectué une analyse de régression. D'une part, nous pouvons voir que la valeur *p* de la variable CBTtherapy est exactement la même que celle du facteur therapy dans notre analyse de variance originale, de sorte que nous pouvons être rassurés que le modèle de régression fait la même chose que l'analyse de variance. D'autre part, ce modèle de régression teste *séparément* le contraste du médicament sansanxiforme et le contraste du médicament joyzepam, comme s'il s'agissait de deux variables complètement différentes. Ce n'est pas surprenant, bien sûr, parce que l'analyse de régression médiocre n'a aucun moyen de savoir que le drugjoyzepam et le druganxifree sont en fait les deux contrastes différents que nous utilisons pour coder notre facteur drug à trois niveaux. Pour autant qu'elle le sache, le drugjoyzepam et le druganxifree ne sont pas plus apparentés que le drugjoyzepam et la therapyCBT. Cependant, nous savons que c'est mieux. À ce stade, nous ne sommes pas du tout intéressés à déterminer si ces deux contrastes sont significatifs individuellement. Nous voulons juste savoir s'il y a un effet « global » du médicament. C'est-à-dire, ce que *nous* voulons que Jamovi fasse, c'est de faire une sorte de test de « comparaison de modèles », un test dans lequel les deux contrastes « liés au médicament » sont mis dans le même panier pour les besoins du test. Ça vous dit quelque chose ? Tout ce que nous avons à faire est de spécifier notre modèle d'hypothèse nulle, qui dans ce cas inclurait le prédicteur de la CBTherapy, et d'omettre les deux variables liées au médicament, comme dans la Figure 14-22.

Bien, c'est mieux comme ça. Notre statistique *F* est de 26,15, les degrés de liberté sont 2 et 14, et la valeur *p* est 0,0000002. Les chiffres sont identiques à ceux que nous avons obtenus pour l'effet principal du facteur drug dans notre analyse de variance originale. Encore une fois, nous constatons que l'analyse de variance et la régression sont fondamentalement les identiques. Il s'agit de deux modèles linéaires, et le mécanisme statistique sous-jacent de l'analyse de variance est identique à celui utilisé pour la régression. L'importance de ce fait ne doit pas être sous-estimée. Tout au long de ce chapitre, nous allons nous appuyer fortement sur cette idée.

Bien que nous ayons passé en revue tous les défauts du calcul de nouvelles variables dans Jamovi pour les contrastes druganxifree et drugjoyzepam, juste pour montrer que l'ANOVA et la régression sont fondamentalement les mêmes, dans l'analyse de régression linéaire de Jamovi il existe en fait un raccourci pratique pour obtenir ceux-ci. voir Figure 14-23.

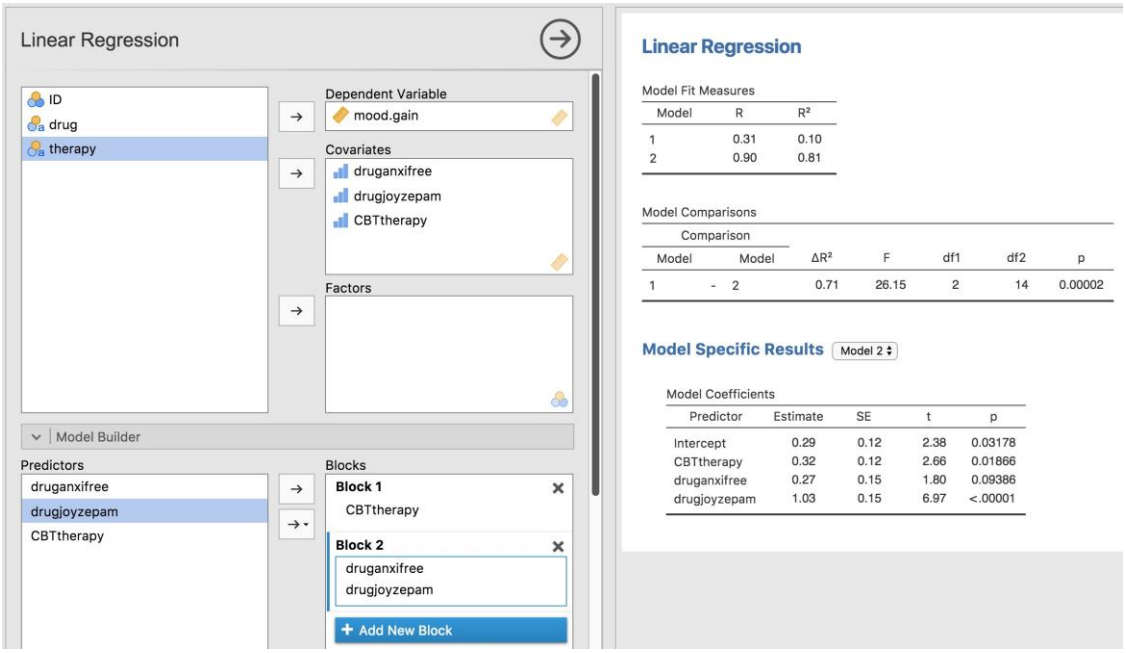


Figure 14-22 : Comparaison des modèles de régression dans Jamovi, modèle d'hypothèse nul 1 vs modèle de contraste 2

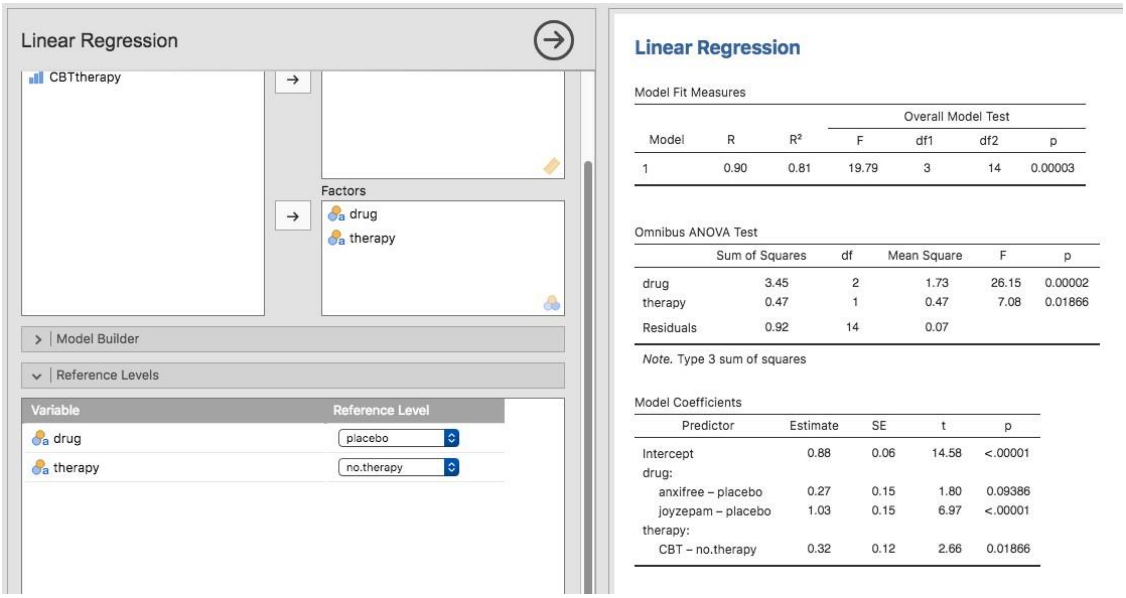


Figure 14-23 : Analyse de régression avec facteurs et contrastes dans Jamovi, y compris omnibus

Ce que Jamovi fait ici, c'est vous permettre d'entrer les variables prédictrices comme des facteurs, attendez, comme des …facteurs ! Intelligent, non. Vous pouvez également spécifier le groupe à utiliser comme niveau de référence, via l'option « Reference Levels ». Nous l'avons changé pour « placebo » et « no.therapy », respectivement, parce que c'est le plus logique.

Si vous cliquez également sur la case à cocher 'ANOVA' sous l'option 'Model Coefficients' - 'Omnibus Test', nous voyons que la statistique *F* est 26,15, les degrés de liberté sont 2 et 14, et la valeur *p* est 0,00002 (Figure 14-23). Les chiffres sont identiques à ceux que nous avons obtenus pour l'effet principal de drug dans notre analyse de variance originale. Encore une fois, nous constatons que l'analyse de variance et la régression sont fondamentalement les mêmes. Il s'agit de deux modèles linéaires, et le mécanisme statistique sous-jacent de l'analyse de variance est identique à celui utilisé pour la régression.

Degrés de liberté comme comptage de paramètres !

Enfin, je peux enfin donner une définition des degrés de liberté dont je suis satisfait. Les degrés de liberté sont définis en fonction du nombre de paramètres qui doivent être estimés dans un modèle. Pour un modèle de régression ou une analyse de variance, le nombre de paramètres correspond au nombre de coefficients de régression (c.-à-d. les valeurs *b*), y compris l'intersection. En gardant à l'esprit que tout *test F* est toujours une comparaison entre deux modèles et le premier *df* est la différence du nombre de paramètres. Par exemple, dans la comparaison de modèles ci-dessus, le modèle nul (mood.gain ~ therapyCBT) a deux paramètres : il y a un coefficient de régression pour la variable therapyCBT, et un autre pour l'interception. Le modèle alternatif (mood.gain ~ druganxifree + drugjoyzepam + therapyCBT) a quatre paramètres : un coefficient de régression pour chacun des trois contrastes, et un autre pour l'interception. Le degré de liberté associé à la *différence* entre ces deux modèles est donc df_1=4-2=2.

Qu'en est-il du cas où il ne semble pas y avoir de modèle d'hypothèse nulle ? Par exemple, vous pouvez penser au test *F* qui apparaît lorsque vous sélectionnez « Test F » dans les options « Linear Regression » - « Model Fit ». J'ai d'abord décrit cela comme un test du modèle de régression dans son ensemble. Toutefois, il s'agit toujours d'une comparaison entre deux modèles. Le modèle de l'hypothèse nulle est le modèle trivial qui ne comprend qu'un seul coefficient de régression, pour le terme d'intersection. Le modèle alternatif contient *K*+1 coefficients de régression, un pour chacune des *K* variables prédicteurs et un autre pour l'interception. Donc la valeur *df* que vous voyez dans ce test *F* est égale à *df1*=*K*+1-1=*K*.

Qu'en est-il de la deuxième valeur *df* qui apparaît dans le test *F* ? Il s'agit toujours des degrés de liberté associés aux résidus. Il est également possible d'envisager cela en termes de paramètres, mais d'une manière légèrement contre-intuitive. Pensez-y comme ça. Supposons que le nombre total d'observations dans l'ensemble de l'étude est *N*. Si vous voulez décrire *parfaitement* chacune de ces *N* valeurs, vous devez le faire en utilisant, eh bien…. *N* nombres. Lorsque vous construisez un modèle de régression, ce que vous faites en réalité, c'est de spécifier que certains des nombres doivent parfaitement décrire les données. Si votre modèle a *K* prédicteurs et une intersection, alors vous avez spécifié *K* + 1 nombres. Donc, sans prendre la peine de déterminer exactement *comment* cela serait fait, combien d'autres chiffres faudra-t-il, selon vous, pour transformer un modèle de régression de paramètres *K+1* en une description parfaite des données brutes ? Si vous pensez *(K+1)+(N-K-1q)=N*, et donc que la réponse devrait être *N*–*K*-1, vous avez gagné ! C'est exactement ça. En principe, vous pouvez imaginer un modèle de régression d'une complexité absurde qui inclut un paramètre pour chaque point de données unique, et qui

fournirait bien sûr une description parfaite des données. Ce modèle contiendrait *N* paramètres au total, mais nous nous intéressons à la différence entre le nombre de paramètres requis pour décrire ce modèle complet (c.-à-d. *N*) et le nombre de paramètres utilisés par le modèle de régression plus simple qui vous intéresse réellement (c.-à-d. *K+1*), et donc le deuxième degré de liberté du test *F* est $df_2 = N-K-1$, où *K* est le nombre de variables explicatives (dans un modèle de régression) ou le nombre de contrastes (dans une ANOVA). Dans l'exemple que j'ai donné ci-dessus, il y a *N*=18 observations dans l'ensemble de données et *K*+1=4 coefficients de régression associés au modèle ANOVA, donc les degrés de liberté des résidus sont df_2=18-4=14.

Différentes façons de spécifier les contrastes

Dans la section précédente, je vous ai montré une méthode pour convertir un facteur en une collection de contrastes. Dans la méthode que je vous ai montrée, nous spécifions un ensemble de variables binaires dans lequel nous avons défini une table comme celle-ci :

drug	druganxifree	drugjoyzepam
"placebo"	0	0
"anxifree"	1	0
"joyzepam"	0	1

Chaque ligne du tableau correspond à l'un des niveaux de facteurs et chaque colonne correspond à l'un des contrastes. Cette table, qui a toujours une ligne de plus que les colonnes, a un nom spécial. C'est ce qu'on appelle une **matrice de contrastes**. Cependant, il existe de nombreuses façons de spécifier une matrice de contrastes. Dans cette section, j'aborde quelques-unes des matrices de contrastes standard utilisées par les statisticiens et la façon dont vous pouvez les utiliser dans les Jamovi. Si vous avez l'intention de lire la section sur l'ANOVA non équilibrée plus loin (Section 14.10), cela vaut la peine de lire attentivement cette section. Si ce n'est pas le cas, vous pouvez vous contenter de la survoler, car le choix des contrastes n'a pas beaucoup d'importance pour des motifs équilibrés.

Les contrastes de traitement

Dans le type particulier de contrastes que j'ai décrit ci-dessus, un niveau du facteur est spécial et agit comme une sorte de catégorie de « référence » (c.-à-d. placebo dans notre exemple), par rapport à laquelle les deux autres sont définis. Le nom de ce type de contrastes est celui de **contraste de traitement**, également connus sous le nom de « faux codage ». Dans ce contraste, chaque niveau du facteur est comparé à un niveau de référence de base, et le niveau de référence de base est la valeur de l'interception.

Le nom reflète le fait que ces contrastes sont tout à fait naturels et raisonnables quand l'une des catégories de votre facteur est vraiment spéciale parce qu'elle représente en fait une

référence. C'est logique dans notre exemple d'essai clinique. L'état placebo correspond à la situation où vous ne donnez pas de vrais médicaments aux gens, et c'est donc particulier. Les deux autres conditions sont définies par rapport au placebo. Dans un cas, vous remplacez le placebo par Anxifree et dans l'autre par Joyzepam.

Le tableau ci-dessus est une matrice des contrastes de traitement pour un facteur à 3 niveaux. Mais supposons que je veuille une matrice des contrastes de traitement pour un facteur à 5 niveaux ? Ce serait quelque chose comme ça :

Level	2	3	4	5
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

Dans cet exemple, le premier contraste est de niveau 2 comparé au niveau 1, le second contraste est de niveau 3 comparé au niveau 1, et ainsi de suite. Notez que, par défaut, le *premier* niveau du facteur est toujours traité comme la catégorie de base (c'est-à-dire celle qui a tous les zéros et qui n'est pas associée à un contraste explicite). Dans Jamovi vous pouvez choisie quelle catégorie est le premier niveau du facteur en manipulant l'ordre des niveaux de la variable affichée dans la fenêtre « Data Variable » (double-cliquez sur le nom de la variable dans la colonne de la feuille de calcul pour faire apparaître la fenêtre « Data Variable ».

Contraste de Helmert

Les contrastes de traitement sont utiles dans de nombreuses situations. Toutefois, elles sont plus sensées dans une situation où il y a vraiment une catégorie de référence, et vous voulez évaluer tous les autres groupes par rapport à cette catégorie. Dans d'autres situations, cependant, il n'existe pas de catégorie de référence et il peut être plus logique de comparer chaque groupe à la moyenne des autres groupes. C'est là que nous rencontrons les **contrastes de Helmert**, générés par l'option « Helmert » dans la boîte de sélection Jamovi « ANOVA » – « Contrasts ». L'idée derrière les contrastes de Helmert est de comparer chaque groupe à la moyenne des « précédents ». C'est-à-dire que le premier contraste représente la différence entre le groupe 2 et le groupe 1, le second contraste représente la différence entre le groupe 3 et la moyenne des groupes 1 et 2, etc. Cela se traduit par une matrice de contraste qui ressemble à celle-ci pour un facteur à cinq niveaux :

1	−1	−1	−1	−1
2	1	−1	−1	−1
3	0	2	−1	−1
4	0	0	3	−1
5	0	0	0	4

Une chose utile à propos des contrastes de Helmert est que chaque contraste est égal à zéro (c'est-à-dire que toutes les colonnes sont égales à zéro). Ceci a pour conséquence que, lorsque nous interprétons l'ANOVA comme une régression, l'intersection correspond à la grande moyenne $\mu_{..}$ si nous utilisons les contrastes de Helmert. Comparez cela aux contrastes de traitement, dans lesquels le terme d'intersection correspond à la moyenne du groupe pour la catégorie de référence. Cette propriété peut être très utile dans certaines situations. Ce n'est pas très important si vous avez un plan équilibré, ce que nous avons supposé jusqu'à présent, mais cela s'avérera important plus tard si nous considérons les plans non équilibrées dans la Section 14.10. En fait, la principale raison pour laquelle j'ai même pris la peine d'inclure cette section est que les contrastes deviennent importants si vous voulez comprendre l'analyse de variance non équilibrée.

Somme des contrastes à zéro

La troisième option que je dois mentionner brièvement est celle des contrastes avec une « somme à zéro », appelés contrastes « simples » en Jamovi, qui sont utilisés pour construire des comparaisons par paires entre groupes. Plus précisément, chaque contraste code la différence entre l'un des groupes et une catégorie de base, qui dans ce cas correspond au premier groupe :

1	−1	−1	−1	−1
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

Tout comme les contrastes de Helmert, nous voyons que chaque colonne a un total à zéro, ce qui signifie que le terme d'intersection correspond à la grande moyenne lorsque ANOVA est traité comme un modèle de régression. Lorsqu'on interprète ces contrastes, il faut reconnaître que chacun de ces contrastes est une comparaison par paires entre le groupe 1 et l'un des quatre autres groupes. Plus précisément, le contraste 1 correspond à une comparaison « groupe 2 moins groupe 1 », le contraste 2 correspond à une comparaison « groupe 3 moins groupe 1 », et ainsi de suite.[128]

Contraste optionnel en Jamovi

Jamovi est fourni également avec une variété d'options qui peuvent générer différents types de contrastes dans l'ANOVA. Celles-ci se trouvent dans l'option « Contrasts » de la fenêtre principale de l'analyse ANOVA, où les types de contraste suivants sont listés :

Type de contraste	
Déviation	Comparer la moyenne de chaque niveau (sauf une catégorie de référence) à la moyenne de tous les niveaux (moyenne générale).
Simple	Tout comme les contrastes du traitement, le contraste simple compare la moyenne de chaque niveau à la moyenne d'un niveau donné. Ce type de contraste est utile lorsqu'il y a un groupe témoin. Par défaut, la première catégorie est la référence. Cependant, avec un simple contraste, l'interception est la grande moyenne de tous les niveaux des facteurs.
Différence	Compare la moyenne de chaque niveau (sauf le premier) à la moyenne des niveaux précédents. (Parfois appelé contraste Helmert inversé)
Helmert	Comparer la moyenne de chaque niveau du facteur (sauf le dernier) à la moyenne des niveaux subséquents.
Répété	Comparer la moyenne de chaque niveau (sauf le dernier) à la moyenne du niveau suivant.

[128] Je vous entends demander : quelle est la différence entre le traitement et les contrastes simples ? Eh bien, à titre d'exemple, considérons un effet principal de genre, avec m=0 et f=1. Le coefficient correspondant au contraste du traitement mesurera la différence de moyenne entre les femmes et les hommes, et l'intersection sera la moyenne des hommes. Cependant, avec un simple contraste, c'est-à-dire m=-1 et f=1, l'intersection est la moyenne des moyennes et l'effet principal est la différence entre la moyenne de chaque groupe et l'intersection.

Polynôme	Compare l'effet linéaire et l'effet quadratique. Le premier degré de liberté contient l'effet linéaire dans toutes les catégories ; le second degré de liberté, l'effet quadratique. Ces contrastes sont souvent utilisés pour estimer les tendances polynomiales

Tests post hoc

Il est temps de changer de sujet. Plutôt que de faire des comparaisons préétablies que vous avez testées en utilisant des contrastes, supposons que vous ayez fait votre analyse de variance et qu'il s'avère que vous avez obtenu certains effets significatifs. Étant donné que les *tests F* sont des tests « omnibus » qui ne testent que l'hypothèse nulle qu'il n'y a pas de différences entre les groupes, l'obtention d'un effet significatif ne vous dit pas quels groupes sont différents des autres. Nous avons discuté de cette question au chapitre 13 et, dans ce chapitre, notre solution consistait à effectuer des *tests t* pour toutes les paires de groupes possibles, en effectuant des corrections pour les comparaisons multiples (p. ex., Bonferroni, Holm) afin de contrôler le taux d'erreur de type I dans toutes les comparaisons. Les méthodes que nous avons utilisées au chapitre 13 ont l'avantage d'être relativement simples et d'être le genre d'outils que vous pouvez utiliser dans un grand nombre de situations différentes où vous testez plusieurs hypothèses, mais ce ne sont pas nécessairement les meilleurs choix si vous êtes intéressé à faire des tests post hoc efficaces dans un contexte ANOVA. Il existe en fait un grand nombre de méthodes différentes pour effectuer des comparaisons multiples dans la littérature statistique (Hsu 1996), et il serait au-delà de la portée d'un texte d'introduction comme celui-ci d'en discuter tous en détail.

Cela dit, il y a un outil sur lequel je veux attirer votre attention, à savoir « Honestly Significant Difference » de Tukey, ou **HSD de Tukey** pour faire court. Pour une fois, je vais vous épargner les formules et m'en tenir aux idées qualitatives. L'idée de base du HSD de Tukey est d'examiner toutes les comparaisons pertinentes par paires entre les groupes, et il n'est vraiment approprié d'utiliser le HSD de Tukey que si ce sont les différences *par paires* qui vous intéressent.[129] Par exemple, nous avons déjà effectué une analyse de variance factorielle à l'aide de l'ensemble de données du clinicaltrial.csv, et après avoir précisé un effet principal du médicament et un effet principal du traitement, nous serions intéressés par les quatre comparaisons suivantes :

- La différence de gain d'humeur entre les personnes ayant reçu Anxifree et celles ayant reçu le placebo.
- La différence de gain d'humeur entre les personnes recevant le Joyzepam et celles recevant le placebo.

---

[129] Si, par exemple, vous souhaitez savoir si le groupe A est significativement différent de la moyenne du groupe B et du groupe C, vous devez utiliser un outil différent (par exemple, la méthode de Scheffe, qui est plus conservatrice, et qui dépasse le cadre du présent livre). Cependant, dans la plupart des cas, vous êtes probablement intéressé par les différences de groupes par paires, donc le HSD de Tukey est une chose assez utile à connaître.

- La différence de gain d'humeur entre les personnes ayant reçu Anxifree et celles ayant reçu Joyzepam.
- La différence de gain d'humeur entre les personnes traitées avec la TCC et celles qui n'ont pas reçu de thérapie.

Pour l'une ou l'autre de ces comparaisons, nous nous intéressons à la différence réelle entre les moyennes des groupes (de population). Le HSD de Tukey construit des **intervalles de confiance simultanés** pour ces quatre comparaisons. Ce que nous entendons par intervalle de confiance « simultané » à 95 %, c'est que, si nous devions répéter cette étude plusieurs fois, alors dans 95 % des résultats de l'étude, les intervalles de confiance contiendraient la vraie valeur pertinente. De plus, nous pouvons utiliser ces intervalles de confiance pour calculer une valeur *p* ajustée pour une comparaison spécifique.

La fonction TukeyHSD dans Jamovi est assez facile à utiliser. Vous spécifiez simplement les termes de modèle ANOVA pour lequel vous voulez exécuter les tests post hoc. Par exemple, si nous cherchions à effectuer des tests post hoc pour les effets principaux mais pas pour l'interaction, nous ouvririons l'option « Post Hoc Tests » dans l'écran d'analyse ANOVA, déplacerions les variables du drug et de la therapy vers la case de droite, puis sélectionnerions la case « Tukey » dans la liste des corrections post hoc qui pourraient être appliquées. La Figure 14-24 illustre ces choix, ainsi que le tableau des résultats correspondant.

Les résultats présentés dans le tableau des résultats des « tests post hoc » sont (je l'espère) assez simples. La première comparaison, par exemple, est la différence Anxifree versus placebo, et la première partie du résultat indique que la différence observée dans les moyennes de groupe est 0,27. Le chiffre suivant est l'erreur-type pour la différence, à partir de laquelle nous pourrions calculer l'intervalle de confiance à 95 % si nous le voulions, bien que Jamovi ne propose pas actuellement cette option. Il y a ensuite une colonne avec les degrés de liberté, une colonne avec la valeur *t*, et enfin une colonne avec la valeur *p*. Pour la première comparaison, la valeur *p* ajustée est .21. Par contre, si vous regardez la ligne suivante, nous voyons que la différence observée entre le joyzepam et le placebo est de 1,03, et ce résultat est significatif *p*<.001q.

Pour l'instant, tout va bien. Qu'en est-il de la situation où votre modèle inclut des termes d'interaction ? Par exemple, l'option par défaut dans Jamovi est de tenir compte de la possibilité qu'il y ait une interaction entre le médicament et la thérapie. Si c'est le cas, le nombre de comparaisons par paires dont nous avons besoin va considérer commence à augmenter.

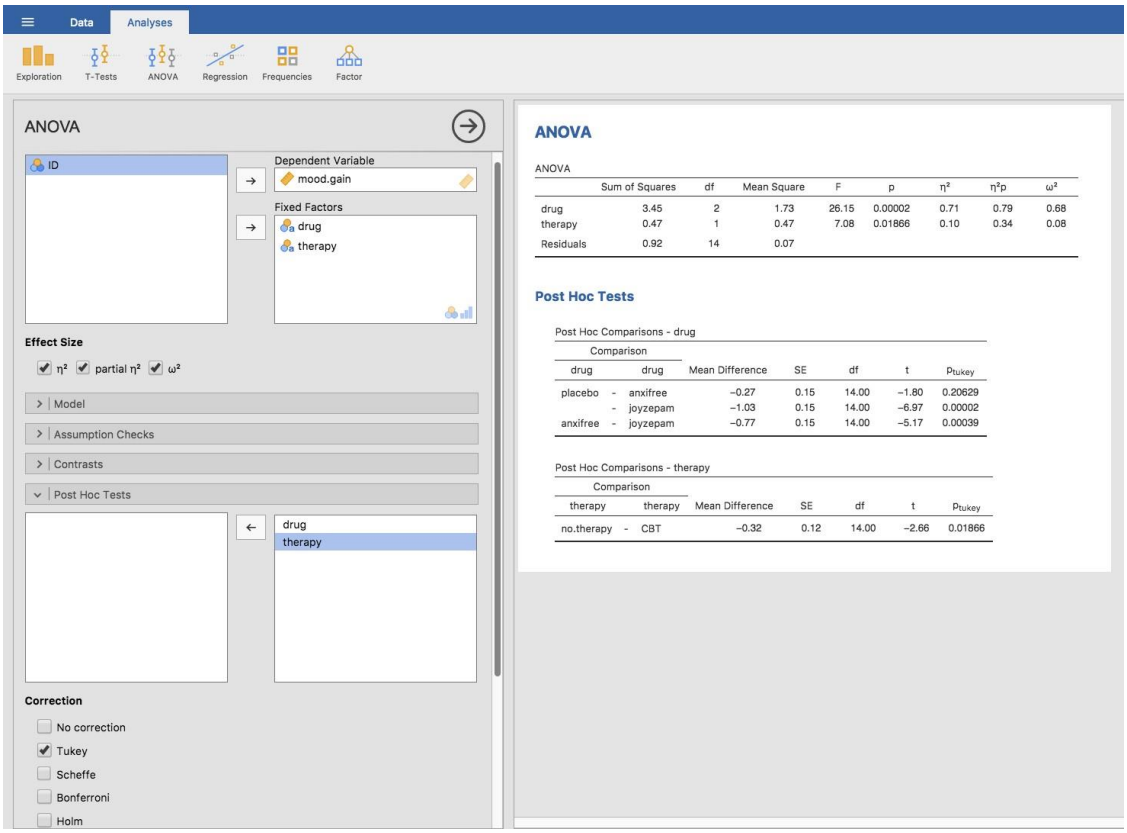


Figure 14-24 : Test post-hoc Tukey HSD dans l'ANOVA factorielle Jamovi, sans terme d'interaction

Comme par le passé, nous devons considérer les trois comparaisons pertinentes à l'effet principal de drug et la comparaison pertinente à l'effet principal du therapy. Mais, si nous voulons envisager la possibilité d'une interaction significative (et essayer de trouver les différences de groupe qui sous-tendent cette interaction significative), nous devons inclure des comparaisons telles que les suivantes :

- La différence de gain d'humeur entre les personnes ayant reçu Anxifree et traitées avec la TCC et les personnes ayant reçu le placebo et traitées avec la TCC.
- La différence de gain d'humeur entre les personnes ayant reçu Anxifree et celles n'ayant reçu aucun traitement, et celles ayant reçu le placebo et n'ayant reçu aucun traitement.
- Etc.

Il y a beaucoup de ces comparaisons dont vous devez tenir compte. Ainsi, lorsque nous effectuons l'analyse post hoc de Tukey pour ce modèle ANOVA, nous constatons qu'il a fait *beaucoup* de comparaisons par paires (19 au total), comme le montre la Figure 14-25. Vous pouvez voir qu'il ressemble beaucoup à ce qu'il était avant, mais avec beaucoup plus de comparaisons faites.

Post Hoc Tests

Post Hoc Comparisons - drug

Comparison			Mean Difference	SE	df	t	ptukey
drug		drug					
placebo	-	anxifree	-0.27	0.13	12.00	-1.98	0.15971
	-	joyzepam	-1.03	0.13	12.00	-7.67	0.00002
anxifree	-	joyzepam	-0.77	0.13	12.00	-5.69	0.00027

Post Hoc Comparisons - therapy

Comparison			Mean Difference	SE	df	t	ptukey
therapy		therapy					
no.therapy	-	CBT	-0.32	0.11	12.00	-2.93	0.01262

Post Hoc Comparisons - drug ✻ therapy

Comparison					Mean Difference	SE	df	t	ptukey
drug	therapy		drug	therapy					
placebo	no.therapy	-	placebo	CBT	-0.30	0.19	12.00	-1.57	0.62800
		-	anxifree	no.therapy	-0.10	0.19	12.00	-0.52	0.99401
		-	anxifree	CBT	-0.73	0.19	12.00	-3.85	0.02187
		-	joyzepam	no.therapy	-1.17	0.19	12.00	-6.12	0.00057
		-	joyzepam	CBT	-1.20	0.19	12.00	-6.30	0.00044
	CBT	-	anxifree	no.therapy	0.20	0.19	12.00	1.05	0.89172
		-	anxifree	CBT	-0.43	0.19	12.00	-2.27	0.27506
		-	joyzepam	no.therapy	-0.87	0.19	12.00	-4.55	0.00676
		-	joyzepam	CBT	-0.90	0.19	12.00	-4.72	0.00507
anxifree	no.therapy	-	anxifree	CBT	-0.63	0.19	12.00	-3.32	0.05298
		-	joyzepam	no.therapy	-1.07	0.19	12.00	-5.60	0.00126
		-	joyzepam	CBT	-1.10	0.19	12.00	-5.77	0.00096
	CBT	-	joyzepam	no.therapy	-0.43	0.19	12.00	-2.27	0.27506
		-	joyzepam	CBT	-0.47	0.19	12.00	-2.45	0.21392
joyzepam	no.therapy	-	joyzepam	CBT	-0.03	0.19	12.00	-0.17	0.99997

Figure 14-25 : Test post hoc de Tukey HSD dans l'ANOVA factorielle de Jamovi avec un terme d'interaction

La méthode des comparaisons planifiées

Dans le prolongement des sections précédentes sur les contrastes et les tests post hoc dans ANOVA, je pense que la méthode des comparaisons planifiées est suffisamment importante pour mériter une discussion rapide. Dans nos discussions sur les comparaisons multiples, dans la section précédente et au chapitre 13, j'ai supposé que les tests que vous voulez effectuer sont vraiment post hoc. Par exemple, dans notre exemple de médicaments ci-dessus, vous pensiez peut-être que les médicaments auraient tous des effets différents sur l'humeur (c.-à-d. que vous avez émis l'hypothèse d'un effet principal du médicament), mais vous n'aviez aucune hypothèse précise sur la façon dont ils diffèreraient, ni aucune idée réelle des comparaisons par paires qu'il serait utile d'examiner. Si c'est le cas, alors vous devez vraiment recourir à quelque chose comme le HSD de Tukey pour faire vos comparaisons par paires.

La situation est assez différente, cependant, si vous aviez réellement des hypothèses spécifiques sur lesquelles les comparaisons sont intéressantes, et que vous n'avez *jamais* eu l'intention de regarder d'autres comparaisons que celles que vous avez spécifiées à l'avance. Quand c'est vrai, et si vous vous en tenez honnêtement et rigoureusement à vos nobles intentions de ne pas faire d'autres comparaisons (même lorsque les données semblent vous montrer des effets délicieusement significatifs pour des choses pour lesquelles vous n'aviez pas de test d'hypothèse), alors cela n'a pas vraiment de sens de faire quelque chose comme le HSD de Tukey, car il apporte des corrections pour toute une série de comparaisons que vous n'avez jamais voulu et auxquelles vous ne vous êtes jamais intéressés. Dans ces circonstances, vous pouvez effectuer un nombre (limité) de tests d'hypothèse sans avoir à faire d'ajustement pour plusieurs tests. Cette situation est connue sous le nom de **méthode des comparaisons planifiées,** et elle est parfois utilisée dans les essais cliniques. Cependant, il n'est pas possible de poursuivre cette réflexion dans ce livre d'introduction, mais au moins, vous savez que cette méthode existe !

ANOVA Factorielle 3 : plans non équilibrés

ANOVA factorielle est une chose très pratique à connaître. C'est l'un des outils standard utilisés pour analyser les données expérimentales depuis de nombreuses décennies, et vous constaterez que vous ne pouvez pas lire plus de deux ou trois articles en psychologie sans y trouver une analyse de variance. Cependant, il y a une énorme différence entre les analyses de variance que vous verrez dans beaucoup d'articles scientifiques réels et les analyses de variance que j'ai décrites jusqu'ici. Dans la vie réelle, nous avons rarement la chance d'avoir des plans parfaitement équilibrés. Pour une raison ou une autre, il est typique de se retrouver avec plus d'observations dans certaines cellules que dans d'autres. Ou, pour le dire autrement, nous avons un **plan non équilibré**.

Les plans non équilibrés doivent être traités avec beaucoup plus de soin que les plans équilibrés, et la théorie statistique qui les sous-tend est beaucoup plus confuse. C'est peut-être une conséquence de ce désordre ou un manque de temps, mais d'après mon expérience, les cours de premier cycle sur les méthodes de recherche en psychologie ont tendance à ignorer complètement cette question. Beaucoup de manuels de statistiques ont aussi tendance à l'ignorer. Le résultat de tout cela, je pense, est que beaucoup de chercheurs actifs dans le domaine ne savent pas vraiment qu'il existe plusieurs « types » d'analyses de variance de pour les plans non équilibrés, et ils produisent des réponses très différentes. En fait, en lisant la littérature psychologique, je suis un peu étonné du fait que la plupart des gens qui rapportent les résultats d'une analyse de variance factorielle non équilibrée ne vous donnent pas assez de détails pour reproduire cette analyse. Je soupçonne secrètement que la plupart des gens ne se rendent même pas compte que leur progiciel statistique prend un grand nombre de décisions importantes d'analyse de données en leur nom. C'est un peu terrifiant quand on y pense. Donc, si vous voulez éviter de confier le contrôle de l'analyse de vos données à un logiciel idiot, lisez ce qui suit.

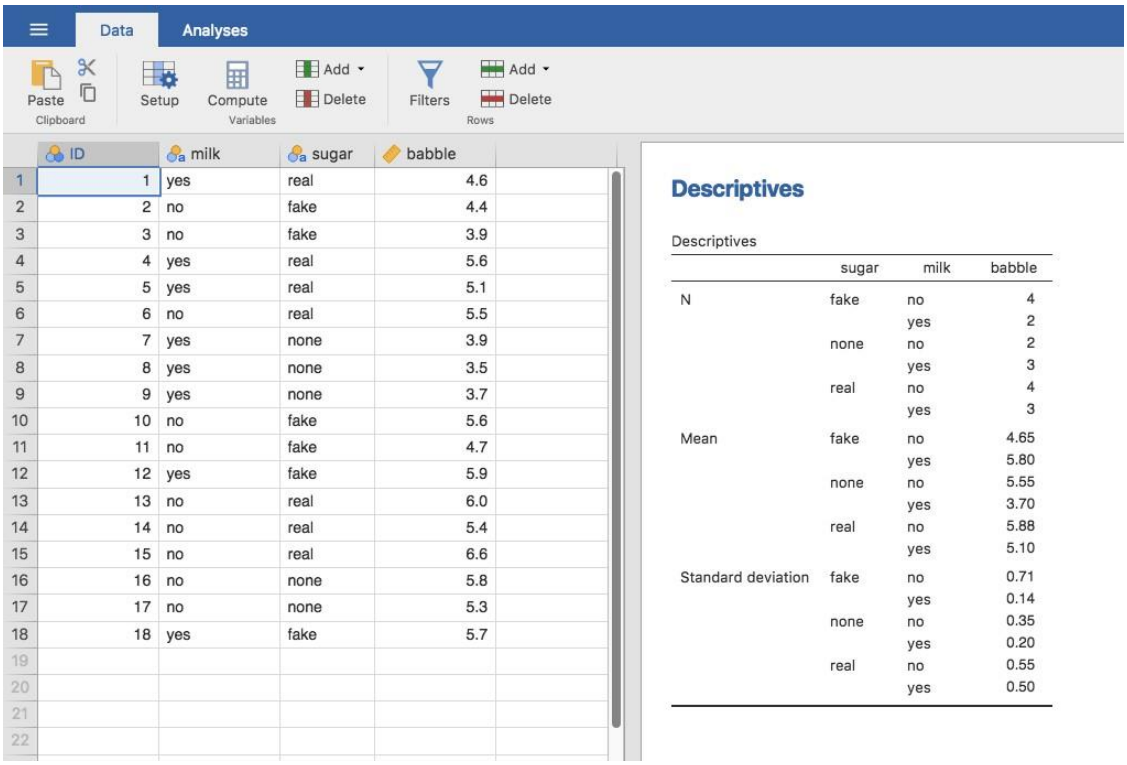
Les données sur le café

Comme d'habitude, nous travaillerons avec des données pour nous aider. Le fichier coffee.csv contient un ensemble de données hypothétiques qui produit une ANOVA 3x2 non

équilibrée. Supposons que nous voulions savoir si la tendance des gens à bavarder lorsqu'ils prennent trop de café est purement un effet du café lui-même, ou s'il y a un effet du lait et du sucre que les gens ajoutent au café. Supposons que nous prenions 18 personnes et leur donnions du café à boire. La quantité de café / caféine a été maintenue constante, et nous avons fait varié si le lait a été ajouté ou non, donc le lait (milk) est un facteur binaire avec deux niveaux, « oui » et « non ». Nous avons également varié le type de sucre en cause. Le café peut contenir du « vrai » sucre ou du « faux » sucre (c'est-à-dire un édulcorant artificiel), ou il peut en contenir « aucun », de sorte que la variable sucre (sugar) est un facteur à trois niveaux. Notre variable de résultat est une variable continue qui fait vraisemblablement référence à une mesure psychologiquement raisonnable du babillage (babble) de quelqu'un. Les détails n'ont pas vraiment d'importance pour notre but. Jetez un coup d'œil aux données du côté tableur de Jamovi, comme dans la Figure 14-26.

En examinant le tableau des moyennes de la Figure 14-26, nous avons l'impression qu'il y a des différences entre les groupes. C'est particulièrement vrai lorsque nous comparons ces moyennes aux écarts-types de la variable babble. Dans l'ensemble des groupes, cet écart-type varie de 0,14 à 0,71, ce qui est assez faible par rapport aux différences dans les moyennes des groupes.[130] Bien que cela puisse sembler à première vue une analyse de variance factorielle simple, un problème se pose lorsque nous examinons le nombre d'observations que nous avons dans chaque groupe. Voir les différents N pour les différents groupes illustrés à la Figure 14-26 Cela va à l'encontre de l'une de nos hypothèses initiales, à savoir que le nombre de personnes dans chaque groupe est le même. Nous n'avons pas vraiment discuté de la façon de gérer cette situation.

[130] Cet écart dans les écarts-types pourrait (et devrait) vous amener à vous demander si nous avons une violation de l'hypothèse d'homogénéité de la variance. Je vais laisser au lecteur le soin de vérifier cela à l'aide de l'option de test de Levene.



	ID	milk	sugar	babble
1	1	yes	real	4.6
2	2	no	fake	4.4
3	3	no	fake	3.9
4	4	yes	real	5.6
5	5	yes	real	5.1
6	6	no	real	5.5
7	7	yes	none	3.9
8	8	yes	none	3.5
9	9	yes	none	3.7
10	10	no	fake	5.6
11	11	no	fake	4.7
12	12	yes	fake	5.9
13	13	no	real	6.0
14	14	no	real	5.4
15	15	no	real	6.6
16	16	no	none	5.8
17	17	no	none	5.3
18	18	yes	fake	5.7
19				
20				
21				
22				

Descriptives

	sugar	milk	babble
N	fake	no	4
		yes	2
	none	no	2
		yes	3
	real	no	4
		yes	3
Mean	fake	no	4.65
		yes	5.80
	none	no	5.55
		yes	3.70
	real	no	5.88
		yes	5.10
Standard deviation	fake	no	0.71
		yes	0.14
	none	no	0.35
		yes	0.20
	real	no	0.55
		yes	0.50

Figure 14-26 : L'ensemble de données de coffee.csv dans Jamovi, avec des informations descriptives agrégées par niveaux de facteurs

Il n'existe pas d'ANOVA standard pour les conceptions non équilibrées.

Des plans non équilibrées nous amènent à la découverte quelque peu troublante qu'il n'y a pas vraiment une seule chose que nous pourrions appeler une ANOVA standard. En fait, il s'avère qu'il y a *trois*[131]façons fondamentalement différentes d'utiliser une ANOVA dans un plan non équilibré. Si vous avez un design équilibré, les trois versions donnent des résultats identiques, avec des sommes de carrés, de valeurs *F,* etc. conformes aux formules que j'ai

[131] En fait, c'est un mensonge. Les analyses de variance peuvent varier d'autres façons que celles dont j'ai parlé dans ce livre. Par exemple, j'ai complètement ignoré la différence entre les modèles à effets fixes dans lesquels les niveaux d'un facteur sont « fixés » par l'expérimentateur ou le monde, et les modèles à effets aléatoires dans lesquels les niveaux sont des échantillons aléatoires d'une population plus large de niveaux possibles (ce livre ne couvre que les modèles à effets fixes). Ne faites pas l'erreur de penser que ce livre, ou tout autre, vous dira « tout ce que vous devez savoir » sur les statistiques, pas plus qu'un seul livre ne pourrait vous dire tout ce que vous devez savoir en psychologie, en physique ou en philosophie. La vie est trop compliquée pour que cela soit vrai. Mais que cela ne soit pas une cause de désespoir. La plupart des chercheurs s'en sortent avec une connaissance pratique basique d'ANOVA qui ne va pas plus loin que ce livre. Je veux juste que vous gardiez à l'esprit que ce livre n'est que le début d'une très longue histoire, pas l'histoire entière.

données au début du chapitre. Cependant, lorsque votre conception est déséquilibrée, ils ne donnent pas les mêmes réponses. En outre, elles ne sont pas toutes adaptées à chaque situation. Certaines méthodes seront plus appropriées à votre situation que d'autres. Compte tenu de tout cela, il est important de comprendre quels sont les différents types d'ANOVA et en quoi ils diffèrent les uns des autres.

Le premier type d'ANOVA est conventionnellement appelé **somme des carrés de type I**. Je suis sûr que vous pouvez deviner comment s'appellent les deux autres. La partie « somme des carrés » du nom a été introduite par le progiciel statistique SAS et est devenue une nomenclature standard, mais c'est un peu trompeur à certains égards. Je pense que la logique pour les désigner comme différents types de somme de carrés est que, lorsqu'on examine les tableaux ANOVA qu'ils produisent, la principale différence dans les chiffres est la valeur SS. Les degrés de liberté ne changent pas, les valeurs MS sont toujours définies comme SS divisées par df, etc. Cependant, la terminologie se trompe parce qu'elle cache la raison *pour laquelle* les valeurs SS sont différentes les unes des autres. À cette fin, il est beaucoup plus utile de considérer les trois différents types d'analyses de variance comme trois *stratégies* différentes *de vérification des hypothèses*. Ces différentes stratégies mènent à des valeurs SS différentes, bien sûr, mais c'est la stratégie qui est importante ici, pas les valeurs SS elles-mêmes. Rappelons à la section 14.6 qu'il est préférable de considérer tout *test F* particulier comme une comparaison entre deux modèles linéaires. Ainsi, lorsque vous regardez un tableau ANOVA, il est utile de se rappeler que chacun de ces *tests F* correspond à une *paire de* modèles qui sont comparés. Cela nous amène naturellement à la question de savoir *quelle* paire de modèles est comparée. C'est la différence fondamentale entre les types I, II et III d'ANOVA : chacun correspond à une manière différente de choisir les paires de modèles pour les tests.

Somme des carrés de type I

La méthode de type I est parfois appelée somme « séquentielle » des carrés, car elle consiste à ajouter au modèle un des termes à la fois. Prenons les données sur le café, par exemple. Supposons que nous voulions exécuter la 3 x 2 ANOVA factorielle complète, y compris les termes d'interaction. Le modèle complet contient la variable de résultat babble, les variables prédicteurs sugar et milk, et l'interaction sugar*milk. Ceci peut s'écrire babble ~ sugar + milk + sugar* milk. La stratégie de type I construit ce modèle séquentiellement, à partir du modèle le plus simple possible et en ajoutant progressivement des termes.

Le modèle le plus simple possible pour les données serait un modèle dans lequel ni le lait ni le sucre n'auraient d'effet sur le babillage. Le seul terme qui serait inclus dans un tel modèle est l'intersection, écrit comme babble ~ 1, c'est notre hypothèse nulle initiale. Le modèle le plus simple suivant pour les données serait un modèle dans lequel un seul des deux effets principaux est inclus. Dans les données sur le café, il y a deux choix possibles, car nous pourrions choisir d'ajouter du lait en premier ou du sucre en premier. L'ordre s'avère important, comme nous le verrons plus tard, mais pour l'instant, faisons un choix arbitraire et choisissons le sucre. Ainsi, le deuxième modèle de notre séquence de modèles est babble ~ sugar, et il forme l'hypothèse alternative pour notre premier test. Nous avons maintenant notre premier test d'hypothèse :

- Modèle d'hypothèse nulle : babble ~ 1
- Modèle alternatif : babble ~ sugar

Cette comparaison forme notre test d'hypothèse sur l'effet principal du sugar. L'étape suivante dans notre exercice de construction de modèle est d'ajouter l'autre terme d'effet principal, donc le modèle suivant dans notre séquence est babble ~ sugar + milk. Le deuxième test d'hypothèse est ensuite formé en comparant les deux modèles suivants :

- Modèle d'hypothèse nulle : babble ~ sugar
- Modèle alternatif : babble ~ sugar + milk

Cette comparaison forme notre test d'hypothèse sur l'effet principal de millk. Dans un sens, cette approche est très élégante : l'hypothèse alternative du premier test forme l'hypothèse nulle pour le second. C'est dans ce sens que la méthode de type I est strictement séquentielle. Chaque test s'appuie directement sur les résultats du dernier. Cependant, dans un autre sens, c'est très peu élégant, parce qu'il y a une forte asymétrie entre les deux tests. Le test de l'effet principal de sugar (le premier test) ignore complètement le lait, alors que le test de l'effet principal du milk (le second test) prend en compte le sugar. En tout cas, le quatrième modèle de notre séquence est maintenant le modèle complet, babble ~ sugar + milk + sugar*milk, et le test d'hypothèse correspondant est :

- Modèle d'hypothèse nulle : babble ~ sugar + milk
- Modèle alternatif : babble ~ sugar + milk + sugar*milk

La somme des carrés de type III est la méthode de test d'hypothèse par défaut utilisée par Jamovi ANOVA, donc pour exécuter une analyse de somme de carrés de type I, nous devons sélectionner « Type 1 » dans la boîte de sélection « Somme des carrés » dans les options Jamovi « ANOVA » – « Model ». Cela nous donne le tableau ANOVA illustré à la figure 14.27.

ANOVA

	Sum of Squares	df	Mean Square	F	p
sugar	3.56	2	1.78	6.75	0.01086
milk	0.96	1	0.96	3.63	0.08106
sugar ✻ milk	5.94	2	2.97	11.28	0.00175
Residuals	3.16	12	0.26		

Figure 14-27 : Tableau des résultats ANOVA utilisant la somme des carrés de type I dans Jamovi

Le gros problème avec l'utilisation de la somme des carrés de type I est le fait que cela dépend vraiment de l'ordre dans lequel vous entrez les variables. Pourtant, dans de nombreuses situations, le chercheur n'a aucune raison de préférer un ordre à un autre. C'est probablement le cas pour notre problème du lait et du sucre. Doit-on ajouter du lait d'abord ou du sucre d'abord ? C'est exactement aussi arbitraire qu'une question d'analyse de

données sur la préparation du café. En fait, il y a peut-être des gens qui ont des opinions fermes sur l'ordre, mais il est difficile d'imaginer une réponse fondée sur des principes à cette question. Pourtant, regardez ce qui se passe lorsque nous changeons l'ordre, comme dans la Figure 14-28.

ANOVA

	Sum of Squares	df	Mean Square	F	p
milk	1.44	1	1.44	5.48	0.03733
sugar	3.07	2	1.53	5.82	0.01708
milk ✻ sugar	5.94	2	2.97	11.28	0.00175
Residuals	3.16	12	0.26		

Figure 14-28 : Tableau des résultats de l'analyse de variance utilisant la somme des carrés de type I dans le Jamovi, mais avec les facteurs saisis dans un ordre différent (milk d'abord).

Les valeurs *p* des deux principaux effets ont changé, et de façon assez spectaculaire. Entre autres choses, l'effet de milk est devenu significatif (bien qu'il faille éviter d'en tirer de fortes conclusions, comme je l'ai mentionné précédemment). Laquelle de ces deux analyses de variance doit-on déclarer ? Ce n'est pas immédiatement évident.

Lorsqu'on regarde les tests d'hypothèse utilisés pour définir le « premier » effet principal et le « second », il est clair qu'ils sont qualitativement différents l'un de l'autre. Dans notre premier exemple, nous avons vu que le test de l'effet principal de sugar ignore complètement milk, alors que le test de l'effet principal de milk prend en compte sugar. En tant que telle, la stratégie d'essai du type I traite vraiment le premier effet principal comme s'il avait une sorte de primauté théorique sur le second. D'après mon expérience, il y a très rarement, voire jamais, une primauté théorique de ce genre qui justifierait de traiter deux effets principaux de façon asymétrique.

La conséquence de tout cela est que les tests du type I présentent très rarement un grand intérêt, et nous devrions donc passer aux essais du type II et du type III.

Type III somme des carrés

Après avoir terminé de parler des tests du type I, vous pourriez penser que la chose naturelle à faire ensuite serait de parler des tests du type II. Cependant, je pense qu'il est en fait un peu plus naturel de discuter des tests de type III (qui sont simples et qui sont par défaut dans l'ANOVA Jamovi) avant de parler des essais de type II (qui sont plus difficiles). L'idée de base des essais du type III est extrêmement simple. Quel que soit le terme que vous essayez d'évaluer, exécutez le test *F* dans lequel l'hypothèse alternative correspond au modèle ANOVA complet tel que spécifié par l'utilisateur, et le modèle d'hypothèse nulle ne fait que supprimer ce terme que vous testez. Dans l'exemple du café, où notre modèle complet était babble ~ sugar + milk + sugar* milk, le test pour un effet principal de sugar correspondrait à une comparaison entre les deux modèles suivants :

- Modèle d'hypothèse nulle : babble ~ milk + sugar* milk

- Modèle alternatif : babble ~ sugar + milk + sugar* milk

De même, l'effet principal du lait est évalué en testant le modèle complet par rapport à un modèle nul qui supprime le terme lait, comme ça :

- Modèle d'hypothèse nulle : babble ~ sugar + sugar* milk
- Modèle alternatif : babble ~ sugar + milk + sugar* milk

Enfin, l'interaction sucre*lait est évaluée exactement de la même manière. Une fois de plus, nous testons le modèle complet par rapport à un modèle nul qui supprime le terme d'interaction sucre*lait, comme ça :

- Modèle d'hypothèse nulle : babble ~ sugar + milk
- Modèle alternatif : babble ~ sugar + milk + sugar* milk

L'idée de base se généralise aux ANOVA d'ordre supérieur. Par exemple, supposons que nous essayions d'exécuter une analyse de variance avec trois facteurs, A, B et C, et que nous voulions considérer tous les effets principaux possibles et toutes les interactions possibles, y compris l'interaction triple A*B*C. Le tableau ci-dessous vous montre à quoi ressemblent les tests du type III dans cette situation :

Term being tested is	Null model is `outcome ~ ...`	Alternative model is `outcome ~ ...`
A	B + C + A*B + A*C + B*C + A*B*C	A + B + C + A*B + A*C + B*C + A*B*C
B	A + C + A*B + A*C + B*C + A*B*C	A + B + C + A*B + A*C + B*C + A*B*C
C	A + B + A*B + A*C + B*C + A*B*C	A + B + C + A*B + A*C + B*C + A*B*C
A*B	A + B + C + A*C + B*C + A*B*C	A + B + C + A*B + A*C + B*C + A*B*C
A*C	A + B + C + A*B + B*C + A*B*C	A + B + C + A*B + A*C + B*C + A*B*C
B*C	A + B + C + A*B + A*C + A*B*C	A + B + C + A*B + A*C + B*C + A*B*C
A*B*C	A + B + C + A*B + A*C + B*C	A + B + C + A*B + A*C + B*C + A*B*C

Aussi moche que ce tableau ait l'air, il est assez simple. Dans tous les cas, l'hypothèse alternative correspond au modèle complet qui contient trois termes d'effets principaux (p. ex. A), trois interactions (p. ex. A*B) et une interaction triple (p. ex. A*B*C). Le modèle de l'hypothèse nulle contient toujours 6 de ces 7 termes, et celui qui manque est celui dont nous essayons de tester la signification.

Au premier abord, les tests de type III semblent être une bonne idée. Premièrement, nous avons éliminé l'asymétrie qui nous causait des problèmes lors de l'exécution des tests de type I. Et parce que nous traitons maintenant tous les termes de la même façon, les résultats des tests d'hypothèse ne dépendent pas de l'ordre dans lequel nous les spécifions. C'est définitivement une bonne chose. Cependant, l'interprétation des résultats des tests pose un gros problème, surtout en ce qui concerne les effets principaux. Pensez aux données sur le café. Supposons qu'il s'avère que l'effet principal de milk ne soit pas significatif selon les tests du type III. Ce que cela nous dit, c'est que babble ~ sugar + sugar*milk est un meilleur modèle pour les données que le modèle complet. Mais qu'est-ce que ça *veut dire* ? Si le terme d'interaction sugar*milk était également non significatif, nous serions tentés de conclure que les données nous disent que la seule chose qui compte est le sucre. Mais

supposons que nous ayons un terme d'interaction significatif, mais un effet principal non significatif du lait. Dans ce cas, faut-il supposer qu'il y a vraiment un « effet du sucre », une « interaction entre le lait et le sucre », mais pas un « effet du lait » ? Cela semble fou. La bonne réponse *doit* simplement être qu'il est inutile de parler de l'effet principal[132] si l'interaction est importante. En général, c'est ce que la plupart des statisticiens nous conseillent de faire, et je pense que c'est le bon conseil. Mais s'il est vraiment inutile de parler d'effets principaux non significatifs en présence d'une interaction significative, il n'est pas du tout évident de savoir pourquoi les essais de type III devraient permettre à l'hypothèse nulle de s'appuyer sur un modèle qui inclut l'interaction mais omet un des principaux effets qui la composent. Lorsqu'elles sont ainsi caractérisées, les hypothèses nulles n'ont vraiment aucun sens.

Nous verrons plus loin que les tests de type III peuvent être utiles dans certains contextes, mais examinons d'abord le tableau des résultats d'ANOVA en utilisant la somme des carrés de type III, voir Figure 14-29.

Mais attention, l'une des caractéristiques perverses de la stratégie d'essais du type III est que les résultats dépendent généralement des *contrastes* que vous utilisez pour coder vos facteurs (voir la section 14.7 si vous avez oublié quels sont les différents types de contrastes).[133]

D'accord, donc si les *valeurs p* qui ressortent généralement des analyses de type III (mais pas dans Jamovi) sont si sensibles au choix des contrastes, est-ce que cela signifie que les essais de type III sont essentiellement arbitraires et ne sont pas fiables ? Dans une certaine mesure, c'est vrai, et lorsque nous parlons des tests de type II, nous verrons que les analyses de type II évitent complètement ce caractère arbitraire, mais je pense que c'est une conclusion trop forte.

ANOVA

	Sum of Squares	df	Mean Square	F	p
milk	1.00	1	1.00	3.81	0.07467
sugar	2.13	2	1.07	4.04	0.04543
milk ✻ sugar	5.94	2	2.97	11.28	0.00175
Residuals	3.16	12	0.26		

Figure 14-29 : Tableau des résultats ANOVA utilisant la somme des carrés de type III in Jamovi

[132] Ou, à tout le moins, rarement d'intérêt.

[133] Cependant, dans Jamovi les résultats pour la somme des carrés de l'ANOVA Type III sont les mêmes quel que soit le contraste choisi, donc Jamovi fait évidemment quelque chose de différent !

Tout d'abord, il est important de reconnaître que certains choix de contrastes produiront toujours les mêmes réponses (ah, c'est donc ce qui se passe dans Jamovi). Il est particulièrement important de noter que si les colonnes de notre matrice de contraste ont toutes une somme à zéro, l'analyse de type III donnera toujours les mêmes réponses.

Type II somme des carrés

Bien, nous avons vu des essais de type I et III maintenant, et les deux sont assez simples. Les essais de type I sont effectués en ajoutant progressivement un des termes à la fois, tandis que les essais de type III sont effectués en prenant le modèle complet et en regardant ce qui se passe lorsque vous retirez chaque terme. Cependant, les deux peuvent avoir certaines limites. Les essais de type I dépendent de l'ordre dans lequel vous entrez les termes, et les essais de type III dépendent de la façon dont vous codez vos contrastes. Les essais du type II sont un peu plus difficiles à décrire, mais ils évitent ces deux problèmes et, par conséquent, ils sont un peu plus faciles à interpréter.

Les essais du type II sont globalement similaires aux essais du type III. Commencer par un modèle « complet » et tester un terme particulier en le supprimant de ce modèle. Cependant, les essais de type II sont basés sur le **principe de marginalité** qui stipule que vous ne devez pas omettre un terme d'ordre inférieur de votre modèle s'il y a des termes d'ordre supérieur qui en dépendent. Ainsi, par exemple, si votre modèle contient l'interaction A*B (un terme de 2e ordre), alors il devrait contenir les effets principaux A et B (termes de 1er ordre). De même, s'il contient un terme d'interaction triple A*B*C, alors le modèle doit également inclure les principaux effets A, B et C ainsi que les interactions plus simples A*B, A*C et B*C. Les essais de type III violent systématiquement le principe de marginalité. Par exemple, examinons le test de l'effet principal de A dans le contexte d'une analyse de variance à trois facteurs qui comprend tous les termes d'interaction possibles. Selon les essais de type III, nos modèles nuls et alternatifs le sont :

- Modèle d'hypothèse nulle : outcome ~ B + C + A*B + A*C + B*C + A*B*C
- Modèle alternatif : outcome ~ A + B + C + C + A*B + A*C + B*C + A*B*C

Notez que l'hypothèse nulle omet A, mais inclut A*B, A*C et A*B*C dans le modèle. D'après les tests du type II, ce n'est pas un bon choix d'hypothèse nulle. Ce que nous devrions plutôt faire, si nous voulons vérifier l'hypothèse nulle selon laquelle A n'est pas pertinent pour notre variable résultat, est de spécifier l'hypothèse nulle qui est le modèle le plus compliqué qui ne repose d'aucune façon sur A, même dans une interaction. L'hypothèse alternative correspond à ce modèle d'hypothèse nulle plus un terme d'effet principal de A. C'est beaucoup plus proche de ce que la plupart des gens penseraient intuitivement d'un « effet principal de A», et elle donne ce qui suit comme notre test de Type II de l'effet principal de A :[134]

[134] Notez, bien sûr, que cela dépend du modèle que l'utilisateur a spécifié. Si le modèle ANOVA original ne contient pas de terme d'interaction pour B*C, il est évident qu'il n'apparaîtra ni dans la valeur nulle ni dans l'alternative. Mais c'est vrai pour les types I, II et

- Modèle nul : outcome ~ B + C + B*C
- Modèle alternatif : outcome ~ A + B + C + C + B*C

Quoi qu'il en soit, pour vous donner une idée du déroulement des essais de type II, voici le tableau complet de modèles qui serait appliqué dans une ANOVA factorielle à trois facteurs :

Term being tested is	Null model is `outcome ~ ...`	Alternative model is `outcome ~ ...`
A	B + C + B*C	A + B + C + B*C
B	A + C + A*C	A + B + C + A*C
C	A + B + A*B	A + B + C + A*B
A*B	A + B + C + A*C + B*C	A + B + C + A*B + A*C + B*C
A*C	A + B + C + A*B + B*C	A + B + C + A*B + A*C + B*C
B*C	A + B + C + A*B + A*C	A + B + C + A*B + A*C + B*C
A*B*C	A + B + C + A*B + A*C + B*C	A + B + C + A*B + A*C + B*C + A*B*C

Dans le contexte de l'analyse de variance à deux facteurs que nous avons utilisée dans les données sur le café, les tests d'hypothèse sont encore plus simples. L'effet principal de sugar correspond à un *test F* comparant ces deux modèles :

- Modèle d'hypothèse nulle : babble ~ milk
- Modèle alternatif : babble ~ sugar + milk

Le test de l'effet principal de milk est le suivant

- Modèle d'hypothèse nulle : babble ~ sugar
- Modèle alternatif : babble ~ sugar + milk

Enfin, le test pour l'interaction sugar*milk est :

- Modèle d'hypothèse nulle : babble ~ sugar + milk
- Modèle alternatif : babble ~ sugar + milk + sugar*mil

L'exécution des tests est à nouveau simple. Il suffit de sélectionner « Type 2 » dans la boîte de sélection « Somme des carrés »' dans les options « ANOVA » - « Modèle » de Jamovi, ce qui nous donne le tableau ANOVA montré dans la Figure 14.30.

III. Ils n'incluent jamais de termes que vous *n'avez pas* inclus, mais ils font des choix différents sur la façon de construire des tests pour ceux que vous avez inclus.

ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	η^2p	ω^2
sugar	3.07	2	1.53	5.82	0.01708	0.23	0.49	0.19
milk	0.96	1	0.96	3.63	0.08106	0.07	0.23	0.05
sugar ✻ milk	5.94	2	2.97	11.28	0.00175	0.45	0.65	0.40
Residuals	3.16	12	0.26					

Figure 14-30 : Tableau des résultats ANOVA utilisant la somme des carrés de type II dans Jamovi

Les essais du type II présentent certains avantages évidents par rapport aux essais du type I et du type III. Ils ne dépendent pas de l'ordre dans lequel vous spécifiez les facteurs (contrairement au Type I), et ils ne dépendent pas des contrastes que vous utilisez pour spécifier vos facteurs (contrairement au Type III). Et bien que les opinions puissent diverger sur ce dernier point, et cela dépendra certainement de ce que vous essayez de faire avec vos données, je pense que les tests d'hypothèse qu'ils spécifient sont plus susceptibles de correspondre à quelque chose qui vous préoccupe vraiment. Par conséquent, je trouve qu'il est généralement plus facile d'interpréter les résultats d'un essai du type II que ceux d'un essai du type I ou III. Pour cette raison, mon conseil provisoire est que, si vous ne pouvez pas penser à des comparaisons de modèles évidentes qui correspondent directement à vos questions de recherche, mais que vous voulez quand même exécuter une analyse de variance dans un plan non équilibré, les tests de type II sont probablement un meilleur choix que de type I ou III.[135]

[135] Je trouve amusant de constater que la valeur par défaut de R est Type I et que la valeur par défaut de SPSS et Jamovi est Type III. Ni l'un ni l'autre ne m'attire tant que ça. Par ailleurs, je trouve déprimant de constater que presque personne dans la littérature psychologique ne se donne la peine de signaler le type de tests qu'ils ont effectués, encore moins l'ordre des variables (pour le type I) ou les contrastes utilisés (pour le type III). Souvent, ils ne signalent pas non plus les logiciels qu'ils ont utilisés. La seule façon de comprendre ce que les gens rapportent habituellement est d'essayer de deviner à partir d'indices annexes quel logiciel ils utilisaient, et de supposer qu'ils n'ont jamais modifié les paramètres par défaut. S'il vous plaît, ne faites pas ça ! Maintenant que vous connaissez ces problèmes, veillez à indiquer le logiciel que vous avez utilisé, et si vous déclarez des résultats ANOVA pour des données non équilibrées, puis précisez le type de tests que vous avez effectués, précisez les informations sur l'ordre si vous avez effectué des tests de type I et précisez les contrastes si vous avez effectué des tests de type III. Ou, mieux encore, faites des tests d'hypothèses qui correspondent à des choses qui vous tiennent vraiment à cœur, puis rapportez-les !

Tailles de l'effet (et sommes non additives de carrés)

Jamovi fournit également les tailles d'effet η^2 et η^2 partiel lorsque vous sélectionnez ces options, comme dans la Figure 14-30. Cependant, quand vous avez un plan non équilibré, c'est un peu plus de complexité.

Si vous vous souvenez de nos premières discussions sur l'analyse de variance, l'une des idées clés derrière les sommes des calculs des carrés est que si nous additionnons tous les termes SS associés aux effets dans le modèle, et que nous ajoutons cela aux SS résiduels, ils sont censés s'additionner pour former la somme totale des carrés. Et, en plus de cela, l'idée derrière η^2 est que, parce que vous divisez l'une des SS par la SS totale, une valeur η^2 peut être interprétée comme la proportion de la variance représentée par un terme particulier. Mais ce n'est pas aussi simple dans les plans non équilibrés parce qu'une partie de la variance est « manquante ».

Cela semble un peu étrange à première vue, mais voici pourquoi. Lorsque vous avez des plans non équilibrés, vos facteurs sont corrélés les uns avec les autres, et il devient difficile de faire la différence entre l'effet du facteur A et l'effet du facteur B. Dans le cas extrême, supposons que nous ayons exécuté un plan 2x2 où le nombre de participants dans chaque groupe était le suivant :

	sugar	no sugar
milk	100	0
no milk	0	100

Nous avons ici un plan spectaculairement déséquilibré : 100 personnes prennent du lait et du sucre, 100 personnes ne prennent ni lait ni sucre, et c'est tout. Il y a 0 personne avec du lait et sans sucre, et 0 personne avec du sucre mais sans lait. Supposons maintenant que, lorsque nous avons recueilli les données, il s'est avéré qu'il existe une différence importante (et statistiquement significative) entre le groupe « lait et sucre » et le groupe « sans lait et sans sucre ». S'agit-il d'un effet principal du sucre ? Un effet principal du lait ? Ou une interaction ? C'est impossible à dire, car la présence de sucre est parfaitement associée à la présence de lait. Supposons maintenant que le design ait été un peu plus équilibré :

	sugar	no sugar
milk	100	5
no milk	5	100

Cette fois-ci, il est techniquement possible de faire la distinction entre l'effet du lait et l'effet du sucre, parce que nous avons quelques personnes qui ont l'un mais pas l'autre. Cependant, il sera encore assez difficile de le faire, car l'association entre le sucre et le lait est encore extrêmement forte, et il y a si peu d'observations dans deux des groupes. Encore une fois, nous sommes très susceptibles d'être dans une situation où nous *savons que les* variables prédicteurs (sugar et milk) sont liées au résultat (babble), mais nous ne savons pas si la *nature* de cette relation est un effet principal de l'un ou l'autre prédicteur, ou l'interaction.

Cette incertitude est à l'origine de la variance manquante. La variance « manquante » correspond à la variation de la variable des résultats qui est clairement attribuable aux prédicteurs, mais nous ne savons pas lequel des effets du modèle est responsable. Lorsque vous calculez la somme des carrés de type I, aucune variance ne disparaît jamais. La nature séquentielle de la somme des carrés de type I signifie que l'analyse de variance attribue automatiquement cette variance aux effets qui sont entrés en premier. Toutefois, les essais de type II et de type III sont plus conservateurs. La variance qui ne peut pas être clairement attribuée à un effet spécifique n'est attribuée à aucun d'entre eux, et elle disparaît.

Résumé

- ANOVA factorielle avec des plans équilibrés, sans interactions (section 14.1) et avec interactions incluses (section 14.2)
- Taille de l'effet, moyennes estimées et intervalles de confiance dans une analyse de variance factorielle (section 14.3)
- Vérification des hypothèses dans l'analyse de variance (Section 14.4)
- Analyse de la covariance (ANCOVA) (Section 14.5)
- Comprendre le modèle linéaire sous-jacent à l'analyse de variance, y compris les différents contrastes (sections 14.6 et 14.7)
- Tests post hoc utilisant le HSD de Tukey (section 14.8) et un bref commentaire sur les comparaisons prévues (section 14.9)
- ANOVA d'usine avec des conceptions non équilibrées (Section 14.10)

Analyse factorielle

Les chapitres précédents ont couvert les tests statistiques visant à déterminer les différences entre deux groupes ou plus. Cependant, dans le cadre d'une recherche, il nous arrive parfois de vouloir examiner la covariabilité de variables multiples. C'est-à-dire, la façon dont elles sont reliées les unes aux autres et de voir si les schémas de relation nous suggèrent quelque chose d'intéressant et ayant dus sens. Par exemple, nous sommes souvent intéressés à déterminer l'existence de **facteurs latents** sous-jacents à partir des variables observées et mesurées directement dans notre ensemble de données. En statistique, les facteurs latents sont des variables cachées qui ne sont pas directement observées, mais plutôt déduites (par analyse statistique) d'autres variables qui sont observées (mesurées directement).

Dans ce chapitre, nous examinerons un certain nombre d'analyses factorielles et de techniques connexes, en commençant par l'analyse factorielle exploratoire ou AFE (en anglais : Exploratory factorial analysis ou EFA). L'AFE est une technique statistique permettant d'identifier les facteurs latents sous-jacents dans un ensemble de données (section 15.1). Ensuite, dans la section 15.2, nous aborderons l'analyse en composantes principales (ACP) qui est une technique de réduction des données qui, à proprement parler, ne permet pas d'identifier les facteurs latents sous-jacents. Au lieu de cela, l'ACP produit simplement une combinaison linéaire de variables observées. Ensuite, la section 15.3 sur l'analyse factorielle confirmatoire (AFC) montre que, contrairement à l'AFE, avec l'AFC, vous commencez par une idée - un modèle - de la façon dont les variables de vos données sont reliées les unes aux autres. Vous testez ensuite votre modèle par rapport aux données observées et évaluez dans quelle mesure il s'adapte au modèle. Une version plus sophistiquée de l'AFC est l'approche dite Multi-Trait Multi-Method (MTMM) (section 15.4), dans laquelle le modèle tient compte à la fois de la variance latente du facteur et de la méthode. Ceci est utile lorsqu'il existe différentes approches méthodologiques utilisées pour la collecte des données et que la variabilité de la méthodologie est donc une question importante. Enfin, nous aborderons une analyse connexe : l'analyse de fiabilité de la cohérence interne teste la cohérence d'une échelle pour la mesure une construction psychologique (section 15.5).

Analyse factorielle exploratoire

L'analyse factorielle exploratoire (AFE) est une technique statistique permettant de révéler tous les facteurs latents cachés qui peuvent être déduits de nos données observées. Cette technique calcule dans quelle mesure un ensemble de variables mesurées, par exemple V1, V2, V3, V4 et V5, peut être représenté comme mesures d'un facteur latent sous-jacent. Ce facteur latent ne peut pas être mesuré au moyen d'une seule variable observée, mais il se manifeste plutôt par les relations qu'il provoque dans un ensemble de variables observées.

Dans la Figure 15-1, chaque variable observée V est « causée « dans une certaine mesure par le facteur latent sous-jacent (F), représenté par les coefficients b1 à b5 (aussi appelés saturations de facteurs). Chaque variable observée a également un terme d'erreur associé,

e1 à e5. Chaque terme d'erreur est la variance de la variable observée associée, *Vi*, qui est inexpliquée par le facteur latent sous-jacent.

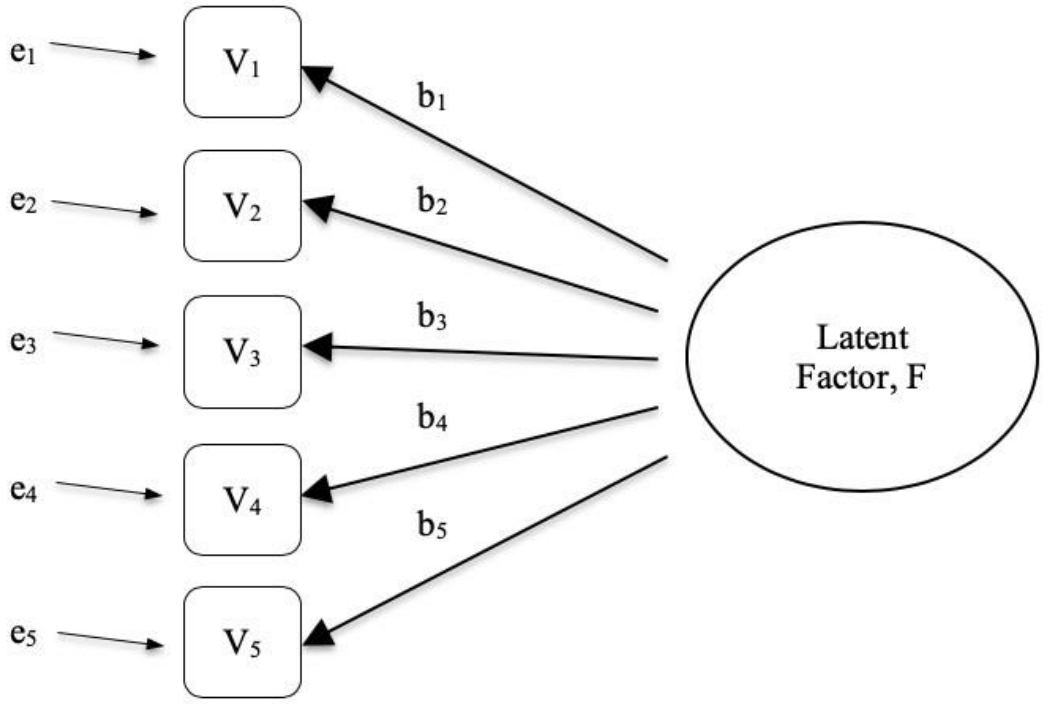


Figure 15-1 : Facteur latent sous-jacent à la relation entre plusieurs variables observées

En psychologie, les facteurs latents représentent des phénomènes ou des constructions psychologiques difficiles à observer ou à mesurer directement. Par exemple, la personnalité, l'intelligence ou le style de pensée. Dans l'exemple de la Figure 15-1, nous avons peut-être posé cinq questions précises sur le comportement ou les attitudes des gens, ce qui nous permet de nous faire une idée d'un concept de personnalité appelé, par exemple, extraversion. Un ensemble différent de questions spécifiques peut nous donner une image de l'introversion d'un individu, ou de sa conscience. Et ces différents facteurs latents peuvent être corrélés entre eux.

Voici un autre exemple : nous pouvons ne pas être en mesure de mesurer directement l'anxiété liée aux statistiques, mais nous pouvons mesurer si l'anxiété liée aux statistiques est élevée ou faible avec un ensemble d'items dans un questionnaire. Par exemple, « Q1 : Faire le devoir d'un cours de statistique », « Q2 : Essayer de comprendre les statistiques décrites dans un article de journal », et « Q3 : Demander à l'enseignant de l'aider à comprendre quelque chose du cours », etc. Les personnes ayant une grande anxiété statistique auront aussi tendance à donner des réponses élevées à ces variables en raison de leur anxiété statistique importante. De même, les personnes dont l'anxiété statistique est faible donneront de la même façon des réponses faibles à ces variables en raison de leur faible anxiété statistique.

Dans l'analyse factorielle exploratoire (AEF), nous explorons essentiellement les corrélations entre les variables observées pour découvrir tout facteur sous-jacent (latent) intéressant et important qui est identifié lorsque les variables observées covarient. Nous pouvons utiliser un logiciel statistique pour estimer tout facteur latent et pour identifier les variables qui ont une saturation importante[136] (p. ex., saturation > 0,5) sur chaque facteur, ce qui suggère qu'elles sont une mesure ou un indicateur utile du facteur latent. Une partie de ce processus comprend une étape appelée rotation, ce qui pour être honnête est une idée assez bizarre, mais heureusement nous n'avons pas à nous soucier de la comprendre ; nous avons juste besoin de savoir qu'elle est utile parce qu'elle rend beaucoup plus clair le modèle des saturations sur différents facteurs. Ainsi, la rotation aide à mieux voir quelles variables sont liées de façon substantielle à chaque facteur. Nous devons également décider combien de facteurs sont raisonnables compte tenu de nos données, et ce qui est utile à cet égard, ce sont les valeurs propres qui indiquent l'ampleur de chaque facteur.

Vérification des hypothèses

Il y a aussi quelques hypothèses qui doivent être vérifiées dans le cadre de cette analyse. La première hypothèse est la **sphéricité**, qui vérifie essentiellement que les variables de votre ensemble de données sont suffisamment corrélées entre elles pour être résumées avec un ensemble plus restreint de facteurs. Le test de sphéricité de Bartlett vérifie si la matrice de corrélation observée s'écarte significativement d'une matrice à corrélation nulle. Ainsi, si le test de Bartlett est significatif (p<.05), cela indique que la matrice de corrélation observée est significativement divergente de la matrice nulle, et donc convient à l'AFE.

La deuxième hypothèse est l'**adéquation de l'échantillonnage** et est vérifiée à l'aide de la mesure de l'adéquation de l'échantillonnage (Measure of sampling adequation ou MSA) de Kaiser-Meyer-Olkin (KMO). L'indice KMO est une mesure de la proportion de variance parmi les variables observées qui pourrait être une variance commune. Il vérifie les corrélations partielles, c'est-à-dire lorsqu'il y a des facteurs qui ne saturent que deux éléments. Nous voulons rarement, sinon jamais, que l'AFE produise beaucoup de facteurs en ne saturant que deux éléments chacun. Le KMO concerne l'adéquation de l'échantillonnage car des corrélations partielles sont généralement observées avec des échantillons inadéquats. Si l'indice KMO est élevé (≈ 1), le CFC est efficace alors que si le KMO est faible (≈ 0), l'AFE n'est pas pertinente. Des valeurs KMO inférieures à 0,5 indiquent que l'AFE n'est pas appropriée et une valeur KMO de 0,6 devrait être confirmée avant que l'AFE soit considérée comme appropriée. Les valeurs entre 0,5 et 0,7 sont considérées comme adéquates, les valeurs entre 0,7 et 0,9 sont bonnes et les valeurs entre 0,9 et 1,0 sont excellentes.

A quoi sert l'AFE ?

Si l'AFE a fourni une bonne solution (c.-à-d. un modèle factoriel), nous devons alors décider quoi faire de nos nouveaux facteurs. Les chercheurs utilisent souvent l'AFE au cours de

---

[136] Les saturations factorielles peuvent être interprétées comme des coefficients de régression normalisés, ce qui est très utile

l'élaboration de l'échelle psychométrique. Ils développeront l'ensemble des items d'un questionnaire qui, selon eux, se rapportent à une ou plusieurs constructions psychologiques, utiliseront l'AFE pour voir quels items « vont ensemble » comme facteurs latents, puis ils évalueront si certains items devraient être supprimés parce qu'ils ne mesurent pas de façon utile ou distincte un des facteurs latents.

Conformément à cette approche, une autre conséquence de l'AFE est de combiner les variables qui saturent des facteurs distincts à travers un score factoriel, parfois appelé score gradué. Il y a deux options pour combiner des variables dans une échelle de notation :

- Créez une nouvelle variable avec une note pondérée par les coefficients de pondération pour chaque élément qui sature un facteur.
- Créez une nouvelle variable en fonction de chaque élément qui sature un facteur, mais en les pondérant également.

Dans la première option, la saturation de chaque élément au score combiné dépend de l'importance de son lien avec le facteur. Dans la deuxième option, nous faisons généralement la moyenne de tous les éléments qui saturent de façon substantielle un facteur pour créer la variable de l'échelle de notation combinée. Le choix est une question de préférence, bien que la première présente l'inconvénient que les saturations peuvent varier considérablement d'un échantillon à l'autre, surtout dans les sciences du comportement et de la santé, où nous sommes souvent intéressés à développer et à utiliser des échelles construites sur des questionnaires composites pour différentes études et différents échantillons. Dans ce cas, il est raisonnable d'utiliser une mesure composite fondée sur la saturation égale des éléments de fond plutôt que sur la pondération par les saturations spécifiques d'un échantillon provenant d'un échantillon différent. Dans tous les cas, il est plus simple et plus intuitif de comprendre une mesure de variable combinée comme une moyenne d'éléments que d'utiliser une combinaison pondérée de façon optimale spécifique à un échantillon.

Une technique statistique plus avancée, qui dépasse la portée du présent ouvrage, consiste à entreprendre une modélisation de régression où les facteurs latents sont utilisés dans les modèles de prédiction d'autres facteurs latents. C'est ce qu'on appelle la « modélisation d'équation structurelle » et il existe des logiciels et des progiciels R spécifiques dédiés à cette approche. Mais n'allons pas trop vite ; ce sur quoi nous devrions vraiment nous concentrer maintenant, c'est sur la manière de faire une AFE avec Jamovi.

Faire une AFE avec Jamovi

D'abord, il nous faut des données. Vingt-cinq éléments d'auto-évaluation de la personnalité (voir Figure 15-2) tirés de l'International Personality Item Pool () ont été inclus dans le cadre du projet SAPA (Synthetic Aperture Personality Assessment) sur le Web (SAPA : .) Les 25 items sont organisés selon cinq facteurs présumés : Agréabilité, Conscience, Extraversion, Névrose et Ouverture.

Variable name	Question / Item (short phrases that you should respond to by indicating how accurately the statement describes your typical behaviour or attitudes)	Coding (R: reverse)
A1	Am indifferent to the feelings of others.	R
A2	Inquire about others' well-being.	
A3	Know how to comfort others.	
A4	Love children.	
A5	Make people feel at ease.	
C1	Am exacting in my work.	
C2	Continue until everything is perfect.	
C3	Do things according to a plan.	
C4	Do things in a half-way manner.	R
C5	Waste my time.	R
E1	Don't talk a lot.	R
E2	Find it difficult to approach others.	R
E3	Know how to captivate people.	
E4	Make friends easily.	
E5	Take charge.	
N1	Get angry easily.	
N2	Get irritated easily.	
N3	Have frequent mood swings.	
N4	Often feel blue.	
N5	Panic easily.	
O1	Am full of ideas.	
O2	Avoid difficult reading material.	R
O3	Carry the conversation to a higher level.	
O4	Spend time reflecting on things.	
O5	Will not probe deeply into a subject.	R

Figure 15-2 : Vingt-cinq items variables observés organisés selon cinq facteurs de personnalité présumés dans l'ensemble de donnéesbfi_sample.csv

Les données sur les items ont été recueillies à l'aide d'une échelle de réponse en 6 points :

1. Très imprécis
2. Modérément inexact
3. Légèrement inexact
4. Légèrement précis
5. Modérément précis
6. Très précis.

Un échantillon de N=250 réponses est contenu dans l'ensemble de données bfi_sample.csv. En tant que chercheurs, nous souhaitons explorer les données pour voir s'il existe des facteurs latents sous-jacents qui sont raisonnablement bien mesurés par les 25 variables

observées dans le fichier de données bfi_sample.csv. Ouvrez l'ensemble de données et vérifiez que les 25 variables sont codées en tant que variables continues (on peut soutenir qu'elles sont ordinales mais pour l'AFE dans Jamovi, cela n'a généralement pas d'importance, sauf si vous décidez de calculer des scores factoriels pondérés, auquel cas des variables continues sont nécessaires). Pour réaliser l'AFE avec Jamovi :

- Sélectionnez Factor - Exploratory Factor Analysis dans la barre principale de boutons de Jamovi pour ouvrir la fenêtre d'analyse de l'AFE (Figure 15-3).
- Sélectionnez les 25 questions de personnalité et transférez-les dans la boîte'Variables'.
- Cochez les options appropriées, y compris les options « Assumptions checks », mais aussi « 'Method » de rotation, « Number of Factors to extract » et « Additional Output ». Voir la Figure 15-3 pour les options suggérées pour cet exemple d'AFE, et notez que la méthode de rotation et le nombre de facteurs à extraire sont habituellement ajustés pendant l'analyse pour trouver le meilleur résultat, tel que décrit ci-dessous.

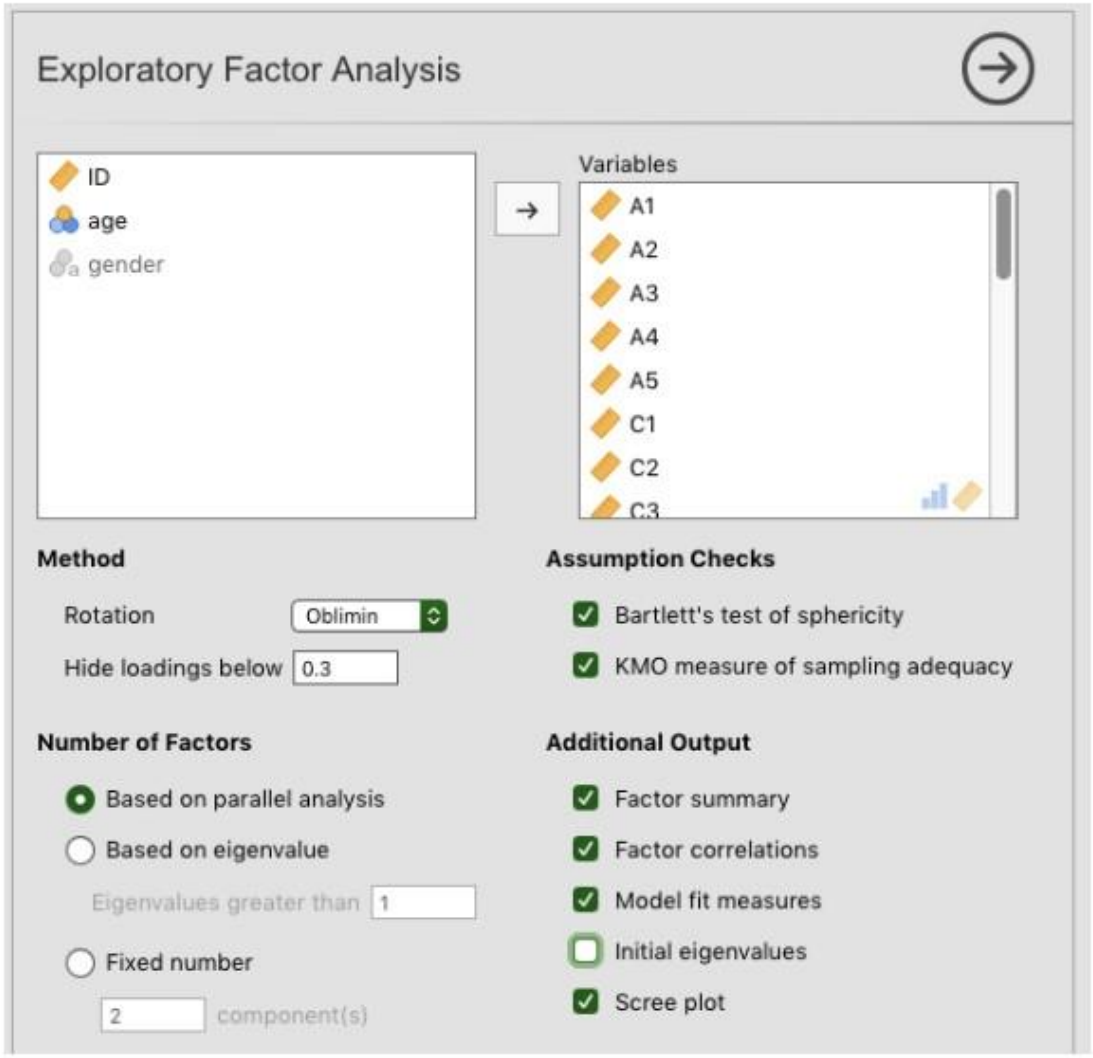


Figure 15-3 : La fenêtre d'analyse de l'AFE de Jamovi

Tout d'abord, vérifions les hypothèses (Figure 15-4). Vous pouvez voir que (1) le test de sphéricité de Bartlett est significatif, donc ce présupposé est satisfait ; et (2) la mesure de l'adéquation de l'échantillonnage (MSA) du KMO est globalement de 0,81, ce qui suggère une bonne adéquation d'échantillonnage. Nous n'avons pas de problème dans ce cas !

Assumption Checks

Bartlett's Test of Sphericity

χ^2	df	p
2204.278	300	<.00001

KMO Measure of Sampling Adequacy

	MSA
Overall	0.808
A1	0.594
A2	0.838
A3	0.820
A4	0.835
A5	0.864
C1	0.800
C2	0.807
C3	0.746
C4	0.793
C5	0.842
E1	0.828
E2	0.845
E3	0.834
E4	0.869
E5	0.910
N1	0.736
N2	0.712
N3	0.768
N4	0.825
N5	0.804
O1	0.840
O2	0.698
O3	0.816
O4	0.743
O5	0.760

Figure 15-4 : Vérification des hypothèse de l'AFE dans Jamovi pour les données du questionnaire de personnalité

Il faut vérifier ensuite le nombre de facteurs à utiliser (ou « extraits » des données). Trois approches différentes sont disponibles :

- Une convention consiste à choisir tous les composants dont les valeurs propres (Eigenvalue) sont supérieures à 1[137]. Cela nous donnerait quatre facteurs avec nos données (essayez pour voir).
- L'examen de l'éboulis, comme sur la Figure 15-5, permet d'identifier le « point d'inflexion ». C'est le point où la pente de la courbe de l'éboulis se stabilise nettement en dessous du « coude ». Cela nous donnerait cinq facteurs avec nos données. L'interprétation des tracés d'éboulis est un peu un art : dans la Figure 15-5, il y a un saut notable de 5 à 6 facteurs, mais dans d'autres tracés d'éboulis, on ne peut pas voir une coupe aussi nette.
- En utilisant une technique d'analyse parallèle, les valeurs propres obtenues sont comparées à celles qui seraient obtenues à partir de données aléatoires. Le nombre de facteurs extraits est le nombre de facteurs dont les valeurs propres sont supérieures à celles que l'on trouverait avec des données aléatoires.

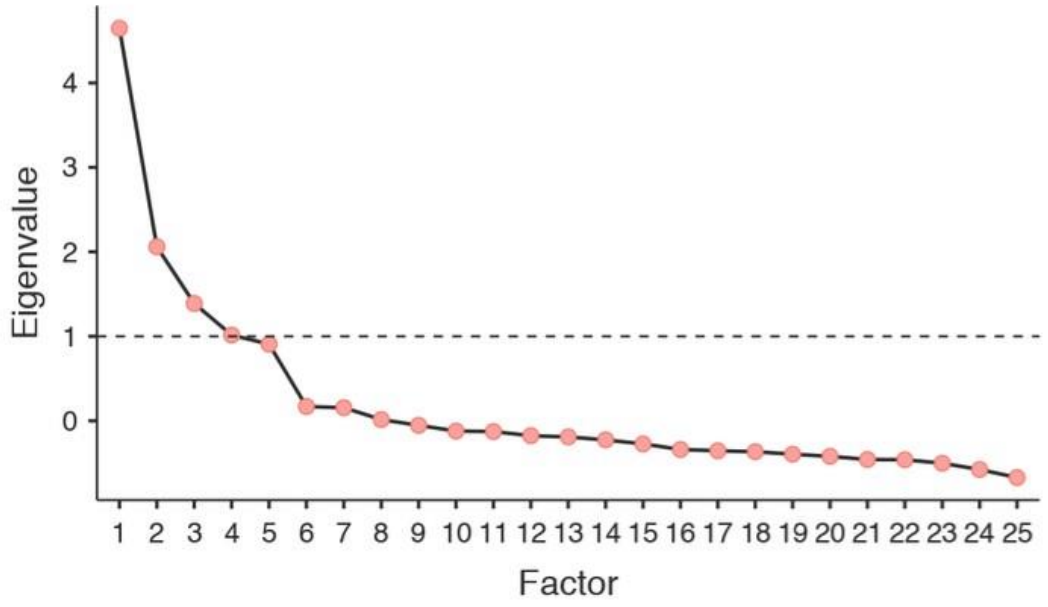


Figure 15-5 : Courbe de l'éboulis de l'EFA avec Jamovi sur les données de personnalité, montrant une inflexion perceptible et une stabilisation après le point 5 (le « coude »)

---

[137] Une valeur propre indique dans quelle mesure la variance des variables observées est prise en compte par un facteur. Un facteur avec une valeur propre > 1 explique plus de variance qu'une seule variable observée.

La troisième approche est satisfaisante selon Fabrigar, Wegener, MacCallum et Strahan (1999) considèrent qu'en pratique, les chercheurs ont tendance à examiner les trois critères précédents et à porter un jugement sur le nombre de facteurs qui sont les plus faciles ou utiles à interpréter. Cela peut être compris comme un « critère de pertinence », et les chercheurs examinent généralement, en plus de la solution de l'une des approches ci-dessus, des solutions comportant un ou deux facteurs plus ou moins importants. Ils adoptent ensuite la solution qui leur semble la plus logique.

Dans le même temps, nous devrions également réfléchir à la meilleure façon d'assurer la rotation de la solution finale. Il existe deux approches principales de la rotation : la rotation orthogonale (p. ex. « varimax ») force les facteurs sélectionnés à ne pas être corrélés, tandis que la rotation oblique (p. ex. « oblimin ») permet la corrélation des facteurs sélectionnés. Les dimensions qui intéressent les psychologues et les spécialistes du comportement ne sont pas souvent des dimensions que l'on s'attendrait à trouver orthogonales, de sorte que les solutions obliques sont sans doute plus raisonnables.

En pratique, si dans une rotation oblique, les facteurs sont substantiellement corrélés (positifs ou négatifs, et > 0.3), comme dans la Figure 15-6 où une corrélation entre deux des facteurs extraits est -0,398, alors cela confirmerait notre intuition de préférer une rotation oblique. Si les facteurs sont, en fait, corrélés, alors une rotation oblique produira une meilleure estimation des vrais facteurs et une meilleure structure simple qu'une rotation orthogonale. Et, si la rotation oblique indique que les facteurs ont des corrélations proches de zéro entre eux, alors le chercheur peut procéder à une rotation orthogonale (qui devrait alors donner la même solution que la rotation oblique).

Factor Statistics

Summary

Factor	SS Loadings	% of Variance	Cumulative %
1	2.578	10.313	10.313
2	2.380	9.518	19.831
3	2.246	8.985	28.816
4	2.337	9.349	38.165
5	1.878	7.514	45.679

Correlation Matrix

	1	2	3	4	5
1	—	–0.195	0.015	0.208	–0.128
2		—	0.184	–0.398	0.261
3			—	–0.300	0.303
4				—	–0.311
5					—

Figure 15-6 : Résumés statistiques des facteurs et corrélations pour une solution à cinq facteurs de l'AFE dans Jamovi

En vérifiant la corrélation entre les facteurs extraits, au moins une corrélation était supérieure à 0,3 (Figure 15-6), de sorte qu'une rotation oblique (oblimin) des cinq facteurs extraits est préférable. Nous pouvons également voir à la Figure 15-6 que la proportion de la variance globale des données attribuable aux cinq facteurs est de 46 %. Le facteur un représente environ 10 % de la variance, les facteurs deux à quatre environ 9 % chacun et le facteur cinq un peu plus de 7 %. Ce n'est pas génial ; il aurait été préférable que la solution globale représente une proportion plus importante de l'écart dans nos données.

Sachez que dans chaque AFE, vous pouvez avoir le même nombre de facteurs que les variables observées, mais chaque facteur supplémentaire que vous incluez ajoutera une variance expliquée plus faible. Si les premiers facteurs expliquent une bonne partie de la variance des 25 variables initiales, alors ces facteurs sont clairement un substitut utile et

plus simple aux 25 variables. Vous pouvez laisser tomber le reste sans perdre trop de la variabilité d'origine. Mais s'il faut 18 facteurs (par exemple) pour expliquer la majeure partie de la variance de ces 25 variables, autant utiliser les 25 variables initiales.

La Figure 15-7 montre les conributions factorielles. C'est-à-dire, comment les 25 éléments de personnalité différents s'appliquent à chacun des cinq facteurs sélectionnés. Nous avons des saturations cachées inférieures à 0,3 (définies dans les options de la Figure 15-3.

Pour les facteurs 1, 2, 3 et 4, le profil des saturations factorielles correspond étroitement aux facteurs présumés indiqués à la Figure 15-2. Ouf ! Et le facteur 5 est assez proche, avec quatre des cinq variables observées qui mesurent supposément « l'ouverture » se chargeant assez bien sur le facteur. La variable 04 ne semble pas tout à fait convenir, car la solution factorielle de la Figure 15-7 suggère qu'elle correle le facteur 3 (quoiqu'avec une corrélation relativement faible) mais pas substantiellement au facteur 5.

Il faut noter également que les variables qui ont été notées « R : codage inversé » à la Figure 15-2 sont celles qui ont des saturations de facteurs négatives. Jetez un coup d'œil aux rubriques A1 (« Je suis indifférent aux sentiments des autres ») et A2 (« Se renseigner sur le bien-être des autres »). Nous pouvons voir qu'un score élevé sur A1 indique un faible degré d'agréabilité, alors qu'un score élevé sur A2 (et toutes les autres variables « A » d'ailleurs) indique un degré d'agréabilité élevé. Par conséquent, A1 sera corrélée négativement avec les autres variables « A », et c'est pourquoi il a une corrélation négative au facteur, comme le montre la Figure 15-7.

Factor Loadings

	Factor					
	1	2	3	4	5	Uniqueness
A1			−0.465			0.738
A2			0.702			0.502
A3			0.687			0.444
A4			0.453			0.667
A5			0.543			0.515
C1		0.720				0.478
C2		0.664				0.542
C3		0.501				0.753
C4		−0.647				0.469
C5		−0.571				0.536
E1				0.561		0.663
E2				0.697		0.352
E3				−0.441	0.419	0.486
E4			0.390	−0.569		0.426
E5				−0.391		0.587
N1	0.818					0.309
N2	0.814					0.359
N3	0.687					0.477
N4	0.461			0.483		0.463
N5	0.451					0.649
O1					0.576	0.626
O2					−0.487	0.689
O3					0.704	0.407
O4			0.307			0.745
O5					−0.493	0.698

Note. 'oblimin' rotation was used

Figure 15-7 : Charges factorielles pour une solution à cinq facteurs dans l'EPT de Jamovi

On peut également voir dans la Figure 15-7 « l'unicité »[138] de chaque variable(uniqueness). L'unicité est la proportion de variance qui est « spécifique » à la variable et qui n'est pas

[138] NdT : j'ai retenu ici le terme proposé par l'ISI glossary (http://isi.cbs.nl/glossary/term3435.htm) pour traduire uniqueness

expliquée par les facteurs[139]. Par exemple, 74 % de la variance de « A1 » ne s'explique pas par les facteurs de la solution à cinq facteurs. Par contre, la variance de « N1 » est relativement faible et n'est pas prise en compte par la solution factorielle (31 %). Il est à noter que plus l'unicité est grande, plus la pertinence ou la saturation de la variable dans le modèle factoriel est faible.

Pour être honnête, il est inhabituel d'obtenir une solution aussi soignée en matière d'AFE. C'est généralement un peu plus compliqué que cela, et souvent l'interprétation de la signification des facteurs est plus difficile. Ce n'est pas souvent que vous avez un ensemble d'éléments aussi bien défini. Le plus souvent, vous aurez tout un tas de variables observées qui, selon vous, peuvent être des indicateurs de facteurs latents, mais vous n'avez pas une idée aussi précise des variables qui y sont associées !

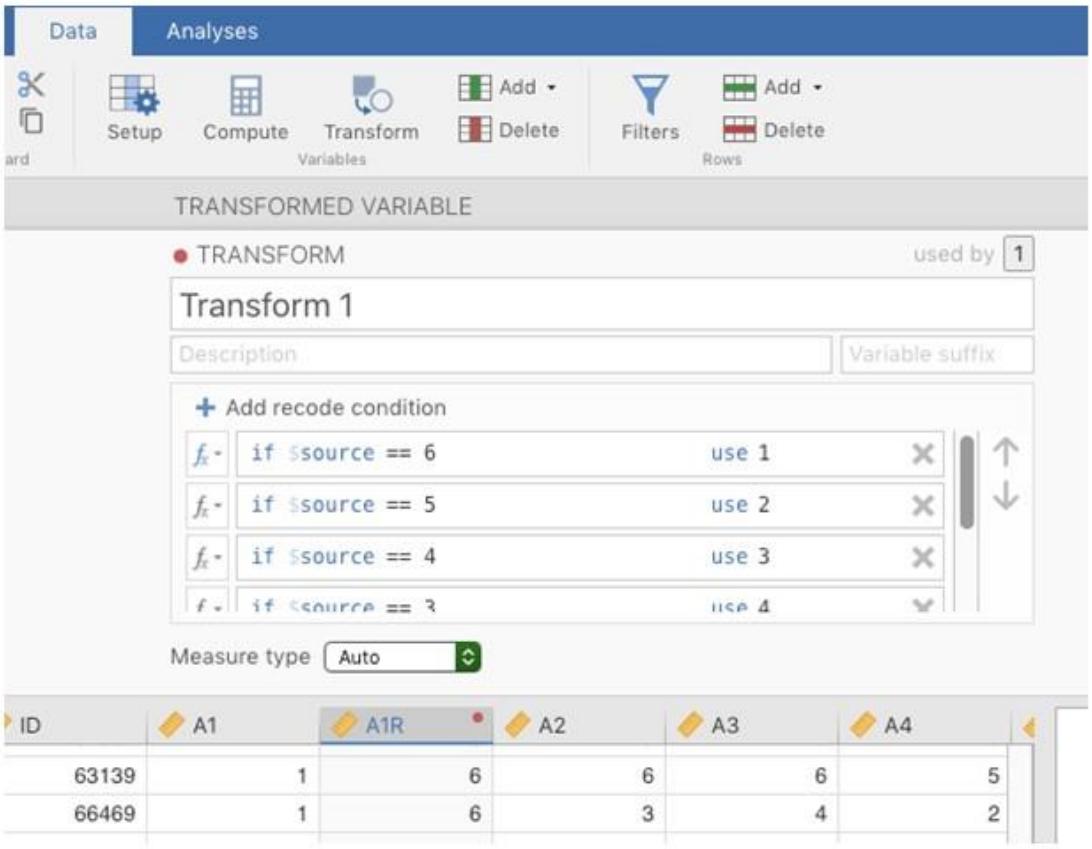


Figure 15-8 : Variable de recodage à l'aide de la commande Jamovi « Transform »

Il semble donc que nous ayons une assez bonne solution à cinq facteurs. Supposons que nous soyons satisfaits de cette solution et que nous souhaitions utiliser nos facteurs dans

---

[139] Dans l'analyse factorielle, le terme « communauté » ou « varaiance commune » est parfois utilisé pour désigner le degré de variance d'une variable qui est pris en compte par la solution factorielle. L'unicité est égale à (1 - *communauté*)

une analyse plus approfondie. L'option simple consiste à calculer un score moyen pour chaque facteur en additionnant le score de chaque variable qui contribue de manière substantielle au facteur et à diviser ensuite par le nombre de variables. Pour chaque personne de notre ensemble de données, cela signifierait, par exemple pour le facteur d'agréabilité, additionner A1 + A2 + A3 + A4 + A5, puis diviser par 5, ce qui signifie que le score factoriel que nous avons calculé est basé sur des scores également pondérés pour chaque variable. On peut le faire avec Jamovi en deux étapes :

1. Recoder A1 en « A1R » en inversant les valeurs de la variable (i.e. 6 = 1; 5 = 2; 4 = 3; 3 = 4; 2 = 5; 1 = 6) en utilisant la commande de transformation de variable de Jamovi (voir Figure 15-8).

2. Calculez une nouvelle variable, appelée « Agreeableness », en calculant la moyenne de A1R, A2, A3, A4 et A5. Pour ce faire, utilisez la commande de calcul de nouvelle variable de Jamovi (voir Figure 15-9).

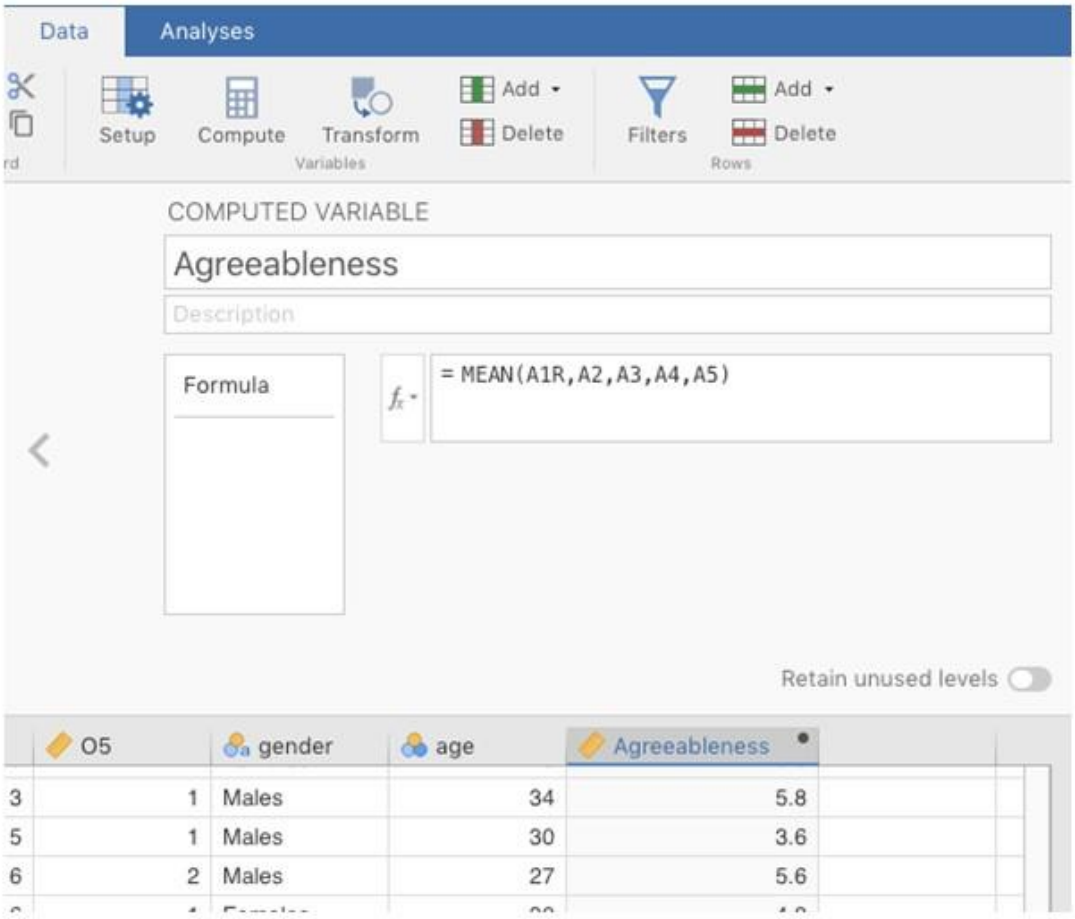


Figure 15-9 : Calcul d'une nouvelle variable score factoriel à l'aide de la commande de Jamovi « Computed variable »

Une autre option consiste à créer un indice de score factoriel pondéré de façon optimale. Nous pouvons utiliser l'éditeur Rj de Jamovi pour le faire dans R. Encore une fois, il y a deux étapes :

1. Utilisez l'éditeur Rj pour exécuter l'AFE dans R selon la même spécification que celle de Jamovi (c'est-à-dire cinq facteurs et rotation oblimin) et calculer les scores factoriels pondérés de façon optimale. Sauvegardez le nouvel ensemble de données, avec les scores factoriels, dans un fichier. Voir la Figure 15-10 et Figure 15-11.

2. Ouvrez le nouveau fichier dans Jamovi et vérifiez que les types de variables ont été correctement définis. Étiqueter les nouvelles variables de score factoriel correspondant aux noms ou définitions des facteurs pertinents (NB : il est possible que les facteurs ne soient pas dans l'ordre prévu, vous devez donc le vérifier).

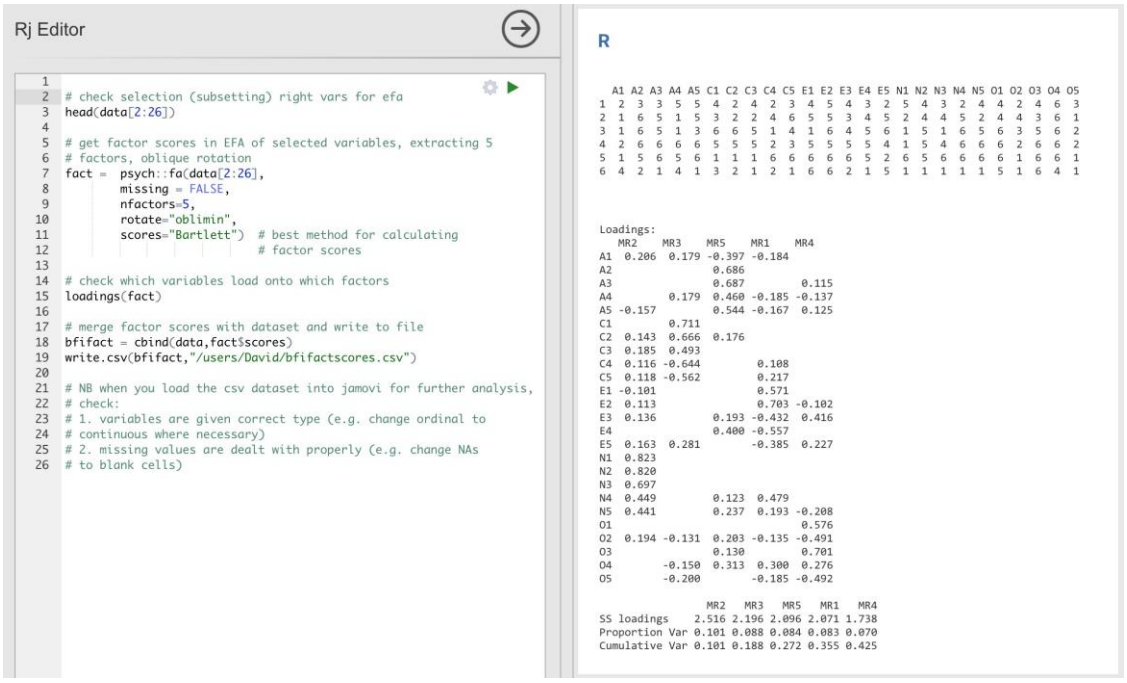


Figure 15-10 : Commandes de l'éditeur Rj pour créer des scores factoriels pondérés de façon optimale pour la solution à cinq facteurs

	O4	O5	gender	age	MR2	MR3	MR5	MR1	MR4
1	6	3	Females	27	0.063	-0.147	-0.656	0.768	-0.061
2	6	1	Females	24	-0.767	-1.812	1.166	1.087	-0.206
3	6	2	Females	19	-0.064	2.037	0.694	3.262	1.179
4	6	2	Females	22	-0.070	0.966	2.388	2.924	1.308
5	6	1	Females	32	0.965	-3.880	2.267	1.423	2.388
6	4	1	Males	24	-2.003	0.071	-3.646	2.090	2.354
7	5	3	Females	29	-1.684	-1.356	0.563	-0.776	-0.480
8	5	5	Females	14	2.100	-0.700	-0.487	-2.038	-0.401
9	5	4	Females	23	-0.772	-1.511	1.704	0.649	-1.940
10	6	1	Females	51	0.588	-1.309	0.826	-1.745	-0.195

Figure 15-11 : Le nouveau fichier de données « bfifactscores.csv » créé dans l'éditeur Rj et contenant les cinq variables de score factoriel. Notez que chacune des nouvelles variables de score factoriel est étiquetée selon l'ordre dans lequel les facteurs sont énumérés dans le tableau des saturations factorielles.

Vous pouvez maintenant procéder à d'autres analyses, soit à l'aide des scores factoriels (une approche basée sur une échelle de score moyen), soit à l'aide des scores factoriels pondérés de manière optimale calculés par l'éditeur Rj. A vous de choisir ! Par exemple, vous pourriez vous interesser aux différences entre les sexes dans chacune de nos échelles de personnalité. Nous l'avons fait pour le score d'agrément que nous avons calculé à l'aide de l'approche factorielle, et bien que le graphique (Figure 15-12) montre que les hommes sont moins agréables que les femmes, cette différence n'est pas significative (Man-Whitney U=5760,5, p = .073).

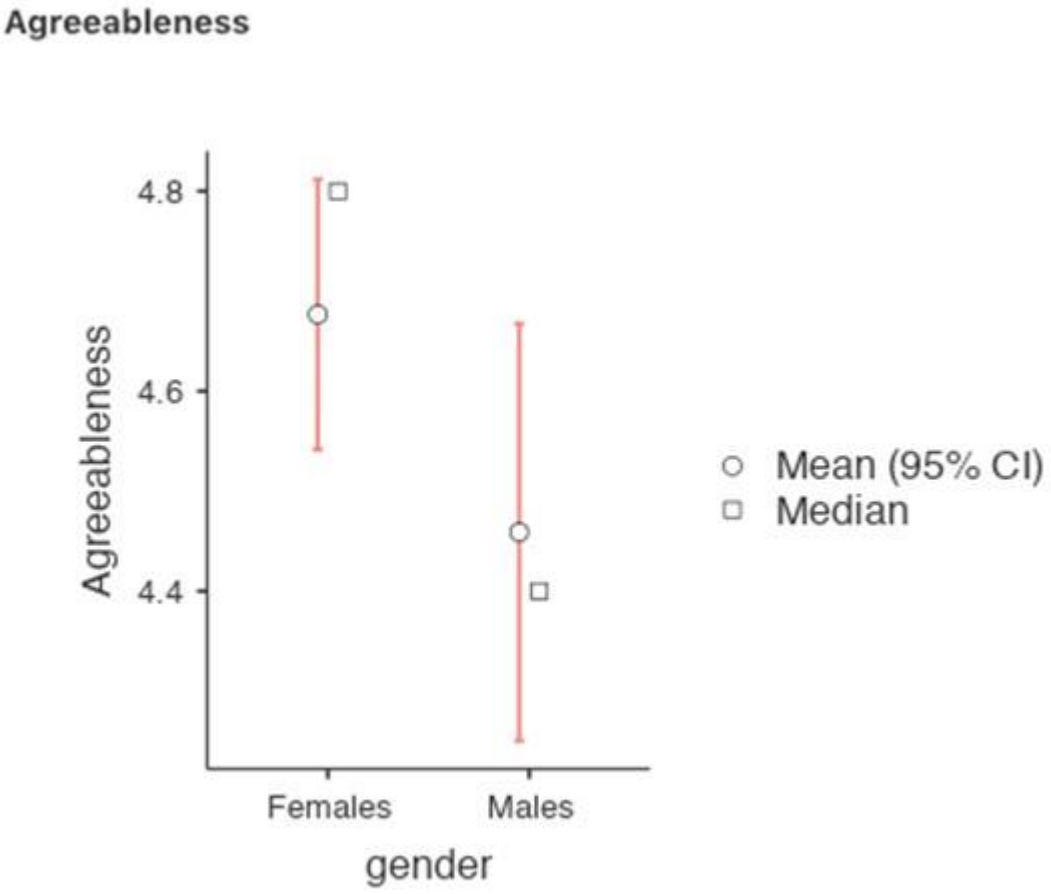


Figure 15-12 : Comparaison des différences entre les scores des hommes et des femmes en fonction du facteur d'agréabilité

J'espère cependant que cela vous a donné une bonne première idée de la manière d'entreprendre l'AEF dans Jamovi. Nous reviendrons sur le fouillis et la nature interprétative (une fois que vous aurez fait le travail technique) de l'analyse factorielle exploratoire dans la vraie vie un peu plus tard, juste avant de nous pencher sur l'analyse factorielle confirmatoire.

Le compte-rendu d'une AFE

Il n'existe pas de méthode officielle normalisée pour rédiger le compte-rendu d'une AFE, et les exemples varient selon les disciplines et les chercheurs. Cela dit, il existe des éléments d'information assez normalisés à l'égard de ce qui suit

Inclure dans votre rédaction

1. Les fondements théoriques pour le domaine que vous étudiez, et plus particulièrement pour les construits que vous souhaitez découvrir à travers l'AFE.
2. Une description de l'échantillon (p. ex. renseignements démographiques, taille de l'échantillon, méthode d'échantillonnage).
3. Une description du type de données utilisées (p. ex. nominales, continues) et des statistiques descriptives.
4. Décrivez comment vous avez procédé pour tester les hypothèses relatives à l'AFE. Les détails concernant les contrôles de sphéricité et les mesures de l'adéquation de l'échantillonnage doivent être rapportés.
5. Expliquer quelle méthode d'extraction de l'analyse factorielle a été utilisée (p. ex., maximum de vraisemblance).
6. Expliquer les critères et le processus utilisés pour décider combien de facteurs ont été extraits dans la solution finale et quels éléments ont été choisis. Expliquer clairement la raison d'être des décisions clés prises au cours du processus de l'EFA.
7. Expliquez quelles méthodes de rotation ont été tentées, les raisons pour lesquelles elles l'ont été et les résultats obtenus.
8. Les saturations factorielles finales (matrice de modèles) doivent être rapportées dans les résultats sous la forme d'un tableau. Ce tableau doit également indiquer l'unicité (ou la communauté) de chaque variable (dans la dernière colonne). Les contributions factorielles doivent être signalées au moyen d'étiquettes descriptives en plus des numéros d'items. Les corrélations entre les facteurs doivent également être incluses, soit au bas de ce tableau, dans un tableau distinct.
9. Les dénominations des facteurs extraits doivent être fournis. Vous souhaiterez peut-être utiliser des noms de facteurs existants, mais en examinant les items et les facteurs réels, vous penserez peut-être qu'un nom différent est plus approprié.

Analyse en composantes principales

Dans la section précédente, nous avons vu que l'AEF cherche à identifier les facteurs latents sous-jacents. Comme nous l'avons vu, dans un exemple, le plus petit nombre de facteurs latents peut être utilisé dans une analyse statistique plus poussée au moyen d'une sorte de score factoriel combiné.

De cette façon, l'AFE est utilisée comme technique de « réduction des données ». Un autre type de technique de réduction des données, parfois considérée comme faisant partie de la famille des AFE, est l'**analyse en composantes principales (ACP)**. Cependant, l'ACP n'identifie pas les facteurs latents sous-jacents. Il crée plutôt un score composite linéaire à partir d'un ensemble plus large de variables mesurées.

L'ACP produit simplement une transformation mathématique des données d'origine sans hypothèses sur la façon dont les variables covarient. Le but de l'ACP est de calculer quelques combinaisons linéaires (composantes) des variables originales qui peuvent être utilisées pour résumer l'ensemble des données observées sans perdre beaucoup d'informations. Toutefois, si l'identification de la structure sous-jacente est un objectif de l'analyse, l'AFE est à privilégier. Comme nous l'avons vu, l'AFE produit des scores factoriels qui peuvent être utilisés à des fins de réduction des données tout comme les scores des composantes principales (Fabrigar et al. 1999).

L'ACP a été populaire en psychologie pour un certain nombre de raisons qu'il vaut la peine d'évoquer. Nous utiliserons les mêmes données bfi_sample.csv. Une grande partie de la procédure est similaire à celle de l'AFE, de sorte que, bien qu'il y ait quelques différences conceptuelles, les étapes sont pratiquement les mêmes[140], et avec de grands échantillons et un nombre suffisant de variables, les résultats de l'ACP et de l'AFE devraient être assez similaires.

Réalisation d'une ACP avec Jamovi

Une fois que vous avez chargé les données bfi_sample.csv, sélectionnez « Factor - Principal Component Analysis » dans la barre de boutons principale de Jamovi pour ouvrir la fenêtre PCA analysis (Figure 15.13). Sélectionnez ensuite les 25 questions de personnalité et transférez-les dans la boîte « Variables ». Cochez les options appropriées, y compris les options « Check assumptions ». Dans « Method » choisissez la méthode de rotation, puis paramétrez « Number of factor to extract » et « Additionnal Outputs ». Voir la figure 15.13 pour les options suggérées pour cette ACP, Comme précédemment, veuillez noter que la méthode de rotation et le nombre de facteurs à extraire sont généralement ajustés pendant l'analyse pour trouver le meilleur résultat, comme décrit ci-dessous.

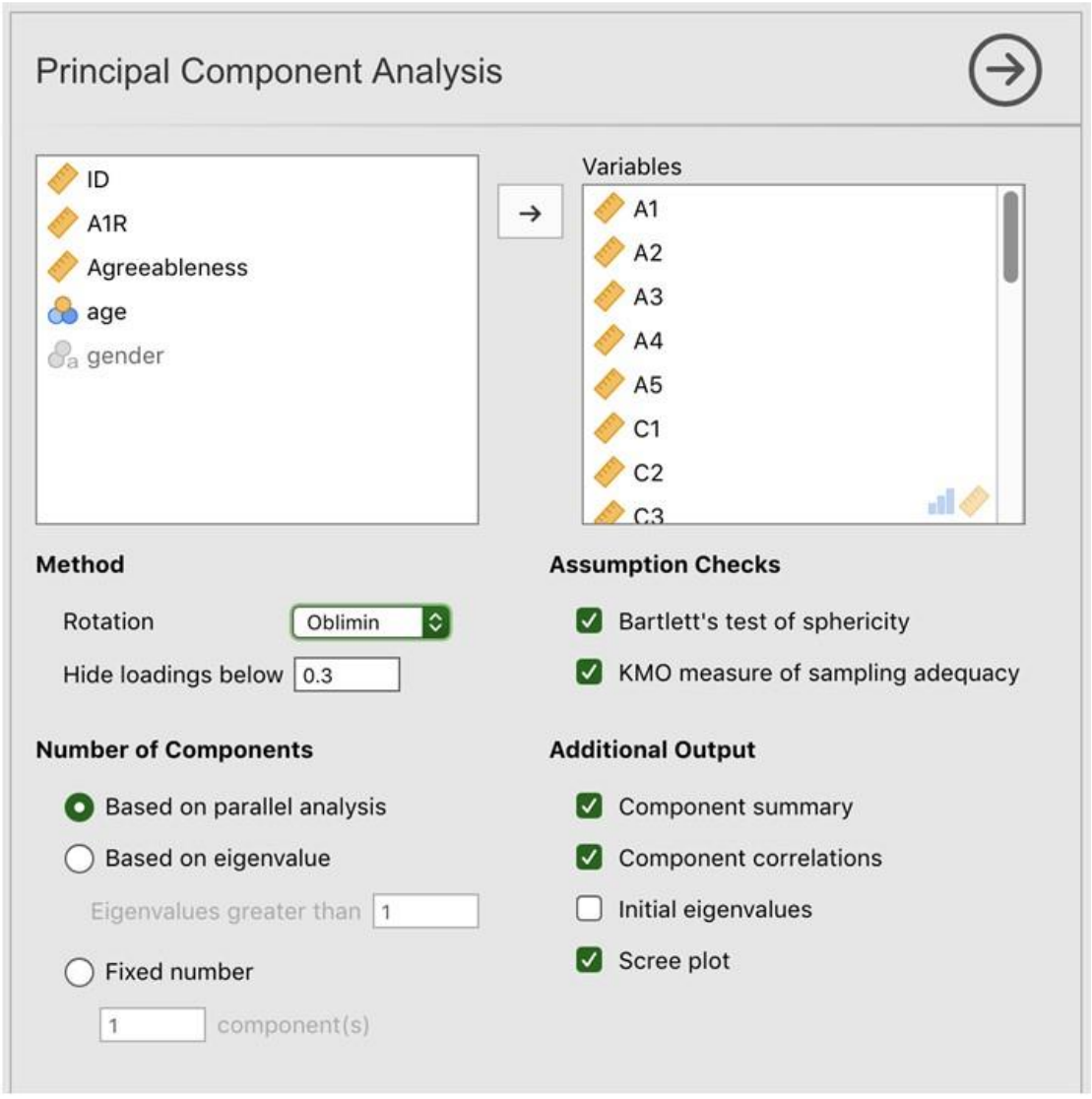
Nous commençons comme précédemment par vérifer les hypothèses sous-jacentes. Comme, vous pouvez voir (1) le test de sphéricité de Bartlett est significatif, cette hypothèse est donc satisfaite ; (2) la mesure de l'adéquation de l'échantillonnage (MSA) du KMO est globalement de 0,81, ce qui suggère une très bonne adéquation d'échantillonnage. Nous pouvons continuer.

La prochaine chose à vérifier est le nombre de composants à utiliser (ou « extraire » des données). Comme pour l'AFE, trois approches différentes sont possibles :

---

[140] … et cela signifie qu'il y a beaucoup de répétitions dans les étapes de l'ACP décrites dans la section suivante. J'en suis et espère que cela conviendra !

- Une convention consiste à choisir toutes les composantes ayant des valeurs propres supérieures à 1, ce qui nous donne deux composantes avec nos données.
- L'examen de l'éboulis, comme sur la Figure 15-15, permet d'identifier le « point d'inflexion ». C'est le point où la pente de la courbe de l'éboulis se stabilise nettement en dessous du « coude ». Encore une fois, cela nous donnerait deux composantes puisque la stabilisation se produit clairement après la deuxième composante.
- En utilisant une technique d'analyse parallèle, les valeurs propres obtenues sont comparées à celles qui seraient obtenues à partir de données aléatoires. Le nombre de composants extraits est le nombre avec des valeurs propres supérieures à ce que l'on trouverait avec des données aléatoires.



- Figure 15-13 : Fenêtre d'analyse en composantes pincipales dans Jamovi

Comme nous l'avons dit en présentant l'AFE, la troisième approche est une bonne approche selon Fabrigar et ses collaborateurs (1999), bien qu'en pratique, les chercheurs aient tendance à examiner les trois et à porter un jugement sur le nombre de composantes qui sont les plus faciles ou les plus utiles à interpréter. Cela peut être compris comme le « critère de pertinence ». Les chercheurs examineront généralement, en plus de la solution de l'une des approches ci-dessus, des solutions comportant une ou deux composantes plus ou moins importantes. Ils adoptent ensuite la solution qui leur semble la plus logique.

Dans le même temps, nous devrions également réfléchir à la meilleure façon d'assurer la rotation de la solution finale. Là encore, comme pour l'AFE, il existe deux approches principales de la rotation : la rotation orthogonale (par exemple « varimax ») force les composantes sélectionnées à ne pas être corrélées, tandis que la rotation oblique (par exemple « oblimin ») permet la corrélation des composantes sélectionnées. Les dimensions qui intéressent les psychologues et les spécialistes du comportement ne sont pas souvent des dimensions que l'on supposerait orthogonales, de sorte que les solutions obliques sont sans doute plus raisonnables. Pratiquement, si dans une rotation oblique, les composantes sont substantiellement corrélées (c.-à-d. > 0.3) alors cela confirmerait notre intuition de préférer une rotation oblique. Si les composantes sont, en fait, corrélées, alors une rotation oblique produira une meilleure estimation des véritables composantes et une meilleure structure simple qu'une rotation orthogonale. Et, si la rotation oblique indique que les composants sont proches de corrélations nulles, alors le chercheur peut procéder à une rotation orthogonale (qui devrait alors donner à peu près la même solution que la rotation oblique).

Assumption Checks

Bartlett's Test of Sphericity

χ^2	df	p
2183.322	300	<.00001

KMO Measure of Sampling Adequacy

	MSA
Overall	0.809
A1	0.568
A2	0.838
A3	0.808
A4	0.837
A5	0.858
C1	0.804
C2	0.797
C3	0.754
C4	0.809
C5	0.851
E1	0.834
E2	0.843
E3	0.854
E4	0.870
E5	0.911
N1	0.737
N2	0.714
N3	0.762
N4	0.819
N5	0.800
O1	0.840
O2	0.688
O3	0.823
O4	0.742
O5	0.738

Figure 15-14 : Vérification des hypothèses de l'ACP dans Jamovi pour les données d'élément de personnalité

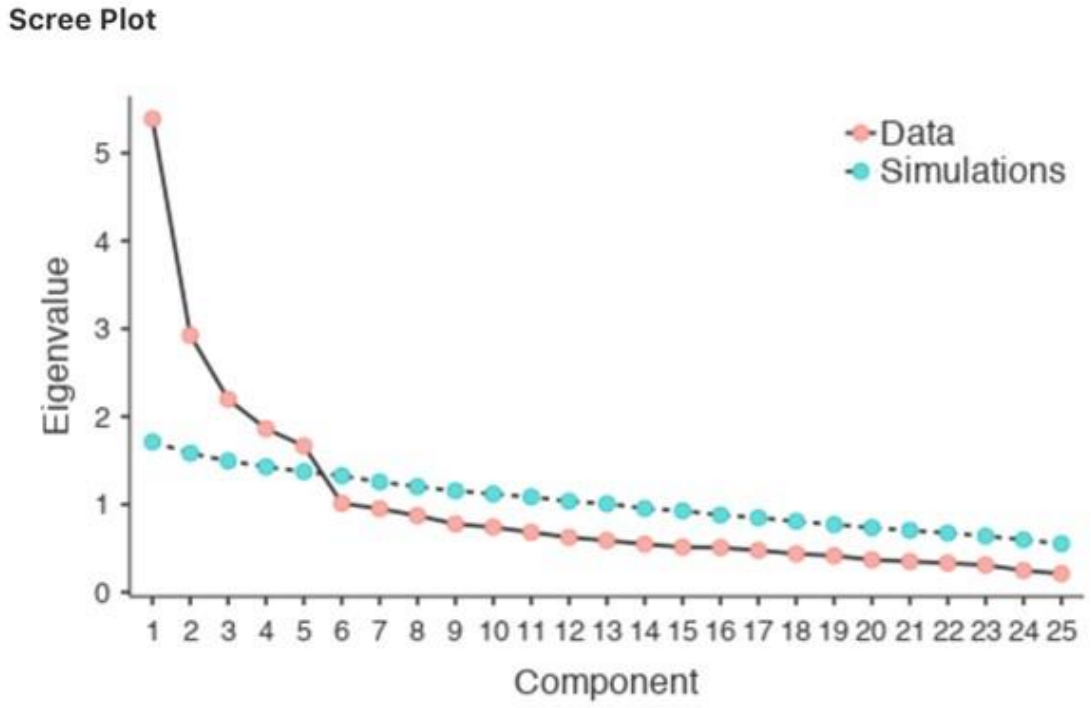


Figure 15-15 : Courbe de l'eboulis de l'ACP dans Jamovi sur les données d'élément de personnalité montrant le point d'inflexion, le « coude », après le composant 5

Dans la Figure 15-16, nous voyons qu'aucune des corrélations n'est > 0,3, il est donc approprié de passer à la rotation orthogonale (varimax).

Dans la Figure 15-16, nous avons également la proportion de la variance globale des données qui est attribuable aux deux composantes. Les composantes une et deux représentent un peu plus de 12 % de la variance chacune. Ensemble, les cinq composantes de la solution représentent un peu plus de la moitié de la variance (56 %) des données observées. Sachez que dans chaque ACP, vous pourriez potentiellement avoir le même nombre de composantes que les variables observées, mais chaque composante supplémentaire que vous incluez ajoutera une plus petite quantité de variance expliquée. Si les premières composantes expliquent une bonne partie de la variance des 25 variables initiales, alors ces composantes sont clairement un substitut utile et plus simple pour les 25 variables. Vous pouvez laisser tomber le reste sans perdre trop de la variabilité d'origine. Mais s'il faut 18 composantes pour expliquer la plus grande partie de la variance de ces 25 variables, autant utiliser les 25 variables originales.

La Figure 15-17 montre les correlations des composants. C'est-à-dire, comment les 25 éléments de personnalité sont corrélés à chacun des composants sélectionnés. Nous avons des corrélations inférieures à 0,4 masquées (définies dans les options illustrées à la Figure 15-13), car nous nous intéressions aux éléments ayant une corrélation importante et le fait

de fixer le seuil à la valeur supérieure 0,4 nous a également fourni une solution plus propre et plus claire.

Pour les composants 1, 2, 3 et 4, la répartition des saturations des composants correspond étroitement aux facteurs présumés indiqués à la Figure 15-2.

Component Statistics

Summary

Component	SS Loadings	% of Variance	Cumulative %
1	3.136	12.542	12.542
2	3.090	12.361	24.903
3	2.856	11.425	36.328
4	2.682	10.729	47.056
5	2.272	9.086	56.143

Correlation Matrix

	1	2	3	4	5
1	—	–0.136	–0.124	0.025	–0.090
2		—	0.259	0.175	0.136
3			—	0.091	0.108
4				—	0.118
5					—

Figure 15-16 : Statistiques sommaires des composants et corrélations pour une solution à cinq composants dans l'ACP dans Jamovi

Le composant 5 est assez proche, avec quatre des cinq variables observées dont on suppose qu'elles mesurent « l'ouverture » sontv assez bien corrélés avec le composant. La variable 04 ne semble pas tout à fait convenir, car la solution des composants de la Figure 15-17 suggère qu'elle est liée au composant 4 (bien qu'avec une corrélation relativement faible) mais pas de manière substantielle sur le composant 5.

On peut également voir dans la Figure 15-17 l'unicité de chaque variable. L'unicité est la proportion de variance qui est « spécifique » à la variable et qui n'est pas expliquée par les composantes. Par exemple, 58 % de la variance de « A1 » ne s'explique pas par les composantes de la solution à cinq composantes. En revanche, la variance de « N1 » est relativement faible et n'est pas prise en compte par la solution du composant (30%). Il est à

noter que plus l'unicité est grande, plus la pertinence ou la corrélation de la variable dans le modèle des composantes est faible.

J'espère que cela vous a donné une première idée claire de la façon d'entreprendre une ACP dans Jamovi, et qu'elle est conceptuellement différente mais pratiquement similaire (si l'on dispose des bonnes données) à l'AFE.

Component Loadings

	Component					
	1	2	3	4	5	Uniqueness
A1				–0.554		0.581
A2				0.743		0.411
A3				0.735		0.384
A4				0.515		0.544
A5				0.597		0.463
C1			0.746			0.392
C2			0.741			0.395
C3			0.621			0.582
C4			–0.673			0.399
C5			–0.634			0.430
E1		–0.664				0.547
E2		–0.746				0.318
E3		0.656				0.387
E4		0.661		0.409		0.362
E5		0.585				0.492
N1	0.809					0.298
N2	0.812					0.314
N3	0.785					0.370
N4	0.652	–0.405				0.386
N5	0.592					0.503
O1					0.634	0.486
O2					–0.632	0.505
O3					0.673	0.361
O4				0.429		0.557
O5					–0.669	0.496

Note. 'varimax' rotation was used

Figure 15-17 : Charges de composants pour une solution à cinq composants dans l'ACP de Jamovi

Vous pouvez ensuite créer les scores des composantes de la même manière que pour l'AFE. Cependant, si vous choisissez de créer un score de composant pondéré de manière optimale, les commandes et la syntaxe dans l'éditeur jamovi Rj sont un peu différentes. Voir la Figure 15-18.

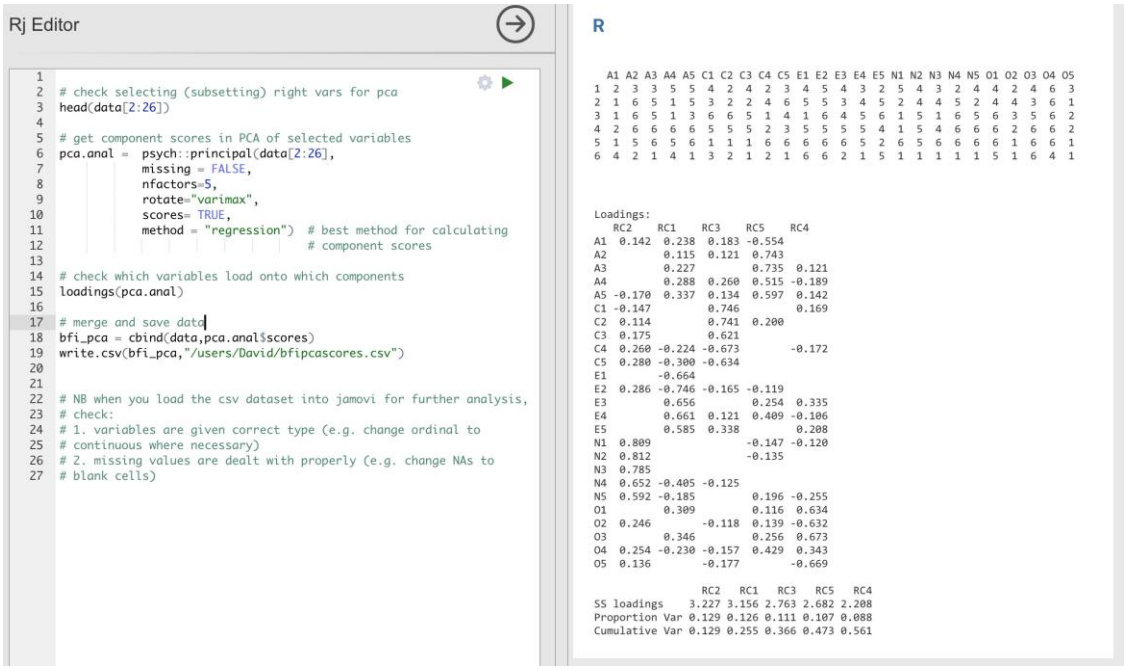


Figure 15-18 : Commandes de l'éditeur Rj pour créer des scores de composants pondérés de manière optimale pour la solution à cinq composants

Analyse factorielle confirmatoire

Notre tentative d'identifier les facteurs latents sous-jacents à l'aide de l'AEF à l'aide de questions soigneusement sélectionnées dans le questionnaire de personnalité nous semble assez réussie. La prochaine étape dans notre quête d'une mesure utile de la personnalité est de vérifier les facteurs latents que nous avons identifiés dans l'AEF originale avec un échantillon différent. Nous voulons voir si les facteurs tiennent le coup, si nous pouvons confirmer leur existence avec des données différentes. Il s'agit d'un contrôle plus rigoureux, comme nous le verrons plus loin. C'est ce qu'on appelle l'**analyse factorielle confirmatoire (AFC),** car nous chercherons, sans le vouloir, à confirmer une structure factorielle latente préétablie.[141]

---

[141] Soit dit en passant, étant donné que nous avions une idée assez précise de nos facteurs « supposés » initiaux, nous aurions simplement pu passer directement à l'AFC et sauter l'étape de l'AFE. La question de savoir si vous utilisez l'AFE et passez ensuite à l'AFC, ou si vous passez directement à l'AFC, est une question de jugement et de confiance dans le fait que vous avez initialement la structure adéquate (en termes de nombre de facteurs et de variables). Au début de l'élaboration des échelles ou pour l'identification des constructions latentes sous-jacentes, les chercheurs ont tendance à utiliser l'AFE. Ensuite, lorsqu'ils se

Dans l'AFC, au lieu de faire une analyse où l'on voit comment les données s'associent d'un point de vue exploratoire, nous imposons plutôt une structure aux données, comme dans la Figure 15-19, et nous voyons dans quelle mesure les données correspondent à notre structure prédéfinie. En ce sens, nous entreprenons une analyse confirmatoire, pour voir dans quelle mesure un modèle préétabli est confirmé par les données observées.

Une simple analyse factorielle de confirmation (AFC) des éléments de personnalité permettrait donc de préciser cinq facteurs latents, comme le montre la Figure 15-19, chacun mesuré par cinq variables observées. Chaque variable est une mesure d'un facteur latent sous-jacent. Par exemple, A1 est prédit par le facteur latent sous-jacent « Agreeableness ». Dans la mesure où A1 n'est pas une mesure parfaite du facteur « Agreeableness », il y a un terme d'erreur, e, qui lui est associé. En d'autres termes, e représente la variance dans A1 qui n'est pas prise en compte par le facteur d'agréabilité. C'est ce qu'on appelle parfois l'**erreur de mesure**.

L'étape suivante consiste à déterminer si les facteurs latents peuvent être corrélés à notre modèle. Comme nous l'avons mentionné précédemment, dans les sciences psychologiques et comportementales, les constructions sont souvent liées les unes aux autres, et nous pensons aussi que certains de nos facteurs de personnalité peuvent être corrélés les uns aux autres. Ainsi, dans notre modèle, nous devrions permettre à ces facteurs latents de varier conjointement, comme le montrent les flèches doubles de la Figure 15-19.

En même temps, nous devrions nous demander s'il existe une raison valable et systématique pour que certains termes d'erreur soient corrélés les uns aux autres. L'une des raisons pourrait être qu'il existe une caractéristique méthodologique commune pour des sous-ensembles particuliers des variables observées, de sorte que les variables observées pourraient être corrélées pour des raisons méthodologiques plutôt que de facteurs latents importants. Nous reviendrons sur cette possibilité dans une section ultérieure, mais, pour l'instant, il n'y a aucune raison claire qui justifierait la corrélation de certains termes d'erreur entre eux.

En l'absence de termes d'erreur corrélés, le modèle que nous testons pour voir dans quelle mesure il correspond à nos données d'observation est conforme à ce qui est indiqué à la Figure 15-19. Seuls les paramètres inclus dans le modèle sont censés se trouver dans les données, de sorte que dans l'ACC, tous les autres paramètres possibles sont mis à zéro. Ainsi, si ces autres paramètres ne sont pas nuls (par exemple, il peut y avoir une charge substantielle de A1 sur le facteur latent Extraversion dans les données observées, mais pas dans notre modèle) alors nous pouvons trouver une mauvaise correspondance entre notre modèle et les données observées.

---

rapprochent d'une échelle finale, ou s'ils veulent vérifier une échelle établie avec un nouvel échantillon, alors l'AFC est une bonne option.

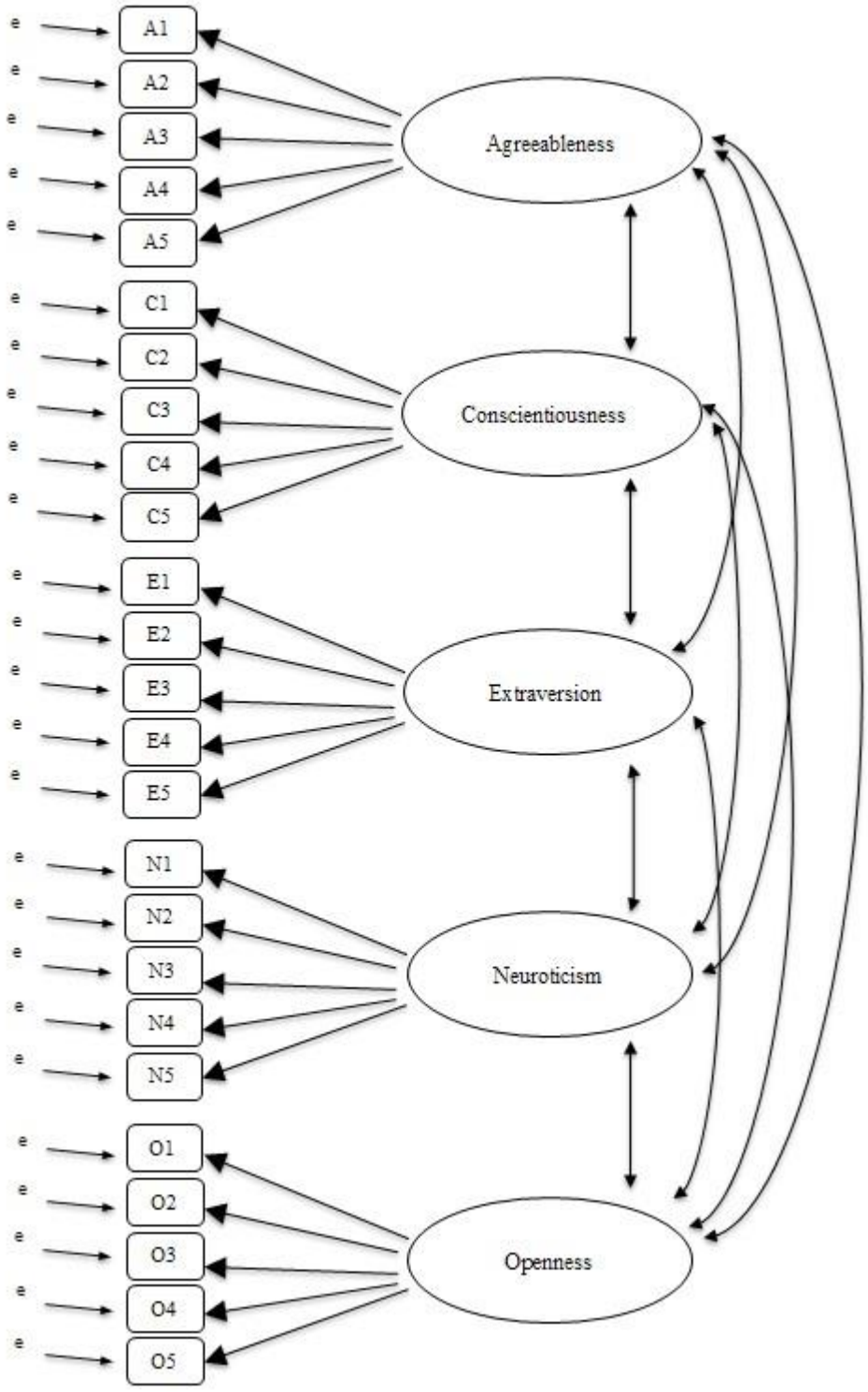


Figure 15-19 : Pré-spécification initiale de la structure des facteurs latents pour les échelles de personnalité à cinq facteurs, à utiliser en CFA

Voyons comment nous avons réalisée cette analyse CFA avec Jamovi.

L'AFC avec Jamovi

Ouvrez le fichier bfisample2.csv, vérifiez que les 25 variables sont codées comme ordinales (ou continues ; cela ne fera aucune différence pour cette analyse) dans Jamovi pour faire l'AFC :

- Sélectionnez « Factor - Confirmatory Factor Analysis » dans la barre de boutons Jamovi principale pour ouvrir la fenêtre CFA analysis (Figure 15-20).
- Sélectionnez les 5 variables A et transférez-les dans la case 'Factors' et donnez ensuite le label « Agreeableness ».
- Créez un nouveau Facteur dans la case « Facteurs » et nommez-le « Conscientiousness ». Sélectionnez les 5 variables C et transférez-les dans la case « Factors » sous le titre « Conscientiousness ».
- Créez un autre nouveau facteur dans la case 'Factors' et nommez-le « Extraversion ». Sélectionnez les 5 variables E et les transférer dans la case « Factors » sous le titre « Extraversion ».
- Créez un autre nouveau facteur dans la boîte « Factors » et nommez-le « Neuroticism ». Sélectionnez les 5 variables N et les transférer dans la case « Factors » sous le titre « Neuroticism ».
- Créez un autre nouveau Facteur dans la case « Factors » et nommez-le « Openness ». Sélectionnez les 5 variables O et les transférer dans la case « Factors » sous le titre « Openness ».
- Cochez les autres options pertinentes, les valeurs par défaut sont correctes pour ce premier travail, vous pouvez également cocher l'option « Path diagram » sous « Plots » pour demander à Jamovi produire un diagramme (assez) similaire à notre Figure 15-19.

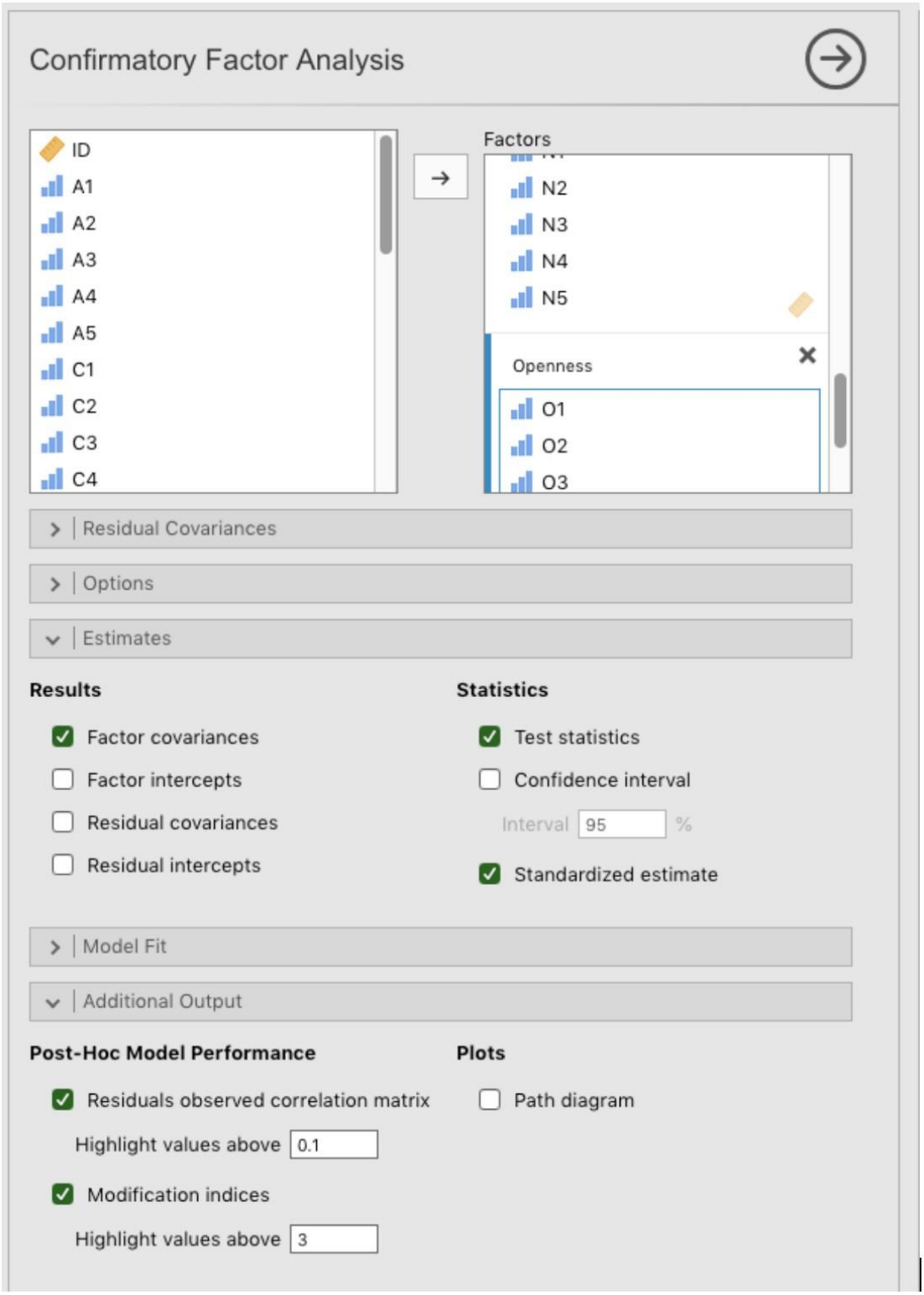


Figure 15-20 : La fenêtre d'analyse pour analyse factorielle confirmatoire (AFC) dans Jamovi

Une fois que nous avons mis en place l'analyse, nous pouvons examiner la fenêtre des résultats de Jamovi et voir ce qu'il en est. La première chose à examiner est l'**ajustement du modèle** (Figure 15-21), car cela nous indique si notre modèle correspond bien aux données observées. Dans notre modèle, seules les covariances prédéfinies sont estimées, y compris les corrélations de facteurs par défaut. Tout le reste est mis à zéro.

Il existe plusieurs façons d'évaluer l'adéquation du modèle. La première est une statistique du chi carré qui, si elle est petite, indique que le modèle est bien adapté aux données. Cependant, la statistique du chi carré utilisée pour évaluer l'ajustement du modèle est assez sensible à la taille de l'échantillon, ce qui signifie qu'avec un grand échantillon, un ajustement suffisamment bon entre le modèle et les données produit presque toujours une valeur du chi carré importante et significative (p.<.05).

Nous avons donc besoin d'autres moyens d'évaluer l'adéquation du modèle. Dans Jamovi plusieurs sont fournis par défaut. Il s'agit de l'indice d'ajustement comparatif (CFI), de l'indice de Tucker Lewis (TLI) et de l'erreur quadratique moyenne quadratique approximative (RMSEA) ainsi que de l'intervalle de confiance à 90 % pour la RMSEA. Quelques règles empiriques utiles nous indiquent qu'un ajustement satisfaisant est indiqué par CFI > 0.9, TLI > 0.9, et RMSEA entre 0,05 et 0,08. Un bon ajustement est CFI > 0.95, TLI > 0.95, et RMSEA et CI supérieur pour RMSEA < 0.05.

Model Fit

Test for Exact Fit

χ^2	df	p
739.726	265	<.00001

Fit Measures

			RMSEA 90% CI	
CFI	TLI	RMSEA	Lower	Upper
0.762	0.731	0.085	0.077	0.092

Figure 15-21 : Résultats de l'ajustement du modèle AFC Jamovi pour notre modèle CFA

Ainsi, en regardant la Figure 15-21, nous pouvons voir que la valeur du chi carré est grande et très significative. La taille de notre échantillon n'est pas trop grande, ce qui indique peut-

être un mauvais ajustement. La CFI est de 0,762 et la TLI est de 0,731, ce qui indique un mauvais ajustement entre le modèle et les données. La RMSEA est de 0,085 avec un intervalle de confiance à 90 % de 0,077 à 0,092, ce qui, encore une fois, n'indique pas un bon ajustement.

Plutôt décevant, hein ? Mais ce n'est peut-être pas trop surprenant étant donné que, dans l'ancien AFE, nous utilisions un ensemble de données semblable (section 15.1), le modèle à cinq facteurs ne représentait que la moitié environ de la variance des données.

Factor Loadings

Factor	Indicator	Estimate	SE	Z	p	Stand. Estimate
Agreeableness	A1	−0.468	0.099	−4.727	<.00001	−0.328
	A2	0.771	0.079	9.761	<.00001	0.621
	A3	1.061	0.082	13.002	<.00001	0.788
	A4	0.749	0.096	7.830	<.00001	0.516
	A5	0.941	0.082	11.443	<.00001	0.711
Conscientiousness	C1	0.802	0.076	10.531	<.00001	0.670
	C2	0.747	0.087	8.558	<.00001	0.569
	C3	0.539	0.087	6.191	<.00001	0.421
	C4	−1.067	0.086	−12.414	<.00001	−0.760
	C5	−1.121	0.104	−10.775	<.00001	−0.678
Extraversion	E1	−0.898	0.106	−8.512	<.00001	−0.550
	E2	−1.261	0.100	−12.667	<.00001	−0.761
	E3	0.911	0.092	9.889	<.00001	0.627
	E4	1.040	0.092	11.278	<.00001	0.692
	E5	0.778	0.083	9.370	<.00001	0.596
Neuroticism	N1	1.339	0.094	14.174	<.00001	0.821
	N2	1.215	0.093	13.067	<.00001	0.765
	N3	1.217	0.097	12.487	<.00001	0.747
	N4	0.956	0.110	8.715	<.00001	0.574
	N5	0.842	0.110	7.631	<.00001	0.494
Openness	O1	0.665	0.082	8.119	<.00001	0.580
	O2	−0.653	0.115	−5.660	<.00001	−0.403
	O3	0.999	0.088	11.301	<.00001	0.831
	O4	0.233	0.085	2.735	0.00624	0.197
	O5	−0.514	0.093	−5.521	<.00001	−0.405

Figure 15-22 : Le tableau des saturation factorielles dans l'AFC réalisée avec Jamovi pour notre modèle

Examinons ensuite les saturations factorielles et les estimations de covariance des facteurs, illustrées aux Figure 15-22 et Figure 15-23. La statistique *Z* et la valeur *p* de chacun de ces paramètres indiquent qu'ils apportent une contribution raisonnable au modèle (c.-à-d. qu'ils ne sont pas nuls), de sorte qu'il ne semble y avoir aucune raison de supprimer du

modèle les relations facteurs-variables ou les corrélations facteur-facteur spécifiés. Souvent, les estimations normalisées sont plus faciles à interpréter, et elles peuvent être précisées dans l'option « Estimates ». Ces tableaux peuvent utilement être incorporés dans un rapport écrit ou un article scientifique.

Factor Covariances

		Estimate	SE	Z	p	Stand. Estimate
Agreeableness	Agreeableness	1.000 [a]				
	Conscientiousness	0.335	0.074	4.507	<.00001	0.335
	Extraversion	0.581	0.063	9.213	<.00001	0.581
	Neuroticism	−0.164	0.077	−2.140	0.03233	−0.164
	Openness	0.424	0.071	5.948	<.00001	0.424
Conscientiousness	Conscientiousness	1.000 [a]				
	Extraversion	0.504	0.065	7.756	<.00001	0.504
	Neuroticism	−0.289	0.074	−3.917	0.00009	−0.289
	Openness	0.277	0.081	3.440	0.00058	0.277
Extraversion	Extraversion	1.000 [a]				
	Neuroticism	−0.238	0.076	−3.122	0.00179	−0.238
	Openness	0.532	0.067	7.949	<.00001	0.532
Neuroticism	Neuroticism	1.000 [a]				
	Openness	−0.182	0.078	−2.320	0.02033	−0.182
Openness	Openness	1.000 [a]				

[a] fixed parameter

Figure 15-23 : Le tableau des covariances des facteurs de notre modèle d'AFC calculées avec Jamovi

Comment pourrions-nous améliorer le modèle ? L'une des options consiste à revenir en arrière et à réfléchir aux éléments ou mesures que nous utilisons et à la façon dont ils pourraient être améliorés ou modifiés. Une autre option est de faire quelques ajustements *post-hoc* sur le modèle pour améliorer l'ajustement. Une façon d'y parvenir est d'utiliser des « indices de modification », demandés dans option « Additionnal Output » dans Jamovi (voir Figure 15-24).

Ce que nous recherchons, c'est la valeur la plus élevée de l'indice de modification (MI). Nous jugerons alors s'il est judicieux d'ajouter ce terme supplémentaire dans le modèle, en utilisant une analyse *post-hoc.* Par exemple, nous pouvons voir à la Figure 15-24 que le MI le plus élevé pour les facteurs qui ne sont pas déjà dans le modèle est une valeur de 28,786 pour la contribution de N4 (« Often feel blue ») sur le facteur latent « Extraversion ». Cela indique que si nous ajoutons cette relation dans le modèle, la valeur du chi carré diminuera d'environ le même montant.

Factor Loadings – Modification Indices

	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
A1		11.138	14.214	1.532	0.391
A2		1.203	1.018	3.645	0.534
A3		1.760	4.695	2.952	0.001
A4		6.992	4.152	0.060	2.299
A5		3.843	11.412	8.179	3.501
C1	4.215		1.962	0.545	0.147
C2	1.399		0.009	12.077	0.007
C3	0.849		1.542	12.774	0.863
C4	0.960		1.053	2.708	0.214
C5	0.226		1.028	4.746	0.206
E1	13.545	0.501		1.357	3.139
E2	5.360	1.670		19.013	2.334
E3	4.395	5.853		6.202	28.017
E4	22.831	1.511		0.302	9.370
E5	2.400	8.775		2.507	2.894
N1	1.147	0.591	1.083		0.147
N2	1.034	9.833	7.580		3.178
N3	0.207	0.007	0.013		0.009
N4	1.653	14.031	28.786		0.652
N5	1.129	0.028	0.843		2.896
O1	0.069	0.190	1.889	0.280	
O2	6.429	2.426	8.060	6.849	
O3	2.711	2.093	7.699	1.037	
O4	1.851	13.387	10.541	8.873	
O5	1.584	4.494	2.974	2.188	

Figure 15-24 : Indices de modification des saturations des facteurs de l'AFC dans Jamovi

Mais dans notre modèle, l'ajout de cette relation n'a sans doute aucun sens théorique ou méthodologique, et n'est donc pas une bonne idée (à moins que vous ne puissiez trouver un argument persuasif selon lequel le fait d'être « Sentir souvent triste » mesure à la fois le névrotisme et l'extraversion). Moi, je ne peux pas. Mais, pour les besoins de l'argumentation, supposons que cela ait un sens et ajoutons ce lien dans le modèle. Retournez à la fenêtre d'analyse CFA (voir Figure 15-20) et ajoutez N4 dans le facteur d'Extraversion. Les résultats du AFC vont maintenant changer (non montrés) ; le khi-carré est descendu à environ 709 (une baisse d'environ 30, à peu près similaire à la taille du MI) et les autres indices d'ajustement se sont également améliorés, mais seulement un peu. Mais ce n'est pas assez ; ce n'est toujours pas un bon modèle.

Si vous ajoutez de nouveaux paramètres à un modèle à l'aide des valeurs MI, vérifiez toujours à nouveau les tables MI après chaque nouvel ajout, car les MI sont actualisés à chaque fois.

C5	E1	E2	E3	E4	E5	N1	N2	N3	N4	N5	O1	O2	O3	O4	O5
5.870	0.040	0.102	7.924	0.057	6.913	1.281	11.771	1.242	1.632	2.131	6.011	0.846	0.803	8.111	1.052
2.673	0.012	3.320	0.835	2.511	1.660	1.165	0.946	0.200	7.178	8.298	0.187	6.272	0.210	12.008	2.159
1.082	5.674	7.547	4.529	3.073	0.820	0.274	0.578	1.074	0.647	0.976	0.028	0.504	0.424	1.252	0.139
2.728	0.352	0.322	0.097	2.191	0.280	3.571	0.086	1.583	2.039	1.176	0.089	2.739	0.284	2.700	1.921
1.783	0.250	3.898	0.426	1.663	2.283	1.927	0.150	0.388	0.187	0.030	0.039	2.981	0.254	0.059	0.168
0.685	0.094	2.141	3.045	0.302	2.286	0.988	0.229	4.310	3.743	1.626	0.173	0.427	0.015	0.056	8.834
0.604	2.037	3.179	0.135	0.562	5.965	0.148	1.892	0.037	2.604	6.324	0.251	9.810	0.692	4.285	0.093
0.851	0.007	0.009	0.318	2.512	2.376	0.915	0.324	0.145	0.179	6.520	0.001	0.538	0.319	4.085	3.643
3.616	0.117	3.046	0.246	1.070	0.069	7.971	1.731	6.531	1.520	4.203	0.011	11.847	5.316	5.080	6.374
	0.002	1.414	0.449	0.382	0.002	6.459	0.807	1.079	10.438	2.168	0.001	1.297	0.342	6.642	1.422
		10.713	0.967	0.036	0.359	2.781	1.630	1.402	4.293	0.057	2.557	0.500	0.027	0.926	0.044
			2.100	1.189	2.349	0.032	2.223	2.668	15.404	7.006	0.132	0.419	0.017	11.442	1.406
				0.930	1.226	1.496	0.054	1.211	1.549	2.720	3.341	2.672	16.896	2.860	0.597
					1.264	3.172e−4	0.866	1.440	2.445	6.863	0.124	10.044	1.115	0.043	3.263
						1.362	6.538	1.536	1.102	0.001	5.970	0.895	0.419	0.330	0.233
							44.997	4.122	25.696	0.250	1.419	0.088	0.210	0.753	0.110
								4.375	3.476	6.245	3.233	2.527	0.817	0.038	4.680
									19.259	0.911	1.188	0.058	0.080	0.021	6.966
										8.571	0.008	0.249	0.742	10.891	0.649
											2.314	3.257	0.128	1.919	2.608
												0.042	4.687	0.332	0.953
													0.060	3.045	12.948
														2.928	3.236
															0.173

Figure 15-25 : Indices de modification des covariances résiduelles dans l'AFC avec Jamovi

Il existe également un tableau des Indices de Modifications de Covariance Résiduelle produit par jamovi (Figure 15-25). En deux mots, c'est un tableau indiquant les erreurs corrélées, qui si elles étaient ajoutées au modèle, amélioreraient le plus l'ajustement du modèle. C'est une bonne idée de vérifier les deux tables de MI en même temps, de repérer le MI le plus grand, de se demander si l'ajout du paramètre suggéré peut être raisonnablement justifié et, si possible, de l'ajouter au modèle. Ensuite, vous pouvez recommencer à chercher le plus grand MI dans les résultats recalculés.

Vous pouvez continuer de cette façon aussi longtemps que vous le souhaitez, en ajoutant des paramètres au modèle basé sur le MI le plus grand, et finalement vous obtiendrez un ajustement satisfaisant. Mais il y aura aussi une forte probabilité qu'en faisant cela vous ayez créé un monstre ! Un modèle laid et déformé qui n'a aucun sens théorique, ni aucune pureté. En d'autres termes, soyez très prudent !

Jusqu'à présent, nous avons vérifié la structure factorielle obtenue dans l'AFE à l'aide d'un deuxième échantillon et de l'AFC. Malheureusement, nous n'avons pas trouvé que la structure factorielle de l'AEF avait été confirmée dans l'AFC, ce qui nous ramène à la case départ pour ce qui est de l'élaboration de cette échelle de personnalité.

Même si nous avions pu modifier l'AFC à l'aide d'indices de modification, il n'y avait pas de bonnes raisons (qui me viennent à l'esprit) d'inclure d'autres saturations factorielles ou covariances résiduelles. Cependant, il y a parfois une bonne raison de permettre aux résidus de covarier (ou de corréler), et un bon exemple en est donné dans la section suivante sur l'**AFC multitraits multiméthodes** (Multi-Trait Multi-Method (MTMM) CFA). Avant de faire cela, expliquons comment communiquer les résultats d'une AFC.

Compte-rendu d'une AFC

Il n'existe pas de méthode officielle normalisée pour faire un compte-rendu d'une AFC, et les exemples varient selon la discipline et le chercheur. Cela dit, il y a des éléments d'information assez standard à inclure dans votre rapport :

1. Une justification théorique et empirique du modèle hypothétique.
2. Une description complète de la façon dont le modèle a été spécifié (c.-à-d. les variables indicatrices pour chaque facteur latent, les covariances entre les variables latentes et toute corrélation entre les termes d'erreur). Il serait bon d'inclure un diagramme des relations, comme celui de la Figure 15-21.
3. Une description de l'échantillon (p. ex. renseignements démographiques, taille de l'échantillon, méthode d'échantillonnage).
4. Une description du type de données utilisées (p. ex. nominales, continues) et des statistiques descriptives.
5. Tests des hypothèses et de la méthode d'estimation utilisées.

6. Une description des données manquantes et la façon dont elles ont été traitées.
7. Le logiciel et la version utilisés pour adapter le modèle.
8. Les mesures et les critères utilisés pour juger de l'adéquation du modèle.
9. Toute modification apportée au modèle original en fonction des indices d'ajustement ou de modification du modèle.
10. Toutes les estimations de paramètres (c.-à-d. les saturations, les variances d'erreur, les (co)variances latentes et leurs erreurs types, probablement dans un tableau).

Multi-Trait Multi-Méthode Multi-Method CFA

Dans cette section, nous allons examiner comment différentes techniques de mesure ou questions peuvent constituer une source importante de variabilité des données, appelée **méthode variance**. Pour ce faire, nous utiliserons un autre ensemble de données psychologiques, celui qui contient des données sur le « style attributif ».

Avec le questionnaire de style attributif, (Attributional Style Questionnaire ou ASQ), Hewitt, Foxcroft et MacDonald (2004) ont collecté des données sur le bien-être psychologique des jeunes au Royaume-Uni et en Nouvelle-Zélande. Ils ont mesuré le style attributif d'événements négatifs, c'est-à-dire la façon dont les gens expliquent habituellement la cause des mauvaises choses qui leur arrivent (Peterson and Seligman 1984). Le questionnaire sur le style attributif (ASQ) mesure trois aspects du style attributif :

- L'internalité est la mesure de la croyance d'une personne que la cause d'un mauvais événement est due à ses propres actions.
- La stabilité mesure la croyance d'une personne que la cause d'un mauvais événement est stable dans le temps.
- La globalité renvoie à la mesure de la croyance d'une personne que la cause d'un mauvais événement dans un domaine affectera d'autres domaines de sa vie.

Il y a six scénarios hypothétiques et pour chaque scénario, les répondants répondent à une question visant à déterminer (a) l'internalité, (b) la stabilité et (c) la globalité. Il y a donc 6 x 3 = 18 items au total. Voir la Figure 15-26 pour plus de détails.

1. YOU HAVE BEEN LOOKING FOR A JOB UNSUCCESSFULLY FOR SOME TIME

(a) Is the cause of your unsuccessful job search due to something about you or something about other people or circumstances?

Totally due to other people or circumstances **1 2 3 4 5 6 7** **Totally due to me**

(b) In the future, when looking for a job, will this cause again be present?

Will never again be present **1 2 3 4 5 6 7** **Will always be present**

(c) Is the cause something that just influences looking for a job, or does it also influence other areas of your life?

Influences just this particular situation **1 2 3 4 5 6 7** **Influences all situations in my life**

Other hypothetical scenarios, each answered with (a), (b) and (c) in the same sort of way

2. A FRIEND COMES TO YOU WITH A PROBLEM AND YOU DON'T TRY TO HELP
3. YOU GIVE AN IMPORTANT TALK IN FRONT OF A GROUP AND THE AUDIENCE REACTS NEGATIVELY
4. YOU MEET A FRIEND WHO IS HOSTILE TOWARDS YOU
5. YOU CAN'T GET ALL THE WORK DONE THAT OTHERS EXPECT OF YOU
6. YOU GO OUT ON A DATE AND IT GOES BADLY

Figure 15-26 : Questionnaire sur le style attributif (QSA) pour les événements négatifs

Les chercheurs ont cherché à vérifier leurs données pour voir s'il y a des facteurs latents sous-jacents qui sont raisonnablement bien mesurés par les 18 variables observées dans le QSA.

Ils ont d'abord, essayé l'AFE avec ces 18 variables (non montrées), mais peu importe la façon dont elles sont extraites ou en rotation, ils n'ont pas trouvé de bonne solution factorielle. Leur tentative d'identifier les facteurs latents sous-jacents dans le questionnaire sur le style attributif (QSA) s'est avérée infructueuse.

Si vous obtenez de tels résultats, soit votre théorie est erronée (il n'y a pas de structure sous-jacente de facteurs latents pour le style attributif, ce qui est possible), soit l'échantillon n'est pas pertinent (ce qui est peu probable étant donné la taille et les caractéristiques de cet échantillon de jeunes adultes du Royaume-Uni et de Nouvelle-Zélande), soit l'analyse ne constitue pas le bon outil pour cet emploi. Nous allons examiner cette troisième possibilité.

Rappelons qu'il y avait trois dimensions mesurées dans l'ASQ : Internalité, Stabilité et Globalité, chacune mesurée par six questions, comme le montre la Figure 15-27.

Internality	Stability	Globality
Q1a	Q1b	Q1c
Q2a	Q2b	Q2c
Q3a	Q3b	Q3c
Q4a	Q4b	Q4c
Q5a	Q5b	Q5c
Q6a	Q6b	Q6c

Figure 15-27 : Six questions sur l'ASQ pour chacune des dimensions Internalité, Stabilité et Globalité

Que se passerait-il si, au lieu de faire une analyse où nous voyons comment les données sont regroupées de façon exploratoire, nous imposions plutôt une structure, comme celle de la Figure 15-27, aux données et voyions dans quelle mesure les données correspondent à notre structure prédéfinie. En ce sens, nous entreprenons une analyse de confirmatoire, pour voir dans quelle mesure un modèle préétabli est confirmé par les données observées.

Une simple analyse factorielle confirmatoire (AFC) de l'ASQ indiquerait donc trois facteurs latents, comme le montrent les colonnes de la Figure 15-27, chacun mesuré par six variables observées.

Nous pourrions les représenter comme dans le diagramme de la Figure 15-28, qui montre que chaque variable est une mesure d'un facteur latent sous-jacent. Par exemple, INT1 est prédit par le facteur latent sous-jacent Internalité. Et comme INT1 n'est pas une mesure parfaite du facteur d'internalité, il y a un terme d'erreur, e1, qui lui est associé. En d'autres termes, e1 représente la variance dans INT1 qui n'est pas prise en compte par le facteur d'internalité. C'est ce qu'on appelle parfois une « erreur de mesure ».

L'étape suivante consiste à déterminer si les facteurs latents peuvent être corrélés dans notre modèle. Comme nous l'avons mentionné précédemment, dans les sciences psychologiques et comportementales, les constructions sont souvent liées les unes aux autres, et nous pensons aussi que l'internalité, la stabilité et la globalité peuvent être corrélées les unes aux autres, de sorte que dans notre modèle nous devrions permettre à ces facteurs latents de varier conjointement, comme le montre la Figure 15-29.

Parallèlement, nous devrions nous demander s'il existe une raison valable et systématique pour que certains termes d'erreur soient corrélés les uns aux autres. En repensant aux questions de l'ASQ, on constate qu'il y a trois sous-questions différentes (a, b et c) pour chaque question principale (1-6). Q1 était relative à la recherche d'emploi infructueuse et il est plausible que cette question ait des particularités artefactuels ou méthodologiques qui la distinguent des autres questions (2-5) et qui aurait quelque chose à voir avec la difficulté à trouver du travail. De même, Q2 était relative à l'aide refusée à un ami qui avait un

problème, et il se peut qu'il n'y ait pas de problèmes artéfactuels ou méthodologiques distinctifs liés au fait de ne pas aider un ami qui ne seraient pas présent dans les autres questions (1 et 3-5).

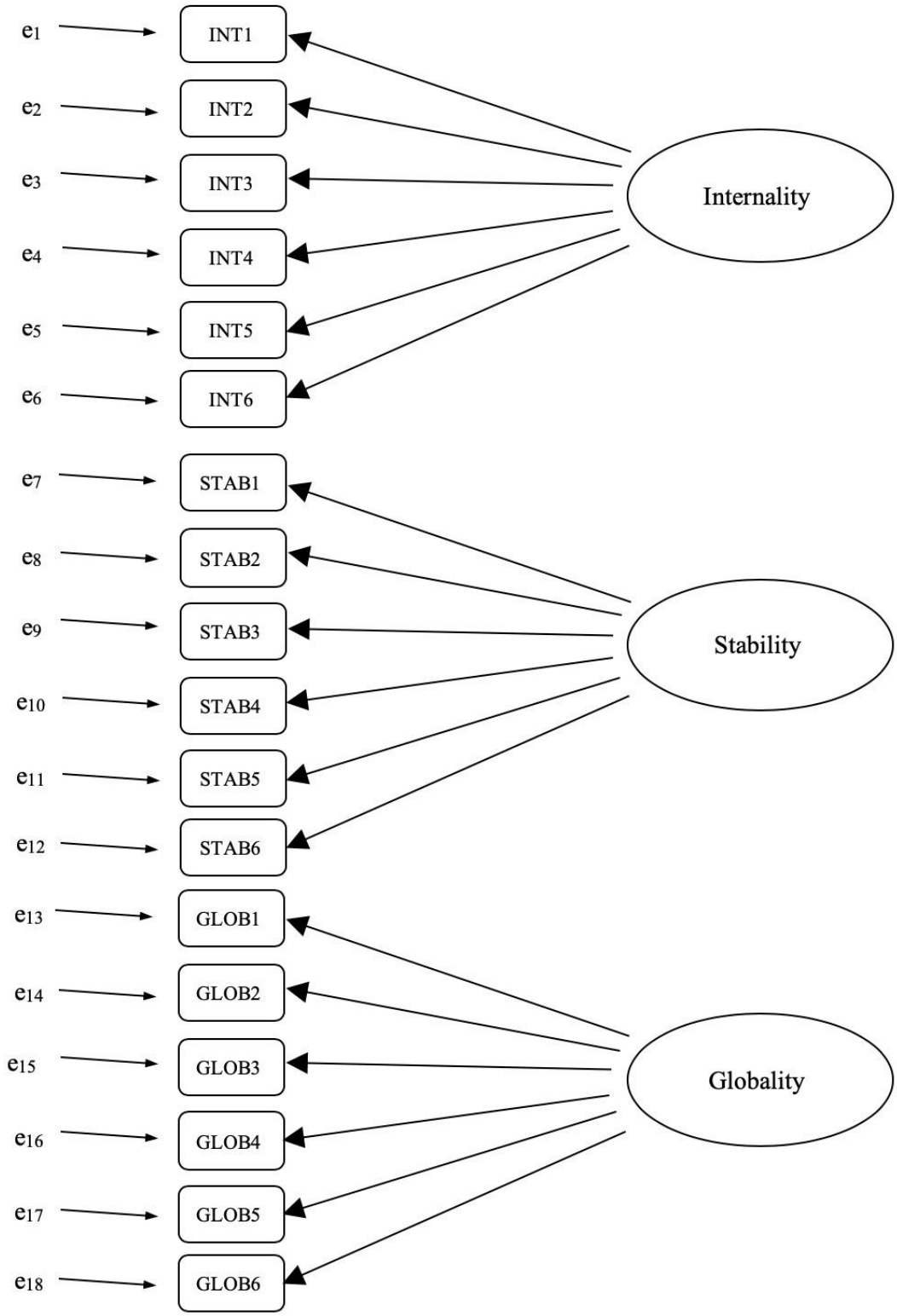


Figure 15-28 : Pré-spécification initiale de la structure factorielle latente pour l'ASQ

Ainsi, en plus de facteurs multiples, nous avons également de multiples caractéristiques méthodologiques dans l'ASQ, où chacune des questions 1 à 6 a une « méthode » légèrement différente, mais chaque « méthode » est partagée entre les sous-questions a, b et c. Afin d'intégrer ces différentes caractéristiques méthodologiques dans le modèle, nous pouvons spécifier que certains termes d'erreur sont corrélés les uns aux autres. Par exemple, les erreurs associées à INT1, STAB1 et GLOB1 devraient être corrélées entre elles pour refléter la variance méthodologique distincte et partagée de Q1a, Q1b et Q1c. Si l'on examine la Figure 15-27, cela signifie qu'en plus des facteurs latents représentés par les colonnes, nous aurons corrélé les erreurs de mesure pour les variables de chaque rangée du tableau.

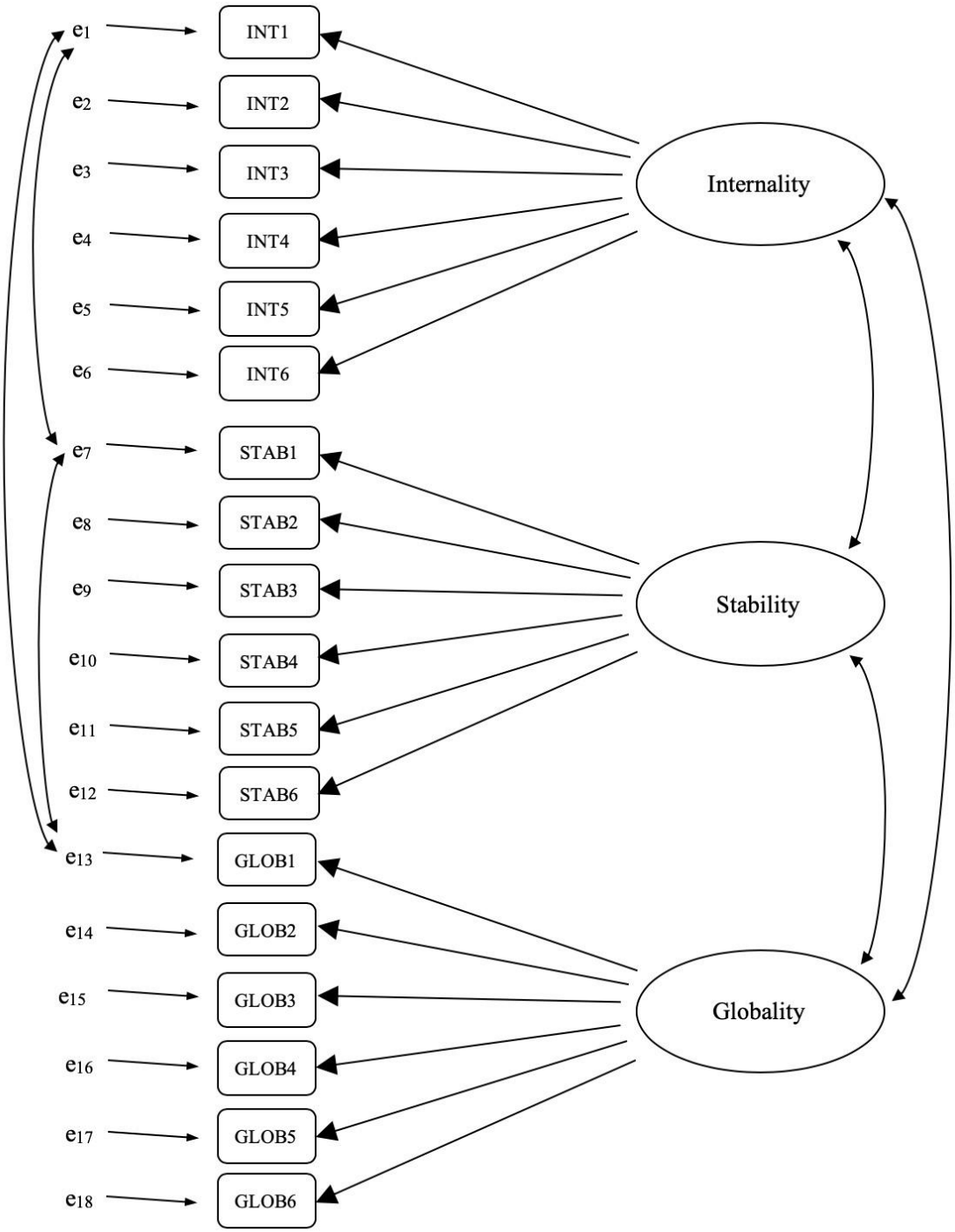


Figure 15-29 : Pré-spécification finale de la structure des facteurs latents pour l'ASQ, y compris les corrélations des facteurs latents et les corrélations des termes d'erreur pour les variables observées INT1, STAB1 et GLOB1, dans un modèle d'AFC MTMM. Par souci de clarté, les autres corrélations de termes d'erreur préétablies ne sont pas indiquées.

Bien qu'un modèle d'AFC de base comme celui illustré à la Figure 15-28 puisse être comparé à nos données observées nous avons en fait mis au point un modèle plus

sophistiqué, comme le montre le diagramme de la Figure 15-29. Ce modèle d'AFC plus sophistiqué est connu sous le nom de modèle **Multi-Trait Multi-Method (MTMM),** et c'est celui que nous allons tester à Jamovi.

Faire une AFC MTMM avec Jamovi

Ouvrez le fichier ASQ.csv et vérifiez que les 18 variables (six variables « Internalité », six variables « Stabilité » et six variables « Globalité ») sont codées comme continue.

Pour effectuer l'AFC MTMM CFA dans Jamovi :

- Sélectionnez Factor - Confirmatory Factor Analysis dans la barre de boutons Jamovi principale pour ouvrir la fenêtre CFA analysis (Figure 15-30).
- Sélectionnez les 6 variables INT et transférez-les dans la case « Factors » et donnez ensuite le label « Internalité ».
- Créez un nouveau facteur dans la case 'Factors' et nommez-le « Stabilité ». Sélectionnez les 6 variables STAB et les transférer dans la case « Facteurs » sous le titre « Stabilité ».
- Créez un autre nouveau Facteur dans la case « Facteurs » et nommez-le « Globalité ». Sélectionnez les 6 variables GLOB et transférez-les dans la case « Facteurs' » sous le titre « Globalité ».
- Ouvrez les options « Residual Covariances », et pour chacune de nos corrélations prédéfinies, déplacez les variables associées dans la case « Residual Covariances » à droite. Par exemple, sélectionnez INT1 et STAB1, puis cliquez sur la flèche pour les déplacer. Faites de même pour INT1 et GLOB1, pour STAB1 et GLOB1, pour INT2 et STAB2, pour INT2 et GLOB2, pour STAB2 et GLOB2, pour INT3 et STAB3, etc.
- Cochez d'autres options appropriées, les valeurs par défaut sont correctes pour ce premier travail, mais vous pouvez cocher l'option « Path diagram » sous « Plots » pour que Jamovi dessine un diagramme (assez) similaire à notre Figure 15-29, mais incluant toutes les corrélations des termes d'erreur que nous avons ajouté ci-dessus.

Une fois que nous avons réalisé l'analyse, nous pouvons porter notre attention sur la fenêtre des résultats de Jamovi et voir ce qu'il en est. La première chose à examiner est « l'ajustement du modèle », car cela nous indique dans quelle mesure notre modèle correspond bien aux données observées. Dans notre modèle, seules les covariances pré-spécifiées sont estimées, tout le reste est mis à zéro, donc l'ajustement du modèle teste à la fois si les paramètres « libres » pré-spécifiés ne sont pas nuls, et inversement si les autres relations dans les données - celles que nous n'avons pas spécifiées dans le modèle - peuvent être considérées à zéro.

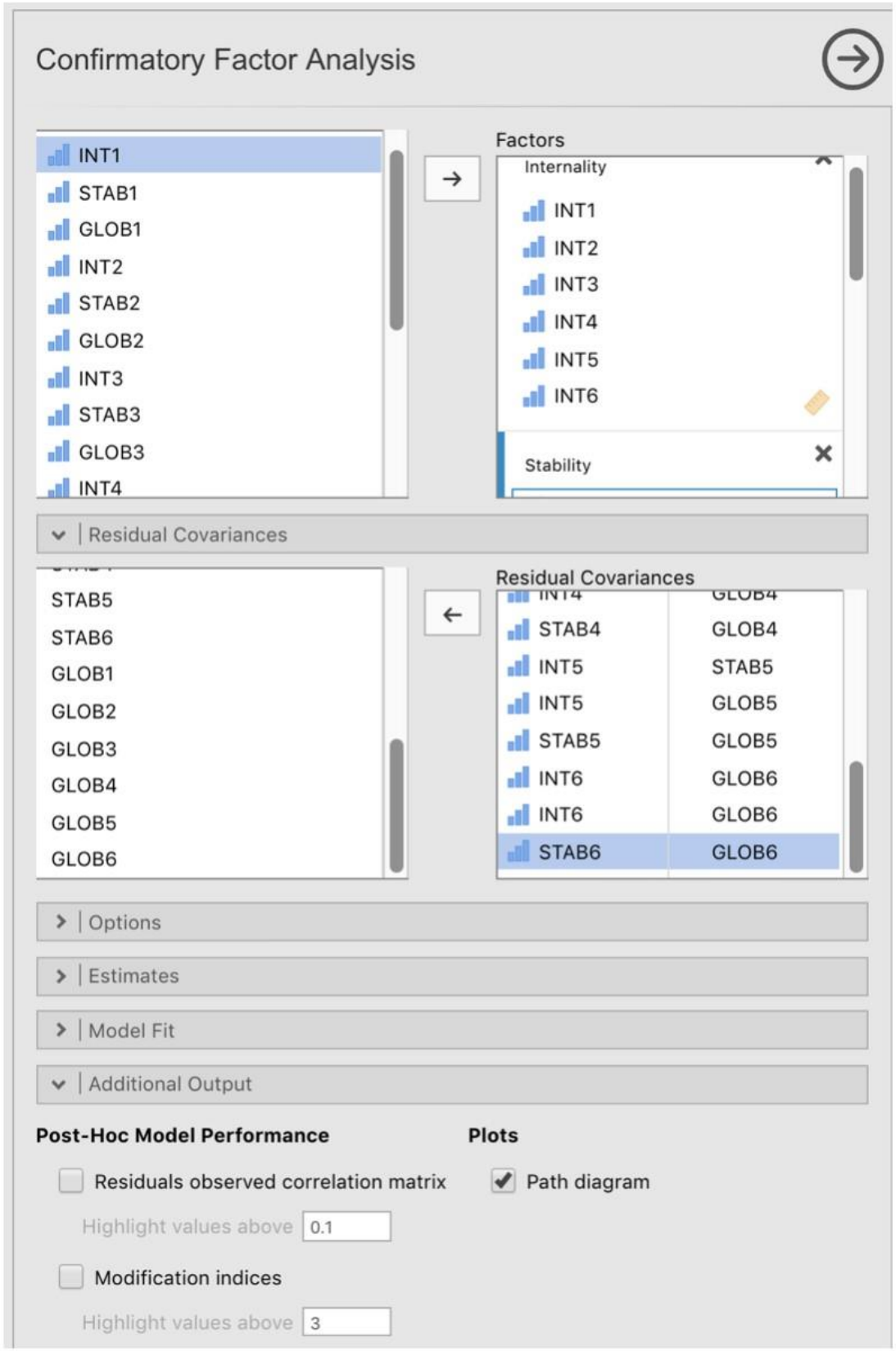

Confirmatory Factor Analysis
INT1
STAB1
GLOB1
INT2
STAB2
GLOB2
INT3
STAB3
GLOB3
INT4
Factors
Internality
INT1
INT2
INT3
INT4
INT5
INT6
Stability
Residual Covariances
STAB5
STAB6
GLOB1
GLOB2
GLOB3
GLOB4
GLOB5
GLOB6
Residual Covariances
STAB4 GLOB4
INT5 STAB5
INT5 GLOB5
STAB5 GLOB5
INT6 GLOB6
INT6 GLOB6
STAB6 GLOB6
Options
Estimates
Model Fit
Additional Output
Post-Hoc Model Performance
Residuals observed correlation matrix
Highlight values above 0.1
Modification indices
Highlight values above 3
Plots
Path diagram

Figure 15-30 : La fenêtre d'analyse CFA Jamovi

Comme pour l'AFC, Il existe plusieurs façons d'évaluer l'adéquation du modèle. La première est une statistique du chi carré, qui, si elle est petite, indique que le modèle est bien adapté aux données. Cependant, la statistique du chi carré utilisée pour évaluer l'ajustement du modèle est très sensible à la taille de l'échantillon, ce qui signifie qu'avec un grand échantillon (plus de 300-400 cas), un ajustement suffisamment bon entre le modèle et les données produit presque toujours une valeur importante et significative du chi carré.

Nous avons donc besoin d'autres moyens d'évaluer l'adéquation du modèle. Dans Jamovi plusieurs sont fournis par défaut. Il s'agit de l'indice d'ajustement comparatif (CFI), de l'indice d'ajustement Tucker (TFI) et de l'erreur quadratique moyenne quadratique approximative (RMSEA) ainsi que de l'intervalle de confiance à 90 % pour la RMSEA. Comme nous l'avons mentionné précédemment, certaines règles empiriques utiles sont qu'un ajustement satisfaisant est indiqué par CFI > 0,9, TFI > 0,9, et RMSEA compris entre 0,05 à 0,08. Un bon ajustement est CFI > 0.95, TFI ą 0.95, et RMSEA et CI supérieur pour RMSEA < 0.05.

Ainsi, en regardant la Figure 15-31, on constate que la valeur du chi carré est très significative, ce qui n'est pas surprenant étant donné la grande taille de l'échantillon (N=2748). La CFI est de 0,98 et l'indice TFI est également de 0,98, ce qui indique un très bon ajustement. La RMSEA est de 0,02 avec un intervalle de confiance de 90 % de 0,02 à 0,02 - est assez étroit !

Dans l'ensemble, je pense que nous pouvons être satisfaits que notre modèle pré-spécifié correspond très bien aux données observées, ce qui appuie notre modèle MTMM pour l'ASQ.

Model Fit

Test for Exact Fit

χ^2	df	p
243.97	114	<.00001

Fit Measures

			RMSEA 90% CI	
CFI	TLI	RMSEA	Lower	Upper
0.98	0.98	0.02	0.02	0.02

Figure 15-31 : Résultats de l'ajustement de notre modèle d'AFC MTMM dans Jamovi

Nous pouvons maintenant examiner les saturations factorielles et les estimations de la covariance des facteurs, comme dans la Figure 15-32.

Souvent, les estimations normalisées sont plus faciles à interpréter, et elles peuvent être précisées dans l'option « Estimates «. Ces tableaux peuvent utilement être incorporés dans un rapport écrit ou un article scientifique.

Vous pouvez voir à la Figure 15-32 que toutes nos saturations factorielles et covariances factorielles préétablies sont significativement différentes de zéro. En d'autres termes, ils semblent tous apporter une contribution utile au modèle.

Nous avons eu beaucoup de chance avec cette analyse, en obtenant un très bon ajustement lors de notre première tentative. C'est assez inhabituel, et souvent, dans l'AFC, des ajustements post hoc supplémentaires sont apportés au modèle pour en améliorer l'ajustement. Une façon d'y parvenir est d'utiliser des « indices de modification » (MI), en les demandant dans « Additionnal Outputs » dans Jamovi.

Ce que nous recherchons, c'est la valeur la plus élevée de l'indice de modification (MI). Nous jugerions alors s'il est judicieux d'ajouter ce terme supplémentaire dans le modèle, en utilisant une analyse *post-hoc*. Par exemple, nous pouvons voir à la Figure 15-33 que le MI le plus élevé pour les charges factorielles qui ne sont pas déjà dans le modèle est une valeur de 24,52 pour la charge de INT6 sur le facteur latent Globalité. Cela indique que si nous ajoutons ce chemin dans le modèle, la valeur du khi-carré diminuera d'environ 25. Mais dans notre modèle, l'ajout de cette relation n'a pas vraiment de sens théorique ou méthodologique, et nous n'inclurons donc pas cette relation dans un modèle révisé.

De même, lorsque nous examinons les MI des termes résiduels (Figure 15-34), le MI le plus élevé qui permet aux erreurs entre INT1 et INT3 de varier (c'est-à-dire d'être incluses dans le modèle) est de 13,48 Ce n'est pas un MI élevé, il n'y a donc pas de justification raisonnable pour inclure ce paramètre dans le modèle, et nous avons déjà un bon ajustement, donc encore une fois notre réponse est de ne pas faire de modification.

Si vous ajoutez de nouveaux paramètres à un modèle à l'aide d'un MI, vérifiez toujours les tables de MI après chaque nouvel ajout (ou exclusion - un MI peut également suggérer des paramètres à supprimer d'un modèle pour améliorer l'ajustement du modèle), car les indicateurs sont actualisés chaque fois.

Analyse de la fiabilité de la cohérence interne

Après avoir suivi le processus initial d'élaboration de l'échelle à l'aide de l'AFE et de l'AFC, vous devriez avoir atteint un stade où l'échelle résiste assez bien à l'AFC avec différents échantillons. Vous pourriez également être intéressé à cette étape de voir comment les facteurs sont mesurés à l'aide d'une échelle qui combine les variables observées.

En psychométrie, nous utilisons l'analyse de fiabilité pour fournir de l'information sur l'uniformité avec laquelle une échelle mesure une construction psychologique (voir la section 2.3). La **cohérence interne** est ce qui nous intéresse ici, c'est-à-dire la cohérence

entre tous les éléments qui composent une échelle de mesure. Ainsi, si nous avons V1, V2, V3, V4 et V5 comme variables d'élément observées, nous pouvons calculer une statistique qui nous indique dans quelle mesure ces éléments sont cohérents sur le plan interne dans la mesure du concept sous-jacent.

Une statistique populaire utilisée pour vérifier la cohérence interne d'une échelle est l'**alpha de Cronbach** (Cronbach 1951). L'alpha de Cronbach est une mesure d'équivalence (c.-à-d. teste si différents ensembles de mesures donnent les mêmes résultats). L'équivalence est testée en divisant les éléments de l'échelle en deux groupes (a « split-half ») et en vérifiant si l'analyse des deux parties donne des résultats comparables.

Confirmatory Factor Analysis

Factor Loadings

Factor	Indicator	Estimate	SE	Z	p	Stand. Estimate
Internality	INT1	0.55	0.05	12.28	<.00001	0.34
	INT2	0.50	0.05	10.52	<.00001	0.28
	INT3	0.61	0.04	13.95	<.00001	0.38
	INT4	0.64	0.05	13.45	<.00001	0.36
	INT5	0.54	0.04	12.52	<.00001	0.33
	INT6	0.66	0.04	16.50	<.00001	0.45
Stability	STAB1	0.53	0.04	14.97	<.00001	0.35
	STAB2	0.48	0.03	14.54	<.00001	0.34
	STAB3	0.69	0.03	20.20	<.00001	0.46
	STAB4	0.65	0.03	20.72	<.00001	0.47
	STAB5	0.67	0.03	21.47	<.00001	0.49
	STAB6	0.66	0.03	22.17	<.00001	0.51
Globality	GLOB1	0.71	0.04	18.16	<.00001	0.40
	GLOB2	0.73	0.04	20.32	<.00001	0.44
	GLOB3	0.93	0.04	25.10	<.00001	0.54
	GLOB4	0.83	0.03	24.99	<.00001	0.53
	GLOB5	0.76	0.03	22.71	<.00001	0.48
	GLOB6	0.96	0.03	27.66	<.00001	0.59

Factor Estimates

Factor Covariances

		Estimate	SE	Z	p	Stand. Estimate
Internality	Internality	1.00 [a]				
	Stability	0.52	0.03	17.10	<.00001	0.52
	Globality	0.45	0.03	14.96	<.00001	0.45
Stability	Stability	1.00 [a]				
	Globality	0.70	0.02	35.47	<.00001	0.70
Globality	Globality	1.00 [a]				

[a] fixed parameter

Figure 15-32 : les tableaux des saturations et des covariances des facteur de l'AFC dans Jamovi avec notre modèle MTMM.

Bien sûr, il existe de nombreuses façons de fractionner un ensemble d'items, mais si tous les fractionnements possibles sont effectués, il est possible de produire une statistique qui reflète le modèle global des coefficients. L'alpha de Cronbach est une telle statistique : une fonction de tous les coefficients de demi-échelle pour une échelle. En même temps, vous pouvez également doubement vérifier si la suppression d'un élément de l'échelle améliore ma fiabilité.

Factor Loadings – Modification Indices

	Internality	Stability	Globality
INT1		1.41	0.16
INT2		4.86e–4	0.03
INT3		5.47	6.61
INT4		1.05e–6	2.66
INT5		0.24	1.68
INT6		12.60	24.52
STAB1	1.48		1.62
STAB2	3.06		0.12
STAB3	1.31		1.08
STAB4	3.10		2.88
STAB5	0.01		8.40
STAB6	1.31		1.05
GLOB1	0.52	0.23	
GLOB2	0.60	0.53	
GLOB3	2.85	0.04	
GLOB4	4.33	12.72	
GLOB5	7.23	0.12	
GLOB6	3.44	6.03	

Figure 15-33 : Indices de modification des charges des facteurs CFA Jamovi

Si un ensemble d'éléments mesurant une construction (p. ex. une échelle d'extraction) a un alpha de 0,80, alors la proportion de variance d'erreur dans l'échelle est de 0,20. En d'autres termes, une échelle avec un alpha de 0,80 comprend environ 20 % d'erreur.

MAIS, (et c'est un GRAND « MAIS »), l'alpha de Cronbach n'est pas une mesure d'unicité (c'est-à-dire un indicateur qu'une échelle mesure un facteur ou une construction unitaire plutôt que plusieurs constructions connexes). Les échelles multidimensionnelles entraîneront une sous-estimation de l'alpha si elles ne sont pas évaluées séparément pour chaque dimension, mais des valeurs élevées pour l'alpha ne sont pas nécessairement des indicateurs d'unicité. Ainsi, un alpha de 0,80 ne signifie pas que 80 % d'une seule construction sous-jacente est prise en compte. Il se peut que les 80 % proviennent de plus d'une construction sous-jacente. C'est pourquoi il est utile de faire avant une AFE et une AFC sont utiles.

INT6	STAB1	STAB2	STAB3	STAB4	STAB5	STAB6	GLOB1	GLOB2	GLOB3	GLOB4	GLOB5	GLOB6
11.83		0.02	2.16	0.18	0.05	0.11		0.13	1.04	1.51	1.76	0.02
1.32	4.33		0.02	0.15	1.73	0.21	0.09		0.13	0.07	2.65	0.84
1.52	0.04	1.67		1.53	0.01	0.26	4.24	0.02		1.27	0.11	0.07
0.30	2.73	3.24	0.46		1.70	0.14	0.05	0.86	0.43		0.89	5.99
2.79	0.74	0.00	1.82	1.86		1.96	0.03	0.22	0.18	0.20		1.73
	0.29	2.92e−4	1.84	0.37	0.01		6.88	0.14	2.75	0.74	1.72	
		5.21	8.32	9.64	2.07	2.54		6.10e−6	0.35	1.34	0.66	1.67
			2.20	11.04	0.04	0.23	3.32		0.06	2.16	0.77	1.88
				7.41	11.37	1.38	4.36	2.45		0.04	10.30	0.62
					0.03	1.99	0.04	0.72	0.78		0.09	0.95
						1.14	0.94	0.11	7.63	4.34		1.64
							3.21	3.49	0.49	0.68	1.34	
								0.02	1.67	1.66	0.16	0.56
									0.88	2.48	1.82	4.37
										3.30	0.94	2.20
											0.41	3.60
												1.49e−6

Figure 15-34 : Indices de modification des covariances résiduelles CFA Jamovi

De plus, une autre caractéristique de l'alpha est qu'il a tendance à être spécifique à l'échantillon : ce n'est pas une caractéristique de l'échelle, mais plutôt une caractéristique de l'échantillon dans lequel l'échelle a été utilisée. Un échantillon biaisé, non représentatif ou de petite taille pourrait produire un coefficient alpha très différent de celui d'un grand échantillon représentatif. L'alpha peut même varier d'un grand échantillon à l'autre. Néanmoins, malgré ces limites, l'alpha de Chronbach est très populaire en psychologie pour l'estimation de la fiabilité de la cohérence interne pour les raisons suivantes. Il est assez facile à calculer, à comprendre et à interpréter, et par conséquent, il peut s'agir d'une vérification initiale utile de la précision de la mesure lorsque vous administrez une échelle avec un échantillon différent, par exemple dans un autre milieu ou une population différents

Une alternative est **Omega de McDonald**, et Jamovi fournit également cette statistique. Alors que l'alpha fait les hypothèses suivantes : (a) aucune corrélation résiduelle, (b) les items ont des saturations identiques, et (c) l'échelle est unitire, l'omega ne les fait pas et est donc une statistique de fiabilité plus robuste. Si ces hypothèses ne sont pas violées, l'alpha et l'oméga seront similaires, mais si elles le sont, alors il faut privilégier l'omega.

Parfois, un seuil pour l'alpha est fourni, ce qui suggère une valeur « suffisante ». Il pourrait s'agir d'alphas de 0,70 ou 0,80 représentant respectivement une fiabilité « acceptable » et « bonne ». Cependant, cela dépend de ce que l'échelle est censée mesurer exactement, de sorte que des seuils comme celui-ci doivent être utilisés avec prudence. Il pourrait être préférable d'indiquer simplement qu'un alpha de 0,70 est associé à une variance d'erreur de 30 % dans une échelle, et qu'un alpha de 0,80 est associé à 20 %.

L'alpha peut-il être trop élevé ? Probablement : si vous obtenez un coefficient alpha supérieur à 0,95, cela indique des corrélations élevées entre les éléments et qu'il peut y avoir trop de spécicités redondantes dans la mesure, avec un risque que la construction mesurée soit peut-être trop restreinte.

Analyse de la fiabilité dans Jamovi

Nous disposons d'un troisième ensemble de données pour entreprendre une analyse de fiabilité : le fichier bfi_sample3.csv. Une fois de plus, vérifiez que les 25 variables de personnalité sont codées comme continues. Pour effectuer des analyses de fiabilité dans Jamovi :

- Sélectionnez Factor - Reliability Analysis dans la barre principale de boutons de Jamovi pour ouvrir la fenêtre d'analyse de fiabilité (Figure 15-35).
- Sélectionnez les 5 variables A et transférez-les dans la boîte « Items ».
- Sous l'option « Reverse Scaled Items », sélectionnez la variable A1 dans la case « Normal Scaled Items » et déplacez-la dans la case « Reverse Scaled Items ».
- Cochez les autres options appropriées, comme dans la Figure 15-35.

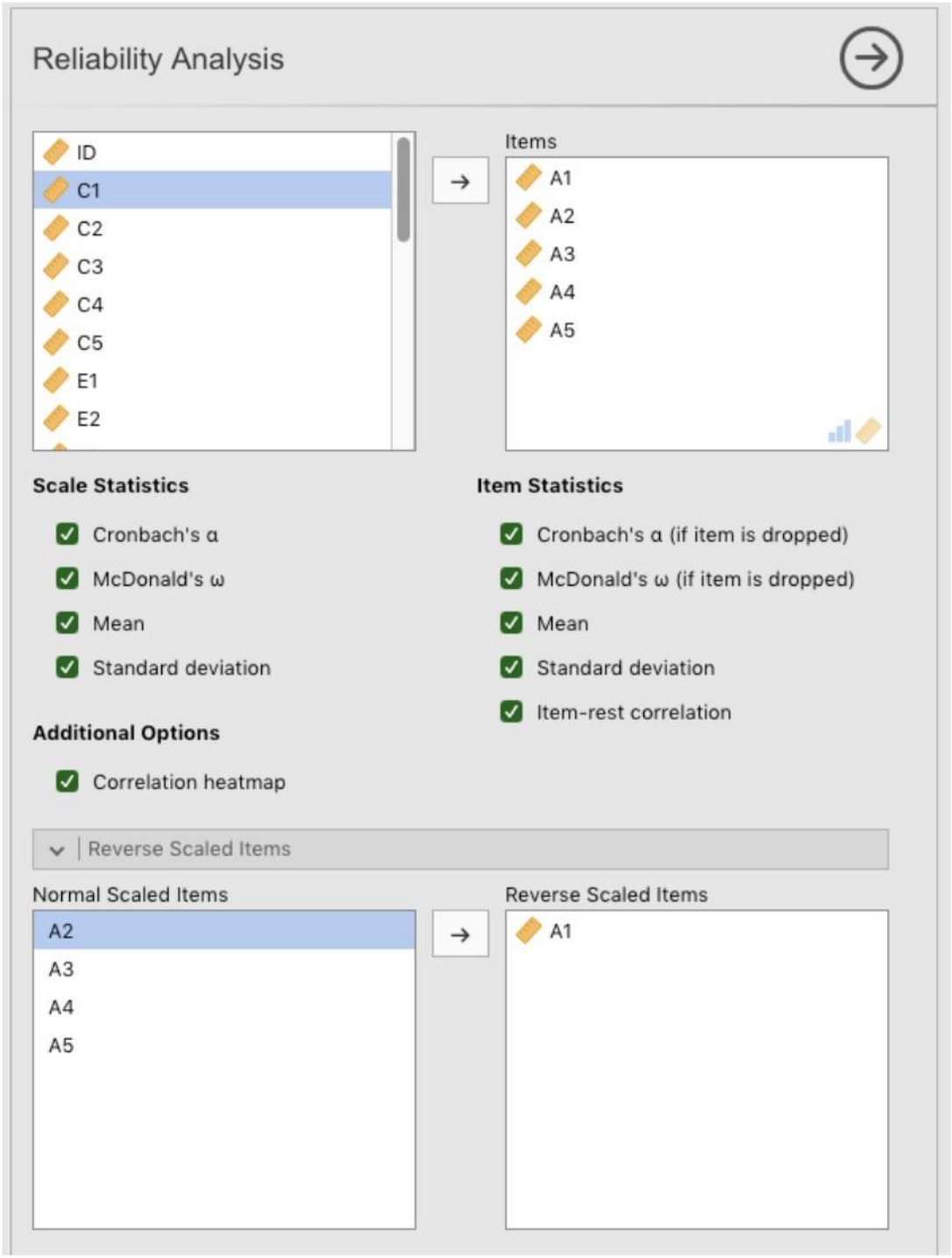


Figure 15-35 : Fenêtre Analyse de fiabilité Jamovi

Une fois fait, regardez la fenêtre des résultats de Jamovi. Vous devriez voir quelque chose comme Figure 15-36. Cela nous indique que le coefficient alpha de Chronbach pour l'échelle de Agreeableness est de 0,70. Cela signifie qu'un peu moins de 30 % de la note de l'échelle Agreeableness représente la variance d'erreur. L'Oméga de McDonald est aussi fourni, et il est de 0,72, ce qui n'est pas très différent de l'alpha.

Nous pouvons également vérifier comment l'alpha ou l'oméga peut être amélioré avec la suppression d'un élément spécifique. Par exemple, l'alpha passerait à 0,72 et l'oméga à 0,74 si nous supprimions le point A1. Il ne s'agit pas d'une augmentation importante, donc cela ne vaut probablement pas la peine d'être fait.

Le processus de calcul et de vérification des statistiques de fiabilité (alpha et oméga) est le même pour toutes les autres échelles (non représentées) : Conscientiousness (alpha = 0,73, oméga = 0,74), Extraversion (alpha = 0,76, oméga = 0,76), Neuroticism (alpha = 0,81, oméga = 0,82) et Openness (alpha = 0,60, oméga = 0,62). Pour l'ouverture, la variance d'erreur dans le score de l'échelle est de 40 %, ce qui est élevé et indique que l'ouverture est beaucoup moins cohérente comme mesure fiable d'un attribut de personnalité que les autres échelles de personnalité.

Reliability Analysis

Scale Reliability Statistics

	mean	sd	Cronbach's α	McDonald's ω
scale	4.650	0.900	0.701	0.720

Item Reliability Statistics

				if item dropped	
	mean	sd	item-rest correlation	Cronbach's α	McDonald's ω
A1 [a]	4.628	1.374	0.287	0.722	0.735
A2	4.676	1.237	0.544	0.619	0.652
A3	4.640	1.301	0.608	0.588	0.607
A4	4.828	1.422	0.404	0.676	0.701
A5	4.480	1.327	0.474	0.645	0.664

[a] reverse scaled item

Figure 15-36 : Résultats de l'analyse de fiabilité de Jamovi pour le facteur Agreeableness

Résumé

- L'analyse factorielle exploratoire (AEF) est une technique statistique permettant d'identifier les facteurs latents sous-jacents d'un ensemble de données. Chaque variable observée est conceptualisée comme représentant dans une certaine mesure le facteur latent, indiqué par une saturation factorielle. Les chercheurs utilisent également l'AEF comme moyen de réduire les données, c'est-à-dire d'identifier les variables observées qui peuvent être combinées en de nouvelles variables pour une analyse ultérieure. (Section 15.1)

- L'analyse en composantes principales (ACP) est une technique de réduction des données qui, à proprement parler, n'identifie pas les facteurs latents sous-jacents. Au

lieu de cela, l'ACP produit simplement une combinaison linéaire de variables observées. (Section 15.2)

- Analyse factorielle de confirmation (AFC). Contrairement à l'AFE, avec l'AFC, vous commencez par une idée - un modèle - de relations entre vos variables observées. Vous testez ensuite votre modèle par rapport aux données observées et évaluez dans quelle mesure il s'adapte au modèle. (Section 15.3)
- Dans l'AFC multitraits multiméthodes (MTMM), la variance du facteur latent et la variance de la méthode sont incluses dans le modèle dans une approche qui est utile lorsqu'il y a différentes approches méthodologiques utilisées et que la variance de la méthode est donc une considération importante (section 15.4).
- Analyse de fiabilité de la cohérence interne. Cette forme d'analyse de fiabilité teste la cohérence d'une échelle pour la mesure d'une mesure consruite (psychologique). (Section 15.5)

Statistiques bayésiennes

Dans nos raisonnements concernant les faits, il y a tous les degrés d'assurance imaginables, de la plus haute certitude à la plus basse espèce de preuve morale. Un homme sage, par conséquent, proportionne sa croyance à l'évidence. - David Hume[142]

Les idées que je vous ai présentées dans ce livre décrivent les statistiques inférentielles du point de vue des fréquentistes. Je ne suis pas le seul à le faire. En fait, presque tous les manuels scolaires donnés aux étudiants de premier cycle en psychologie présentent le point de vue du statisticien fréquentiste comme *la* théorie de la statistique inférentielle, la seule véritable façon de faire les choses. J'ai enseigné de cette façon pour des raisons pratiques. La vision fréquentiste de la statistique a dominé le champ académique de la statistique pendant la majeure partie du XXe siècle, et cette domination est encore plus extrême chez les scientifiques appliqués. C'était et c'est une pratique courante chez les psychologues d'utiliser des méthodes fréquentistes. Parce que les méthodes fréquentistes sont omniprésentes dans les articles scientifiques, chaque étudiant en statistique doit comprendre ces méthodes, sinon il sera incapable de comprendre ce que disent ces articles ! Malheureusement, du moins à mon avis, la pratique actuelle de la psychologie est souvent malavisée et le recours aux méthodes fréquentistes est en partie responsable. Dans ce chapitre, j'explique pourquoi je pense cela et je donne une introduction aux statistiques bayésiennes, une approche qui, à mon avis, est généralement supérieure à l'approche orthodoxe.

Ce chapitre est divisé en deux parties. Dans les sections 16.1 à 16.3, j'explique en quoi consistent les statistiques bayésiennes, les règles mathématiques de base de leur fonctionnement ainsi que les raisons pour lesquelles l'approche bayésienne me semble si

---

142 http://en.wikiquote.org/wiki/David_Hume

utile. Par la suite, je donne un bref aperçu de la façon dont vous pouvez faire des versions bayésiennes de *t-tests* (Section 16.4).

Raisonnement probabiliste des agents rationnels

D'un point de vue bayésien, l'inférence statistique est une question de *révision des croyances*. Je pars d'un ensemble d'hypothèses candidates *h* sur le monde. Je ne sais pas laquelle de ces hypothèses est vraie, mais j'ai des croyances sur les hypothèses qui sont plausibles et celles qui ne le sont pas. Quand j'observe les données, *d*, je dois réviser ces croyances. Si les données sont cohérentes avec une hypothèse, ma croyance en cette hypothèse est renforcée. Si les données ne concordent pas avec l'hypothèse, ma croyance en cette hypothèse s'en trouve affaiblie. A la fin de cette section, je donnerai une description précise du fonctionnement du raisonnement bayésien, mais je veux d'abord travailler sur un exemple simple pour présenter les idées clés. Considérons le problème de raisonnement suivant.

Je prends un parapluie. Croyez-vous qu'il va pleuvoir ?

Dans ce problème, je vous ai présenté un seul élément de données (*d* = Je prends le parapluie), et je vous demande de me dire votre croyance ou votre hypothèse sur la pluie. Vous avez deux alternatives, *h* : soit il pleuvra aujourd'hui, soit il ne pleuvra pas. Comment résoudre ce problème ?

A priori : ce que vous croyiez avant

La première chose que vous devez faire est d'ignorer ce que je vous ai dit au sujet du parapluie, et de noter vos croyances préexistantes sur la pluie. C'est important. Si vous voulez être honnête sur la façon dont vos croyances ont été révisées à la lumière de nouvelles preuves (données) alors vous *devez* dire quelque chose sur ce que vous croyiez avant que ces données n'apparaissent ! Bien, que pourriez-vous croire à propos du fait qu'il va pleuvoir aujourd'hui ? Vous savez probablement que je vis en Australie et qu'une grande partie de l'Australie est chaude et sèche. La ville d'Adélaïde où je vis a un climat méditerranéen, très similaire à celui de la Californie du Sud, de l'Europe du Sud ou de l'Afrique du Nord. J'écris ceci en janvier et vous pouvez donc supposer que c'est le milieu de l'été. En fait, vous avez peut-être décidé de jeter un coup d'œil sur Wikipédia[143] et découvert qu'Adélaïde reçoit en moyenne 4,4 jours de pluie pendant les 31 jours de janvier. Sans rien savoir d'autre, vous pourriez conclure que la probabilité de pluie en janvier à Adélaïde est d'environ 15%, et la probabilité d'une journée sèche est de 85%. Si c'est vraiment ce que vous croyez au sujet de la pluie d'Adélaïde (et maintenant que je vous l'ai dit, je parie que c'*est* vraiment ce que vous croyez) alors ce que j'ai écrit ici est votre **distribution à priori**, écrite *P(h)*:

hypothesis	Degree of belief
Rainy day	0.15

[143] https://en.wikipedia.org/wiki/Climate_of_Adelaide

Dry day 0.85

Probabilités : théories sur les données

Pour résoudre le problème de raisonnement, vous avez besoin d'une théorie sur mon comportement. Quand Dan prend-il un parapluie ? Vous devinerez peut-être que je ne suis pas un idiot[144], et j'essaie de ne prendre des parapluies que les jours de pluie. D'un autre côté, vous savez aussi que j'ai de jeunes enfants, et vous ne seriez pas surpris d'apprendre que j'ai tendance à oublier ce genre de choses. Supposons que les jours de pluie, je me souviens de mon parapluie environ 30% du temps (je suis vraiment nul à ce jeu).

Mais disons que par temps sec, je n'ai que 5% de chances de prendre un parapluie. Vous pourriez écrire une petite table comme celle-ci :

```
<th class="tg-cly1"></th>
<th class="tg-nrix" colspan="2">Data</th>

<td class="tg-cly1">Hypothesis</td>
<td class="tg-nrix">Umbrella</td>
<td class="tg-nrix">No umbrella</td>

<td class="tg-cly1">Rainy day</td>
<td class="tg-nrix">0.30</td>
<td class="tg-nrix">0.70</td>

<td class="tg-cly1">Total</td>
<td class="tg-nrix">0.05</td>
<td class="tg-nrix">0.95</td>
```

Il est important de se rappeler que chaque cellule de ce tableau décrit vos croyances sur les données d qui seront observées, *étant donné* la véracité d'une hypothèse particulière h. Cette « probabilité conditionnelle » est écrite $P(d \mid h)$, que vous pouvez lire comme « la probabilité de d donnée h ». Dans les statistiques bayésiennes, il s'agit de la **probabilité** des données d compte tenu de l'hypothèse h.[145]

---

[144] C'est un acte de foi, je sais, mais il faut s'en accommoder, si vous êtes d'accord

[145] Je déteste soulever cette question, mais certains statisticiens s'opposeraient à ce que j'utilise le mot "probabilité" ici. Le problème, c'est que le mot "probabilité" a une signification très spécifique dans les statistiques fréquentistes, et ce n'est pas tout à fait la même chose que dans les statistiques bayésiennes. D'après ce que je peux dire, les Bayésiens n'avaient pas à l'origine de nom convenu pour la probabilité, et il est donc devenu pratique courante pour les gens d'utiliser la terminologie fréquentiste. Cela n'aurait pas posé de problème si la façon dont les Bayésiens utilisent le mot ne s'avèrait être très différente de celle des fréquentistes. Ce n'est pas l'endroit pour une autre longue leçon d'histoire, mais, pour le dire grossièrement, quand un Bayésien parle « d'une fonction de vraisemblance », il se réfère habituellement à l'une des lignes du tableau. Quand un fréquentiste dit la même chose, il se réfère au tableau lui-même, mais pour aux « une

La probabilité conjointe des données et des hypothèses

À ce stade, tous les éléments sont en place. Après avoir noté les à priori et la probabilité, vous avez toute l'information dont vous avez besoin pour faire le raisonnement bayésien. La question est maintenant de *savoir comment* utiliser cette information. Il s'avère qu'il y a une équation très simple que nous pouvons utiliser ici, mais il est important que vous compreniez pourquoi nous l'utilisons, alors je vais essayer de la construire à partir d'idées plus simples.

Commençons par une des règles de la théorie des probabilités. Je l'ai énuméré dans le Tableau 7-1, mais je n'en ai pas fait grand cas à l'époque et vous l'avez probablement ignoré. La règle en question est celle qui parle de la probabilité que *deux* choses soient vraies. Dans notre exemple, vous voudrez peut-être calculer la probabilité qu'aujourd'hui soit pluvieux (c'est-à-dire que l'hypothèse *h* soit vraie) *et* que je prenne un parapluie (c'est-à-dire que les données *d* soient observées). La **probabilité conjointe** de l'hypothèse et des données est écrite *P(d,h)*, et vous pouvez la calculer en multipliant l'à priori *P(h)* par la probabilité $P(d \mid h)$. Mathématiquement, nous disons que

$$P(d,h) = P(d|h)P(h)$$

Alors, quelle est la probabilité qu'aujourd'hui soit un jour de pluie *et* que je me souvienne de prendre un parapluie ? Comme nous l'avons vu plus haut, l'à priori nous dit que la probabilité d'un jour de pluie est de 15%, et la probabilité que je me souvienne de prendre mon parapluie un jour de pluie est de 30%. Ainsi, la probabilité que ces deux évènements soient vrais est calculée en multipliant les deux valeurs suivantes

$$\begin{aligned} P(\text{rainy, umbrella}) &= P(\text{rainy} \mid \text{umbrella}) \times P(\text{rainy}) \\ &= 0{,}30 \times 0{,}15 \\ &= 0{,}045 \end{aligned}$$

En d'autres termes, avant de savoir ce qui s'est réellement passé, vous pensez qu'il y a une probabilité de 4,5% qu'aujourd'hui soit un jour de pluie et que je me souvienne du parapluie. Cependant, il y a bien sûr *quatre* évènements qui peuvent arriver. Répétons donc l'exercice pour les quatre. Si nous faisons cela, nous obtenons le tableau suivant :

	Umbrella	No umbrella
Rainy	0.0450	0.1050
Dry	0.0425	0.8075

---

fonction de vraisemblance » fait référence presque toujours à une des colonnes. Cette distinction est importante dans certains contextes, mais elle ne l'est pas pour notre propos.

Ce tableau saisit toutes les informations sur les probabilités des des quatre possibilités. Pour avoir une vue d'ensemble, cependant, il est utile d'additionner les totaux des lignes et des colonnes. Cela nous donne cette table :

	Umbrella	No umbrella	Total
Rainy	0.0450	0.1050	0.15
Dry	0.0425	0.8075	0.85
Total	0.0875	0.9125	1

C'est un tableau très utile, il vaut donc la peine de prendre un moment pour réfléchir à ce que tous ces chiffres nous disent. Tout d'abord, notez que les sommes des lignes ne nous disent rien de nouveau du tout. Par exemple, la première rangée nous dit que si nous ignorons toute cette affaire de parapluie, la probabilité qu'aujourd'hui soit un jour de pluie est de 15%. Ce n'est pas surprenant, bien sûr, puisque c'est notre à priori.[146] L'important n'est pas les nombres eux-même. Au contraire, l'important, c'est que cela nous donne une certaine confiance que nos calculs sont raisonnables ! Maintenant, jetez un coup d'oeil aux sommes de la colonne et remarquez qu'elles nous disent quelque chose que nous n'avons pas encore explicitement indiqué. De la même manière que les sommes des lignes nous indiquent la probabilité de pluie, les sommes des colonnes nous indiquent la probabilité que je prenne un parapluie. Plus précisément, la première colonne nous indique qu'en moyenne (c.-à-d. en ignorant si c'est un jour de pluie ou non) la probabilité que je prenne un parapluie est de 8,75 %. Enfin, notons que lorsque nous additionnons les quatre événements logiquement possibles, le total vaut 1, c'est-à-dire que nous avons écrit une distribution de probabilités correcte, définie sur toutes les combinaisons possibles de données et d'hypothèses.

Maintenant et parce que ce tableau est si utile, je veux m'assurer que vous comprenez à quoi correspondent tous les éléments et comment ils sont notés :

	Umbrella	No umbrella	
Rainy	P(umbrella, Rainy)	P(No Umbrella, Rainy)	P(Rainy)
Dry	P(umbrella, Dry)	P(No umbrella, Dry)	P(Dry)
	P(umbrella)	P(No umbrella)	

Enfin, nous pouvons utiliser une notation statistique « correcte ». Dans le problème des jours de pluie, les données correspondent à l'observation que j'ai ou non un parapluie. Nous

---

[146] Pour être clair, l'information « à priori » est une connaissance ou une croyance préexistante, avant que nous recueillions ou utilisions toute donnée pour améliorer cette information.

allons noter d_1 la possibilité que vous m'observiez en prenant un parapluie, et d_2 à la possibilité que vous m'observiez en n'en prenant pas. De même, h_1 est votre hypothèse qu'aujourd'hui il pleut et h_2 est l'hypothèse que ce n'est pas le cas. En utilisant cette notation, le tableau ressemble à ceci :

	d_1	d_2	
h_1	$P(h_1,d_1)$	$P(h_1,d_2)$	$P(h_1)$
h_2	$P(h_2,d_1)$	$P(h_2,d_2)$	$P(h_2)$
	$P(d_1)$	$P(d_2)$	

Révision des croyances à l'aide de la règle de Bayes

Le tableau que nous avons présenté dans la dernière section est un outil très puissant pour résoudre le problème des jours de pluie, parce qu'il considère les quatre possibilités logiques et indique exactement dans quelle mesure vous avez confiance en chacune d'elles avant de recevoir des données. Il est maintenant temps de réfléchir à ce qu'il adviendra de nos croyances lorsque nous recevrons les données. Dans le problème des jours de pluie, on vous dit que je prends vraiment un parapluie. C'est un événement assez surprenant. Selon notre tableau, la probabilité que je porte un parapluie n'est que de 8,75 %. Mais est-ce logique ? Un type qui prend un parapluie un jour d'été dans une ville chaude et sèche est assez inhabituel, et vous ne vous y attendiez pas du tout. Néanmoins, les données vous disent que c'est vrai. Aussi improbable que vous pensiez que soit, vous devez maintenant ajuster vos croyances pour tenir compte du fait que vous *savez* maintenant que j'ai un parapluie.[147] Pour tenir compte de ces nouvelles connaissances, notre tableau *révisé* doit comporter les chiffres suivants :

	Umbrella	No umbrella
Rainy		0
Dry		0
Total	1	0

En d'autres termes, les faits ont éliminé toute possibilité de «ne pas prendre de parapluie », nous devons donc mettre des zéros dans toutes les cases du tableau qui impliquent que je ne prends pas de parapluie. Maintenant que vous savez que je prends un parapluie, la somme de la colonne de gauche doit être 1 pour décrire correctement le fait que *P(umbrella)*=1.

Quels sont les deux nombres à mettre dans les cellules vides ? Encore une fois, ne nous préoccupons pas des mathématiques, mais pensons plutôt à nos intuitions. Lorsque nous avons rédigé notre tableau la première fois, il s'est avéré que ces deux cellules avaient des

---

[147] Si nous étions un peu plus sophistiqués, nous pourrions étendre l'exemple pour tenir compte de la possibilité que je mente au sujet du parapluie. Mais gardons les choses simples, d'accord ?

nombres presque identiques. Nous avons calculé que la probabilité commune de « pluie et parapluie » était de 4,5%, et la probabilité commune de « temps sec et parapluie » de 4,25%. En d'autres termes, avant de vous dire que je porte en fait un parapluie, vous auriez dit que ces deux événements étaient de probabilité presque identiques. Mais remarquez que ces *deux* possibilités sont cohérentes avec le fait que je prends un parapluie. Du point de vue de ces deux possibilités, très peu de choses ont changé. J'espère que vous conviendrez qu'il est *toujours* vrai que ces deux possibilités sont tout aussi plausibles. Nous nous attendons donc à voir dans notre tableau final quelques chiffres qui préservent le fait que « pluie et parapluie » est *légèrement* plus plausible que « sec et parapluie », tout en assurant que les chiffres dans le tableau s'additionnent. Quelque chose comme ça, peut-être ?

	Umbrella	No Umbrella
Rainy	0.514	0
Dry	0.486	0
Dry day	1	0

Ce que ce tableau vous dit, c'est qu'après avoir appris que je prends un parapluie, vous croyez qu'il y a 51,4 p.100 de chances qu'aujourd'hui soit un jour de pluie, et 48,6 p.100 de chances qu'il ne le soit pas. C'est la réponse à notre problème ! La **probabilité à posteriori** de pluie *P(h|d)* étant donné que je prends un parapluie est de 51,4%.

Comment ai-je calculé ces chiffres ? Vous pouvez probablement le deviner. Pour calculer qu'il y avait une probabilité de 0,514 de « pluie », je n'ai fait que prendre la probabilité de 0,045 de « pluie et parapluie » et la diviser par la probabilité de 0,0875 de « parapluie ». Il en résulte un tableau qui répond à notre besoin d'avoir la somme total de 1, et à notre besoin de ne pas interférer avec la plausibilité relative des deux événements qui sont en fait compatibles avec les données. Pour dire la même chose en utilisant un jargon statistique fantaisiste, ce que j'ai fait ici est de diviser la probabilité commune de l'hypothèse et des données *P(d,h)* par la **probabilité marginale** des données *P(d)*, et c'est ce qui nous donne la probabilité à posteriori de l'hypothèse *étant donné* les données qui ont été observées que nous pouvons écrire sous forme d'équation[148]

$$P(h|d) = \frac{P(d,h)}{P(d)}$$

Cependant, vous vous rappelez ce que j'ai dit au début de la dernière section que la probabilité commune *P(d,h)* est calculée en multipliant le *P(h)* précédent par la probabilité $P(d \mid h)$. Dans la vraie vie, les choses que nous savons vraiment notées sont les a priori et

---

[148] Vous remarquerez peut-être que cette équation est en fait une reformulation de la même règle de base que celle que j'ai énoncée au début de la dernière section. Si vous multipliez les deux côtés de l'équation par *P(d)*, alors vous obtenez $P(d)P(h \mid d) = P(d,h)$, qui est la règle de calcul des probabilités conjointes. Je n'introduis donc pas de « nouvelles » règles ici, j'utilise simplement la même règle d'une manière différente.

les probabilités, alors substituons-les dans l'équation. Cela nous donne la formule suivante pour la probabilité à posteriori

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

Et cette formule, les amis, est connue sous le nom de **règle de Bayes**. Elle décrit comment un apprenant commence avec des croyances à piorir sur la plausibilité de différentes hypothèses, et vous dit comment ces croyances devraient être révisées face aux données. Dans le paradigme bayésien, toute inférence statistique découle de cette règle simple.

Tests d'hypothèses bayésiennes

Au chapitre 9, j'ai décrit l'approche orthodoxe de la vérification des hypothèses. Il a fallu un chapitre entier pour le décrire, parce que la vérification d'hypothèses nulles est une machinerie très élaborée que les gens trouvent très difficile à comprendre. En revanche, l'approche bayésienne de la vérification des hypothèses est incroyablement simple. Prenons des paramètres très proches du scénario orthodoxe. Il y a deux hypothèses que nous voulons comparer, une hypothèse nulle h_0 et une hypothèse alternative h_1. Avant de lancer l'expérience, nous avons quelques croyances $P(h)$ sur les hypothèses qui sont vraies. Nous faisons une expérience et obtenons des données *d.* Contrairement aux statistiques fréquentistes, les statistiques bayésiennes nous permettent de parler de la probabilité que l'hypothèse nulle soit vraie. Mieux encore, il permet de calculer la **probabilité à postériori de l'hypothèse nulle**, en utilisant la règle de Bayes

$$P(h_0|d) = \frac{P(d|h_0)P(h_0)}{P(d)}$$

Cette formule nous dit exactement comment quantifier la croyance nous devrions avoir dans l'hypothèse nulle après avoir observé les données *d. De* même, nous pouvons calculer comment quantifier la croyance nous devons placer dans l'hypothèse alternative en utilisant pour presque la même équation. Tout ce qu'on fait, c'est changer l'indice

$$P(h_1|d) = \frac{P(d|h_1)P(h_1)}{P(d)}$$

C'est si simple que je me sens comme un idiot de prendre la peine d'écrire ces équations. Je ne fais que copier la règle de Bayes de la section précédente[149].

---

[149] Évidemment, c'est une histoire très simplifiée. Toute la complexité de la vérification des hypothèses bayésiennes dans la vie réelle se résume à la façon dont vous calculez la probabilité P(d|h) lorsque l'hypothèse h est une chose complexe et vague. Je ne vais pas parler de ces complexités dans ce livre, mais je tiens à souligner que même si cette histoire simple est vraie, la vraie vie est plus compliquée que ce que je suis capable de traiter dans un manuel de statistiques d'introduction.

Le facteur Bayes

En pratique, la plupart des analystes bayésiens de données ont tendance à ne pas parler en termes de probabilités à posteriori brutes $P(h_0 \mid d)$ et $P(h_1 \mid d)$. Nous avons plutôt tendance à parler en termes de risque relatif à postériori (posterior odds ratio). Pensez-y comme un pari. Supposons, par exemple, que la probabilité à postériori de l'hypothèse nulle soit de 25 % et que la probabilité à postériori de l'alternative soit de 75 %. L'hypothèse alternative est trois fois plus probable que l'hypothèse nulle, donc nous disons que le risque relatif est de 3:1 en faveur de l'alternative. Mathématiquement, tout ce que nous avons à faire pour calculer le risque relatif à poseriori est de diviser une probabilité a posteriori par l'autre.

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{0{,}75}{0{,}25} = 3$$

Ou, pour écrire la même chose en termes des équations ci-dessus

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0} \times \frac{P(h_1)}{P(h_0)}$$

En fait, cette équation vaut la peine d'être développée. Il y a ici trois termes différents que vous devriez connaitre. Sur le côté gauche, nous avons le **risque relatif a posterori** (Posterior odds ou ratio a posteriori), qui vous dit ce que vous pensez de la plausibilité relative de l'hypothèse nulle et de l'hypothèse alternative *après avoir* vu les données. Sur le côté droit, nous avons le ratio des **probabilités a priori** (Prior odds), qui indiquent ce que vous pensiez *avant de* voir les données. Au milieu, nous avons le **facteur Bayes**, qui décrit la quantité de preuves fournies par les données.

$$\underbrace{\frac{P(h_1|d)}{P(h_0|d)}}_{\text{Posterior odds}} = \underbrace{\frac{P(d|h_1)}{P(d|h_0)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(h_1)}{P(h_0)}}_{\text{Prior odds}}$$

Le facteur Bayes (parfois abrégé en **BF**) occupe une place particulière dans les tests d'hypothèse bayésiens, car il joue un rôle similaire à la *valeur p* dans les tests d'hypothèse orthodoxes. Le facteur Bayes quantifie la force des preuves fournies par les données et, à ce titre, c'est le facteur Bayes que les gens ont tendance à déclarer lorsqu'ils font un test d'hypothèse bayésien. La raison pour laquelle les facteurs Bayes sont rapportés plutôt que les ratios a posteriori est que les chercheurs n'ont pas tous les mêmes a priori. Certaines personnes peuvent avoir un préjugé fort pour croire que l'hypothèse nulle est vraie, d'autres peuvent avoir un préjugé fort pour croire qu'elle est fausse. Pour cette raison, la façon polie qu'un chercheur sérieux doit adopter, c'est de déclarer le facteur Bayes. De cette

façon, quiconque lit le journal peut multiplier le facteur Bayes par ses propres probabilités a priori *personnelles* et déterminer par lui-même quelles seraient les ratios a porteriori. Quoi qu'il en soit, par convention, nous aimons prétendre que nous accordons la même importance à l'hypothèse nulle et à l'alternative, auquel cas la probabilité a priori est égale à 1, et le ratio a posteriori est égal au facteur Bayes.

Interprétation des facteurs Bayes

L'un des aspects les plus intéressants du facteur Bayes, c'est que les chiffres sont intrinsèquement significatifs. Si vous faites une expérience et que vous calculez un facteur Bayes de 4, cela signifie que la preuve fournie par vos données correspond à un risque relatif de 4:1 en faveur de l'alternative. Toutefois, certains auteurs ont tenté de quantifier des valeurs normatives de preuve qui seraient considérées comme significatives dans un contexte scientifique. Les deux plus largement utilisés sont ceux de Jeffreys (1998) et Kass et Raftery (1995). Entre les deux, j'ai tendance à préférer le tableau de Kass et al (1995) (1995) parce qu'il est un peu plus conservateur. Le voici :

Bayes factor	Interpretation
1-3	Negligible evidence
3-20	Positive evidence
20-150	Strong evidence
>150	Very strong evidence

Et pour être tout à fait honnête, je pense que même les normes de Kass et al (1995) sont un peu charitables. Si ça ne tenait qu'à moi, j'aurais appelé la catégorie « preuves positives » « preuves faibles ». Pour moi, tout ce qui se situe entre 3:1 et 20:1 est au mieux une preuve « faible » ou « modeste ». Mais il n'y a pas de règle absolue. Ce qui compte comme preuve forte ou faible dépend entièrement de votre prudence et des normes sur lesquelles votre communauté insiste avant de vouloir qualifier une conclusion de « vraie ».

Quoi qu'il en soit, notez que tous les chiffres énumérés ci-dessus ont un sens si le facteur Bayes est supérieur à 1 (c.-à-d. que les données probantes favorisent l'hypothèse alternative). Cependant, l'un des grands avantages pratiques de l'approche bayésienne par rapport à l'approche orthodoxe est qu'elle permet également de quantifier les preuves en faveur de l'hypothèse nulle. Vous pouvez choisir de déclarer un facteur Bayes inférieur à 1, mais pour être honnête, je trouve cela déroutant. Par exemple, supposons que la probabilité des données sous l'hypothèse nulle $P(d \mid h_0)$ est égale à 0,2, et que la probabilité correspondante $P(d \mid h_1)$ sous l'hypothèse alternative est 0,1. En utilisant les équations données ci-dessus, le facteur de Bayes ici serait

$$BF = \frac{P(d|h_1)}{P(d|h_0)} = \frac{0,1}{0,2} = 0,5$$

Si on lit littéralement, ce résultat indique que les preuves en faveur de l'alternative sont de 0,5 à 1, ce que je trouve difficile à comprendre. Pour moi, il est beaucoup plus logique de renverser l'équation et de rapporter le montant de la preuve op en faveur de la *l'hypothèse nulle*. En d'autres termes, ce que nous calculons est

$$BF' = \frac{P(d|h_0)}{P(d|h_1)} = \frac{0{,}2}{0{,}1} = 2$$

Et ce que nous signalons, c'est un facteur Bayes de 2:1 en faveur de l'hypothèse nulle. Beaucoup plus facile à comprendre, et vous pouvez l'interpréter à l'aide du tableau ci-dessus.

Pourquoi être Bayésien ?

Jusqu'à présent, je me suis concentré exclusivement sur la logique qui sous-tend les statistiques bayésiennes. Nous avons parlé de l'idée de la « probabilité comme degré de croyance », et de ce qu'elle implique sur la façon dont un agent rationnel devrait raisonner le monde. La question à laquelle vous devez répondre vous-même est la suivante : comment voulez-vous faire vos statistiques ? Voulez-vous être un statisticien orthodoxe qui se fie aux distributions d'échantillonnage et aux valeurs *p* pour guider ses décisions ? Ou voulez-vous être un Bayésien, en vous appuyant sur des choses comme les croyances à priori, les facteurs Bayes et les règles de révision rationnelle des croyances ? Pour être tout à fait honnête, je ne peux pas répondre à cette question pour vous. En fin de compte, cela dépend de ce que vous pensez être juste. C'est votre décision et votre décision seule. Cela dit, je peux vous expliquer un peu pourquoi *je* préfère l'approche bayésienne.

Des statistiques qui signifient ce que vous pensez qu'elles signifient

Tu n'arrêtes pas d'utiliser ce mot. Je ne pense pas que ça veuille dire ce que tu penses que ça veut dire. - Inigo Montoya, La princesse mariée[150]

Pour moi, l'un des plus grands avantages de l'approche bayésienne est qu'elle répond aux bonnes questions. Dans le cadre bayésien, il est tout à fait raisonnable et possible de faire référence à « la probabilité qu'une hypothèse soit vraie ». Vous pouvez même essayer de calculer cette probabilité. En fin de compte, n'est-ce pas ce que vous *voulez* que vos tests statistiques vous disent ? Pour un être humain, cela semble être *l'objectif* des statistiques, c'est-à-dire de déterminer ce qui est vrai et ce qui ne l'est pas. Chaque fois que vous n'êtes pas sûr de ce qu'est la vérité, vous devriez utiliser le langage de la théorie des probabilités pour dire des choses comme « il y a 80% de chances que la théorie A soit vraie, mais 20% de chances que la théorie B soit vraie à la place ». Cela semble si évident pour un humain, mais c'est pourtant explicitement interdit dans le cadre orthodoxe. Pour un fréquentiste, de telles affirmations sont un non-sens car « la théorie est vraie » n'est pas un événement répétable. Une théorie est vraie ou elle ne l'est pas, et aucune déclaration probabiliste n'est permise, peu importe de quel point vous voulez le faire. Il y a une raison pour laquelle, à la section 9.5, je vous ai averti à plusieurs reprises de *ne pas* interpréter la *valeur p* comme la probabilité que l'hypothèse nulle soit vraie. Il y a une raison pour laquelle presque tous les

---

[150] Je dois noter au passage que je ne suis pas la première personne à utiliser cette citation pour me plaindre des méthodes fréquentistes. Rich Morey et ses collègues ont eu l'idée en premier. Je la vole sans vergogne parce que c'est une citation à utiliser dans ce contexte et je refuse de rater une occasion de citer *The Princess Bride*.

manuels de statistiques sont obligés de répéter cet avertissement. C'est parce que les gens *veulent* désespérément que ce soit la bonne interprétation. Malgré le dogme frequentiste, une vie entière d'expérience dans l'enseignement aux étudiants de premier cycle et dans l'analyse de données sur une base quotidienne me suggère que la plupart des humains pensent que « la probabilité que l'hypothèse soit vraie » n'est pas seulement du sens, c'est l'idée à laquelle nous tenons le *plus*. C'est une idée tellement séduisante que même des statisticiens formés sont victimes de l'erreur d'essayer d'interpréter une valeur *p* de cette façon. Par exemple, voici une citation d'un rapport officiel de Newspoll en 2013, expliquant comment interpréter l'analyse de leurs données (fréquentistes) :[151]

Tout au long du rapport, des changements statistiquement significatifs ont été notés, le cas échéant. Tous les tests de signification ont été basés sur le niveau de confiance de 95 %. **Cela signifie que si un changement est noté comme étant statistiquement significatif, il y a une probabilité de 95 % qu'un changement réel se soit produit**, et non simplement dû à une variation aléatoire. (non souligné dans l'original)

Non ! Ce *n'est pas* ce que $p<.05$ signifie. Ce *n'est pas* ce que 95 % de confiance signifie pour un statisticien fréquentiste. La phrase en gras est tout simplement fausse. Les méthodes orthodoxes ne peuvent pas vous dire « qu'il y a 95% de chances qu'un changement réel se soit produit », car ce n'est pas le genre d'événement auquel les probabilités fréquentistes peuvent être attribuées. Pour un théoricien fréquentiste, cette phrase devrait être dénuée de sens. Même si vous êtes un fréquentiste plus pragmatique, c'est toujours la mauvaise définition d'une valeur *p*. Il n'est tout simplement pas permis ou correct de dire des choses si l'on veut se fier à des outils statistiques orthodoxes.

D'un autre côté, supposons que vous soyez Bayésien. Bien que le passage en gras soit la mauvaise définition d'une valeur *p*, c'est à peu près exactement ce que ceut dire un Bayésien lorsqu'on dit que la probabilité à posteriori de l'hypothèse alternative est supérieure à 95%. Voilà le truc. Si le la probabilité à postérioiri du bayésien est réellement ce que vous vous *voulez*, pourquoi essayer d'utiliser des méthodes orthodoxes ? Si vous voulez faire des affirmations bayésiennes, tout ce que vous avez à faire est d'être un Bayésien et d'utiliser des outils bayésiens.

Pour ma part, j'ai trouvé que passer à la vision bayésienne était le point de vue le plus libérateur. Une fois que vous avez sauté le pas, vous n'avez plus à vous soucier des définitions contre-intuitives des valeurs *p*. Vous n'avez pas besoin de vous rappeler pourquoi vous ne pouvez pas dire que vous êtes sûr à 95 % que la vraie moyenne se situe dans un certain intervalle. Tout ce que vous avez à faire, c'est d'être honnête sur ce que vous croyiez avant de mener l'étude et de rapporter ce que vous avez appris en le faisant. Ça a l'air sympa, n'est-ce pas ? Pour moi, c'est la grande promesse de l'approche bayésienne. Vous faites l'analyse que vous voulez vraiment faire et vous exprimez ce que vous croyez vraiment que les données vous disent.

---

[151] http://about.abc.net.au/reports-publications/appreciation-survey-summary-report-2013/

Des normes de preuve auxquelles vous pouvez croire

Si[p] est inférieur à .02, il est fortement suggéré que l'hypothèse[nulle] ne correspond pas à l'ensemble des faits. Nous ne nous égarerons pas souvent si nous traçons une ligne conventionnelle à .05 et en considérant que [de plus petites valeurs de p] indiquent un écart réel. - Sir Ronald Fisher (1925)

Considérons la citation ci-dessus de Sir Ronald Fisher, l'un des fondateurs de ce qui est devenu l'approche orthodoxe de la statistique. Si quelqu'un a déjà eu le droit d'exprimer une opinion sur la fonction prévue des valeurs *p*, c'est bien Fisher. Dans ce passage, tiré de son guide classique *Statistical Methods for Research Workers*, il est assez clair sur ce que signifie le rejet d'une hypothèse nulle à $p<.05$. A son avis, si nous prenons $p<.05$ pour signifier qu'il y a « un effet réel », alors « nous ne serons pas souvent égarés ». Ce point de vue n'est pas inhabituel. D'après mon expérience, la plupart des praticiens expriment des points de vue très semblables à ceux de Fisher. Essentiellement, la $p<.05$ est supposée représenter une norme de preuve assez stricte.

Quelle est la réalité de cela ? Une façon d'aborder cette question est d'essayer de convertir les valeurs *p* en facteurs Bayes et de voir comment on peut comparer les deux. Ce n'est pas une chose facile à faire parce qu'une valeur *p* est un type de calcul fondamentalement différent d'un facteur Bayes, et ils ne mesurent pas la même chose. Cependant, il y a eu quelques tentatives pour établir la relation entre les deux, et c'est quelque peu surprenant. Par exemple, Johnson (2013) présente un cas assez convaincant où (du moins pour les *tests t*) la $p< .05$ correspond approximativement à un facteur Bayes compris entre 3:1 et 5:1 en faveur de l'alternative. Si c'est vrai, alors la proposition de Fisher est un peu exagérée. Supposons que l'hypothèse nulle soit vraie environ la moitié du temps (c.-à-d. que la probabilité antérieure de H_0 est de 0,5), et que nous utilisons ces nombres pour calculer la probabilité a posteriori de l'hypothèse nulle puisqu'elle a été rejetée à $p<.05$. En utilisant les données de Johnson (2013), nous voyons que si vous rejetez l'hypothèse nulle à $p< .05$, vous aurez raison environ 80% du temps. Je ne sais pas ce que vous en pensez, mais, à mon avis, une norme de preuve qui garantit que vous vous trompez sur 20 % de vos décisions n'est pas suffisante. Il n'en demeure pas moins que, tout à fait contrairement à ce que prétend Fisher, si vous rejetez à $p<.05$ vous vous égarerez très souvent. Ce n'est pas du tout un seuil de preuve très strict.

La valeur *p* est un mensonge.

Le gâteau est un mensonge. Le gâteau est un mensonge. Le gâteau est un mensonge. Le gâteau est un mensonge. - Portal[152]

Bien, à ce stade, vous pensez peut-être que le vrai problème ne vient pas des statistiques orthodoxes, mais juste du $p<.05$ standard. En un sens, c'est vrai. La recommandation de Johnson (2013) n'est pas que « tout le monde doit être bayésien à partir de maintenant ». Au contraire, la suggestion est qu'il serait plus sage de changer la norme conventionnelle et de choisir un niveau comme un $p < .01$. Ce n'est pas un point de vue déraisonnable, mais à mon sens, le problème est un peu plus grave que cela. À mon avis, il y a un assez gros

---

[152] http://knowyourmeme.com/memes/the-cake-is-a-lie

problème dans la construction de la plupart (mais pas tous) les tests d'hypothèses orthodoxes. Ils sont tout à fait naïfs quant à la façon dont les humains font de la recherche, et c'est pour cette raison que la plupart des valeurs *p* sont fausses.

Cela vous parait une affirmation absurde ? Eh bien, considérez le scénario suivant. Vous avez formulé une hypothèse de recherche très intéressante et vous concevez une étude pour la mettre à l'épreuve. Vous êtes très diligent, alors vous effectuez une analyse de puissance pour déterminer quelle devrait être la taille de votre échantillon, et vous effectuez l'étude. Vous exécutez votre test d'hypothèse et vous obtenez une valeur *p* de 0,072. Vraiment très ennuyeux, non ?

Que devriez-vous faire ? Voici quelques possibilités :

1. Vous concluez qu'il n'y a pas d'effet et essayez de le publier comme résultat nul.
2. Vous supposez qu'il pourrait y avoir un effet et essayez de le publier comme un résultat « significatif limite ».
3. Vous abandonnez et essayez une nouvelle étude
4. Vous collectez quelques données supplémentaires pour voir si la valeur de *p* augmente ou (de préférence !) descend en dessous du critère « magique » de $p<.05$

Laquelle choisiriez-vous ? Avant de poursuivre la lecture, je vous invite à prendre le temps d'y réfléchir. Soyez honnête avec vous-même. Mais n'en faites pas trop, parce que vous êtes foutu, peu importe ce que vous choisissez. En me basant sur mes propres expériences en tant qu'auteur, rélecteur et éditeur, ainsi que sur les histoires que j'ai entendues d'autres personnes, voici ce qui va se passer dans chaque cas :

- Commençons par l'option 1. Si vous essayez de le publier comme résultat nul, l'article aura du mal à être publié. Certains critiques penseront que $p = .072$ n'est pas vraiment un résultat nul. Ils diront que c'est à la limite de l'important. D'autres examinateurs conviendront qu'il s'agit d'un résultat nul, mais prétendront que même si certains résultats nuls *sont* publiables, le vôtre ne l'est pas. Un ou deux relecteurs peuvent même être de votre côté, mais vous devrez livrer une bataille ardue pour y arriver.

- Réfléchissons à l'option numéro 2. Supposons que vous essayez de le publier comme un résultat significatif limite. Certains critiques diront qu'il s'agit d'un résultat nul et qu'il ne devrait pas être publié. D'autres prétendront que les preuves sont ambiguës et que vous devriez recueillir plus de données jusqu'à ce que vous obteniez un résultat significatif clair. Encore une fois, le processus de publication n'est pas de votre côté.

- Compte tenu des difficultés à publier un résultat « ambigu » comme $p = .072$, l'option numéro 3 peut sembler tentante : abandonner et faire autre chose. Mais c'est la recette du suicide professionnel. Si vous abandonnez et essayez un nouveau projet chaque fois que vous vous trouvez face à une ambiguïté, votre travail ne sera jamais publié. Et si vous êtes dans le milieu universitaire sans avoir publié d'articles, vous pouvez perdre votre emploi. Donc, cette option n'existe pas.

- On dirait que vous êtes coincé avec l'option 4. Vous n'avez pas de résultats concluants, alors vous décidez de recueillir d'autres données et de réexécuter l'analyse. Cela

> semble raisonnable, mais malheureusement pour vous, si vous faites cela, toutes vos *valeurs p* sont maintenant incorrectes. *Toutes les trois*. Pas seulement les *valeurs p que* vous avez calculées pour *cette* étude. Toutes les trois. Toutes les *valeurs p que* vous avez calculées dans le passé et toutes les *valeurs p que* vous calculerez dans le futur. Heureusement, personne ne le remarquera. Vous serez publié, et vous aurez menti.

Mais attendez ! Comment cette dernière partie peut-elle être vraie ? Je veux dire, ça semble être une stratégie parfaitement raisonnable. Vous avez recueilli des données, les résultats n'étaient pas concluants, alors maintenant ce que vous voulez faire, c'est recueillir plus de données jusqu'à ce que les résultats *soient* concluants. Qu'est-ce qu'il y a de mal à ça ?

Honnêtement, il n'y a rien de mal à ça. C'est une chose raisonnable, raisonnable et rationnelle à faire. Dans la vraie vie, c'est exactement ce que chaque chercheur fait. Malheureusement, la théorie des tests d'hypothèse nulle telle que je l'ai décrite au chapitre 9 vous *interdit* de le faire. La raison[153] est que la théorie suppose que l'expérience est terminée et que toutes les données sont disponibles. Parce qu'elle suppose que l'expérience est terminée, elle n'envisage que *deux* décisions possibles. Si vous utilisez la méthode conventionnelle $p<.05$, ces décisions sont :

Outcomes	Action
p less than .05	Reject the null
p greater than .05	Retain the null

Ce que *vous* faites, c'est ajouter une troisième action possible au problème de prise de décision. Plus précisément, ce que vous faites, c'est d'utiliser la valeur *p* elle-même comme raison pour justifier la poursuite de l'expérience. Par conséquent, vous avez transformé la procédure de prise de décision en une procédure qui ressemble davantage à celle-ci :

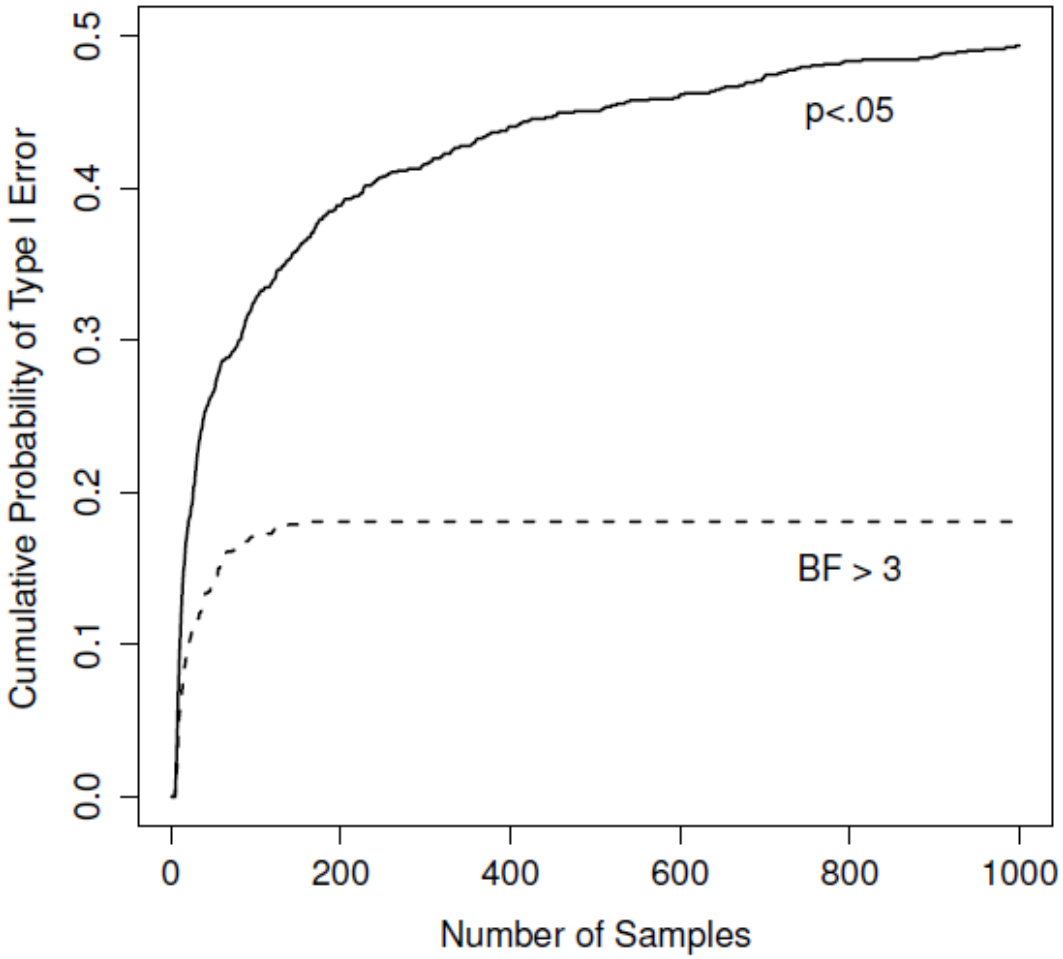
Outcomes	Action
p less than .05	Reject the null
p between .05 and 1	Continue the experiment
p greater than 1	Stop the experiment and retain the null

---

[153] Pour être tout à fait honnête, je dois reconnaître que tous les tests statistiques orthodoxes ne reposent pas sur cette hypothèse stupide. Il existe un certain nombre d'outils d'analyse séquentielle qui sont parfois utilisés dans les essais cliniques et autres. Ces méthodes sont fondées sur l'hypothèse que les données sont analysées au fur et à mesure qu'elles arrivent, et ces tests ne sont pas remis en cause comme je le prétends ici. Cependant, les méthodes d'analyse séquentielle sont construites d'une manière très différente de la version « standard » des tests d'hypothèse nulle. Ils ne sont pas inclus dans les manuels d'introduction, et ils ne sont pas très largement utilisés dans la littérature psychologique. La préoccupation que je soulève ici est valable pour chaque test orthodoxe que j'ai présenté jusqu'ici et pour presque tous les tests que j'ai vus rapportés dans les journaux que j'ai lus.

La théorie « de base » des tests d'hypothèse nulle n'est pas construite pour gérer ce genre de choses, pas sous la forme que j'ai décrite au chapitre 9. Si vous êtes le genre de personne qui choisirait de « collecter plus de données » dans la vie réelle, cela implique que vous *ne* prenez *pas de* décisions en accord avec les règles du test d'hypothèse nulle. Même si vous arrivez à la même décision que le test d'hypothèse, vous ne suivez pas le *processus de* décision qu'il implique, et c'est cette incapacité à suivre le processus qui pose problème.[154] Vos valeurs *p* sont des mensonges.

Pire encore, c'est un mensonge dangereux, parce qu'ils sont tous *trop petits*. Pour vous donner une idée de la gravité de la situation, considérez le scénario suivant (le pire des cas). Imaginez que vous êtes un vrai un chercheur super enthousiaste avec un budget serré qui n'a pas prêté attention à mes mises en garde ci-dessus.



[154] Un problème connexe : http://xkcd.com/1478/

Figure 16-1 : Dans quelle mesure les choses peuvent-elles mal tourner si vous réexécutez vos tests chaque fois que de nouvelles données arrivent ? Si vous êtes un fréquentiste, la réponse est « très mauvaise ».

Vous concevez une étude comparant deux groupes. Vous voulez désespérément voir un résultat significatif à la $p < .05$, mais vous ne voulez vraiment pas collecter plus de données qu'il n'en faut (parce que c'est cher). Afin de réduire les coûts, vous commencez à collecter des données, mais chaque fois qu'une nouvelle observation arrive, vous effectuez un test *t* sur vos données. Si le test *t* donne $p<.05$, vous arrêtez l'expérience et rapportez un résultat significatif. Si ce n'est pas le cas, vous continuez à collecter des données. Vous continuez à le faire jusqu'à ce que vous atteigniez votre limite de dépenses prédéfinie pour cette expérience. Supposons que cette limite s'applique à $N=1000$ observations. Il s'avère que la vérité est qu'il n'y a pas d'effet réel à trouver : l'hypothèse nulle est vraie. Alors, quelle est la probabilité que vous arriviez à la fin de l'expérience et que vous concluiez (correctement) qu'il n'y a aucun effet ? Dans un monde idéal, la réponse ici devrait être 95%. Après tout, l'intérêt de la $p< .05$ est de contrôler le taux d'erreur de type I à 5 %, alors ce que nous espérons, c'est qu'il n'y a que 5 % de chances de rejeter faussement l'hypothèse nulle dans cette situation. Cependant, il n'y a aucune garantie que ce soit vrai. Vous enfreignez les règles. Parce que vous effectuez des tests de façon répétée, en jetant un coup d'oeil à vos données pour voir si vous avez obtenu un résultat significatif, tous les paris sont annulés.

A quel point est-ce grave? La réponse est représentée par la ligne noire pleine de la Figure 16-1, et c'est *incroyablement* mauvais. Si vous jetez un coup d'oeil à vos données après chaque observation, il y a 49% de chance que vous fassiez une erreur de type I. C'est, euh, un peu plus que les 5% que vous devriez avoir. A titre de comparaison, imaginez que vous avez utilisé la stratégie suivante. Vous commencez à recueillir des données. Chaque fois qu'une observation arrive, effectuez un test *t* bayésien (section 16.4) et examinez le facteur Bayes. Je suppose que Johnson (2013) a raison, et je traiterai un facteur de Bayes de 3:1 comme étant à peu près équivalent à une valeur *p* de 0,05.[155] Cette fois-ci, notre chercheur à la gâchette heureuse utilise la procédure suivante. Si le facteur de Bayes est de 3:1 ou plus en faveur de la l'hypothèse nulle, vous arrêter l'expérience et conserver la valeur nulle. S'il est de 3:1 ou plus en faveur de l'alternative, vous arrêtez l'expérience et rejetez l'hypothèse nulle. Dans le cas contraire, vous continuez le test. Maintenant, comme la dernière fois, supposons que l'hypothèse nulle soit vraie. Que se passe-t-il ? Il se trouve que j'ai également effectué les simulations pour ce scénario, et les résultats sont représentés par la ligne pointillée de la Figure 16-1. Il s'avère que le taux d'erreur de type I est beaucoup plus faible que le taux de 49 % que nous obtenions en utilisant le test *t* orthodoxe.

D'une certaine façon, c'est remarquable. Le *but* des tests d'hypothèse nulle orthodoxes est de contrôler le taux d'erreur de type I. Les méthodes bayésiennes ne sont pas du tout

---

[155] Certains lecteurs pourraient se demander pourquoi j'ai choisi 3:1 plutôt que 5:1, étant donné que Johnson (2013) suggère que $p = .05$ se trouve quelque part dans cette intervalle. Je l'ai fait dans le but d'être charitable avec la *valeur p*. Si j'avais choisi un facteur Bayes de 5:1, les résultats seraient encore plus favorables à l'approche bayésienne.

conçues pour cela. Or, il s'avère que, face à un chercheur à la « gâchette heureuse » qui continue à faire des tests d'hypothèse au fur et à mesure que les données arrivent, l'approche bayésienne est beaucoup plus efficace. Même la norme 3:1, que la plupart des Bayésiens considéreraient comme inacceptablement laxiste, est beaucoup plus sûre que la norme $p<.05$.

C'est vraiment si grave ?

L'exemple que j'ai donné dans la section précédente est une situation assez extrême. Dans la vraie vie, les gens ne font pas de tests d'hypothèse chaque fois qu'une nouvelle observation arrive. Il n'est donc pas juste de dire que la $p>.05$ correspond à un taux d'erreur de type I de 49 % (c.-à-d., $p=.49$). Mais le fait est que si vous voulez que vos valeurs p soient honnêtes, vous devez soit changer complètement la façon de faire des tests d'hypothèse, soit appliquer une règle stricte de *ne pas regarder avant*. Vous *n'êtes pas* autorisé à utiliser les données pour décider quand mettre fin à l'expérience. Vous *n'êtes pas* autorisé à regarder une valeur p « limite » et à décider de collecter plus de données. Vous n'êtes même pas autorisé à modifier votre stratégie d'analyse de données après avoir examiné les données. Vous êtes strictement tenu de suivre ces règles, sinon les valeurs p que vous calculerez seront absurdes.

Oui, ces règles sont étonnamment strictes. En classe, il y a quelques années, j'ai demandé aux élèves de réfléchir à ce scénario. Supposons que vous avez commencé votre étude avec l'intention de collecter $N = 80$ personnes. Au début de l'étude, vous suivez les règles, refusant d'examiner les données ou d'effectuer des tests. Mais quand vous atteignez $N = 50$, votre volonté cède et vous jetez un coup d'oeil. Et là surprise, vous avez un résultat significatif ! Bien sûr, vous aviez dit que vous continueriez à mener l'étude jusqu'à une taille d'échantillon de $N=80$, mais cela semble un peu inutile maintenant. Le résultat est significatif avec une taille d'échantillon de $N = 50$, donc à quoi servirait de continuer à collecter des données ? N'êtes-vous pas tenté d'arrêter ? Juste un peu ? N'oubliez pas que si c'est le cas, votre taux d'erreur de type I à $p> .05$ vient de s'envoler à 8%. Lorsque vous rendez compte dans votre journal de $p>.05$ dans votre journal, ce que vous êtes en train de dire en vrai, c'est $p<.08$. Cela illustre à quel point les conséquences « d'un simple coup d'oeil » peuvent être graves.

Considérez maintenant que la littérature scientifique est remplie de tests t, d'analyses de variance, de régressions et de tests khi-deux. Quand j'ai écrit ce livre, je n'ai pas choisi ces tests arbitrairement. La raison pour laquelle ces quatre outils apparaissent dans la plupart des textes d'introduction à la statistique est qu'il s'agit des outils de base en science. Aucun de ces outils n'inclut de correction pour faire face au « coup d'œil dans les données » : ils supposent tous que vous ne le faites pas. Mais dans quelle mesure cette hypothèse est réaliste ? Dans la vie réelle, combien de personnes, selon vous, ont « jeté un coup d'œil » à leurs données avant la fin de l'expérience et adapté leur comportement ultérieur après avoir vu à quoi elles ressemblaient ? Sauf lorsque la procédure d'échantillonnage est fixée par une contrainte externe, je suppose que la réponse est « la plupart des gens l'ont fait ». Si cela s'est produit, vous pouvez en déduire que les valeurs de p rapportées sont fausses. Pire encore, comme nous ne savons pas quel processus de décision ils ont suivi, nous n'avons aucun moyen de savoir quelles *auraient dû* être les valeurs *p.* Vous ne pouvez pas calculer

une valeur *p* si vous ne connaissez pas la procédure de prise de décision utilisée par le chercheur. Il s'en suit que la valeur *p* rapportée n'est pas fiable.

Compte tenu de tout ce qui précède, quel est le message à retenir ? Ce n'est pas que les méthodes bayésiennes sont infaillibles. Si un chercheur est déterminé à tricher, il peut toujours le faire. La règle de Bayes ne peut pas empêcher les gens de mentir, ni les empêcher de truquer une expérience. Ce n'est pas ce que je veux dire. Mon argument est le même que celui que j'ai fait valoir au tout début du livre à la section 1.1 : la raison pour laquelle nous faisons des tests statistiques est de nous protéger de nous-mêmes. Et la raison pour laquelle « l'examen des données » est une telle préoccupation, c'est qu'il est si tentant, *même pour les chercheurs honnêtes*. Une théorie de l'inférence statistique doit le reconnaître. Oui, vous pourriez essayer de défendre les valeurs *p* en disant que c'est la faute du chercheur s'il ne les utilise pas correctement, mais à mon avis, ce n'est pas la question. Une théorie de l'inférence statistique qui est tellement naïve à l'égard des humains qu'elle ne tient même pas compte de la possibilité que le chercheur puisse *examiner ses propres données* n'est pas une théorie valable. La citation suivante résume ce que je veux dire essentiellement :

Les bonnes lois ont leur origine dans les mauvaises moeurs. - Ambrosius Macrobius[156]

De bonnes règles pour les tests statistiques doivent reconnaître la fragilité humaine. Aucun de nous n'est sans péché. Aucun de nous n'est à l'abri de la tentation. Un bon système d'inférence statistique devrait fonctionner même lorsqu'il est utilisé par de vrais humains. Les tests d'hypothèse nulle orthodoxe ne le font pas.[157]

Tests *t* bayésiens

Un type important de problème d'inférence statistique discuté dans ce livre est la comparaison entre deux moyennes, discutée en détail dans le chapitre sur les tests *t* (chapitre 11). Si vous vous souvenez bien, il existe plusieurs versions du test *t*. Je vais parler

---

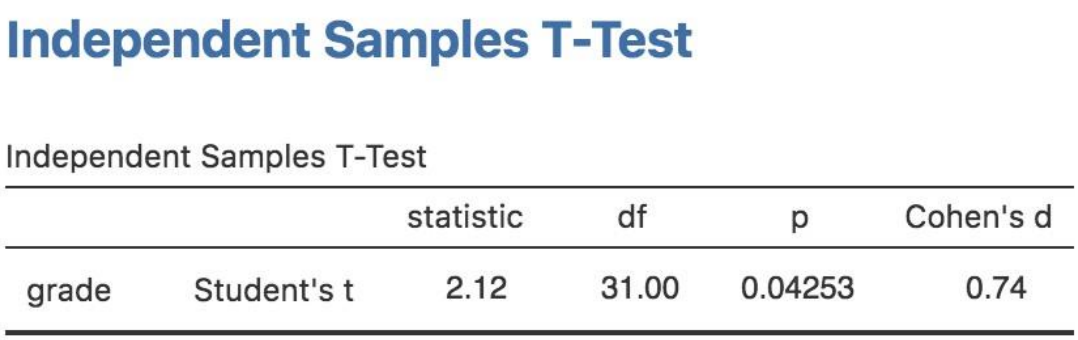
[156] http://www.quotationspage.com/quotes/Ambrosius_Macrobius/

[157] Bien sûr, je sais que certains fréquentistes avertis vont lire ceci et commencer à contester les propos de cette section. Je ne suis pas stupide. Je sais parfaitement que si vous adoptez une perspective d'analyse séquentielle, vous pouvez éviter ces erreurs dans le cadre orthodoxe. Je sais aussi que vous pouvez explicitement concevoir des études avec des analyses intermédiaires à l'esprit. Donc oui, dans un sens, j'attaque une version « bouc émissaire » des méthodes orthodoxes. Cependant, le bous émissaire que j'attaque est celui qui est utilisé par presque tous les praticiens. Si jamais les méthodes séquentielles deviennent la norme chez les psychologues expérimentaux et que je ne suis plus obligé de lire 20 analyses de variance extrêmement douteuses par jour, je promets de réécrire cette section et de réduire le vitriol. Mais d'ici là, je maintiens mon affirmation selon laquelle les méthodes utilisant le facteur Bayes par défaut sont beaucoup plus robustes face aux pratiques d'analyse des données telles qu'elles existent dans le monde réel. Les méthodes orthodoxes par défaut craignent, et nous le savons tous.

un peu des versions bayésiennes des tests *t* pour des échantillons indépendants et pour des échantillons appariés *t-tests* dans cette section.

Test *t* pour échantillons indépendants

Le type de test *t* le plus courant est le test *t* d'échantillons indépendants, et il apparaît lorsque vous disposez de données comme dans l'ensemble de données harpo.csv que nous avons utilisé dans le chapitre précédent sur les tests *t* (chapitre 11). Dans cet ensemble de données, nous avons deux groupes d'élèves, ceux qui ont reçu des leçons d'Anastasia et ceux qui ont pris leurs cours avec Bernadette. La question à laquelle nous voulons répondre est de savoir s'il y a une différence dans les notes obtenues par ces deux groupes d'élèves. Au chapitre 11, j'ai suggéré d'analyser ce type de données à l'aide du test *t* pour des Échantillons indépendants dans Jamovi, qui nous a donné les résultats de la Figure 16-2. Comme nous obtenons une valeur *p* inférieure à 0,05, nous rejetons l'hypothèse nulle.



Independent Samples T-Test

		statistic	df	p	Cohen's d
grade	Student's t	2.12	31.00	0.04253	0.74

Figure 16-2 : Échantillons indépendants - résultat du test t dans Jamovi

A quoi ressemble la version bayésienne du test *t*? Nous pouvons obtenir l'analyse du facteur de Bayes en cochant la case « Bayes Factor » sous l'option « Tests », et en acceptant la valeur à priori par défaut suggérée dans l'option « Prior ». Ceci donne les résultats présentés dans le tableau de la Figure 16-3. Ce que nous obtenons dans ce tableau est une statistique du facteur Bayes de 1,75, ce qui signifie que les preuves fournies par ces données sont d'environ 1,8:1 en faveur de l'hypothèse alternative.

Avant de poursuivre, il convient de souligner la différence entre les résultats des tests orthodoxes et ceux des tests bayésiens. Selon le test orthodoxe, nous avons obtenu un résultat significatif, mais à peine. Néanmoins, beaucoup de gens seraient heureux d'accepter *p*=.043 comme preuve raisonnablement solide d'un effet. Par contre, notez que le test bayésien n'atteint même pas un seuil de 2:1 en faveur d'un effet, et serait considéré comme une preuve très faible au mieux. D'après mon expérience, c'est un résultat assez typique. Les méthodes bayésiennes exigent habituellement plus de preuves avant de rejeter l'hypothèse nulle.

Independent Samples T-Test

Independent Samples T-Test

		statistic	±%	df	p	Cohen's d
grade	Student's t	2.12		31.00	0.04253	0.74
	Bayes factor$_{10}$	1.75	7.57e−6			

Figure 16-3 : Analyse des facteurs Bayes en parallèle avec le test t d'échantillons indépendants

Test *t* pour échantillons appariés

A la section 11.5, j'ai discuté de l'ensemble de données chico.csv dans lequel les notes des élèves ont été mesurées lors de deux tests, et nous voulions savoir si les notes avaient augmenté du test 1 au test 2. Comme chaque élève a fait les deux tests, l'outil utilisé pour analyser les données était un test *t* pour des échantillons appariés de *tests*. La Figure 16-4 montre le tableau des résultats de Jamovi pour le test *t* apparié conventionnel à côté de l'analyse du facteur de Bayes. A ce stade, j'espère que vous pourrez lire cette sortie sans aucune difficulté. Les données fournissent des preuves d'environ 6000:1 en faveur de l'alternative. Nous pourrions probablement rejeter l'hypothèse nulle avec un peu plus de confiance !

Paired Samples T-Test

			statistic	±%	df	p
grade_test2	grade_test1	Student's t	6.48		19.00	< .00001
		Bayes factor$_{10}$	5991.58	6.09e-10		

Figure 16-4 : Résultats pour un test t pour échantillons appariés et le Facteur de Bayes dans Jamovi

Résumé

La première moitié de ce chapitre a porté principalement sur les fondements théoriques des statistiques bayésiennes. J'ai présenté les principes mathématiques de fonctionnement de l'inférence bayésienne (section 16.1), et j'ai donné un aperçu très simple de la façon dont la vérification des hypothèses bayésiennes est généralement effectuée (section 16.2). Enfin, j'ai consacré un peu d'espace à expliquer pourquoi je pense que les méthodes bayésiennes valent la peine d'être utilisées (Section 16.3).

Puis j'ai donné un exemple pratique, un *t-test* bayésien (Section 16.4). Si vous souhaitez en savoir plus sur l'approche bayésienne, il existe de nombreux ouvrages intéressants que vous pouvez consulter. Le livre de John Kruschke, *Doing Bayesian Data Analysis,* est un bon point de départ (Kruschke 2011) et est un bon mélange de théorie et de pratique. Son

approche est un peu différente de celle du « facteur de Bayes » dont j'ai parlé ici, de sorte que vous ne couvrirez pas le même terrain. Si vous êtes un psychologue cognitif, vous voudrez peut-être consulter le livre de Michael Lee et E.J. Wagenmakers intitulé *Bayesian Cognitive Modeling* (Lee and Wagenmakers 2014). J'ai choisi ces deux livres parce que je pense qu'ils sont particulièrement utiles pour les gens de ma discipline, mais il y a beaucoup de bons livres, alors regardez autour de vous !

Épilogue

« Commencez par le commencement », dit le roi, très sérieusement, et continuez jusqu'à la fin, puis arrêtez ». - Lewis Carroll

C'est un peu étrange d'écrire ce chapitre, et plus qu'un peu inapproprié. Un épilogue, c'est ce que vous écrivez quand un livre est terminé, et ce livre n'est vraiment pas fini. Il manque encore *beaucoup* de choses dans ce livre. Il n'a pas encore d'index. Il manque beaucoup de références. Il n'y a pas d'exercices « faites-le vous-même ». Et en général, je pense qu'il y a beaucoup de choses qui ne vont pas dans la présentation, l'organisation et le contenu de ce livre. Compte tenu de tout cela, je ne veux pas essayer d'écrire un « bon » épilogue. Je n'ai pas encore terminé la rédaction du contenu de fond, donc il n'est pas logique d'essayer de tout rassembler. Mais cette version du livre sera mise en ligne pour que les élèves puissent l'utiliser, et vous pourrez peut-être en acheter une copie papier aussi, alors je veux au moins lui donner un vernis de fermeture. Alors, allons-y, d'accord ?

Les statistiques non découvertes

Tout d'abord, je vais parler un peu du contenu que j'aurais aimé avoir la chance de creuser dans cette version du livre, juste pour que vous puissiez vous faire une idée des autres notions qui existent dans le monde des statistiques. Je pense que ce serait important même si ce livre se rapproche d'un produit final. Les étudiants ne réalisent souvent pas que leurs cours d'introduction à la statistique ne sont qu'une introduction. Si vous voulez aller faire de l'analyse de données réelles, vous devez apprendre un tas de nouveaux outils qui étendent le contenu de vos cours de premier cycle de multiples façons différentes. Ne présumez pas qu'on ne peut pas faire quelque chose simplement parce qu'il n'a pas été couvert par le programme de premier cycle. Ne présumez pas qu'une chose est bonne à faire juste parce qu'elle a été traitée dans un cours de premier cycle. Pour vous éviter d'être victime de ce piège, je pense qu'il est utile de donner un aperçu de certaines des autres idées existantes.

Omissions à l'intérieur des sujets couverts

Même parmi les sujets que j'ai couverts dans le livre, il y a beaucoup d'omissions que j'aimerais corriger dans la future version du livre. Je m'en tiendrai à ce qui est purement statistique (plutôt qu'à ce qui est associé à Jamovi), voici une liste représentative, mais non exhaustive, de sujets sur lesquels j'aimerais m'étendre à un moment donné :

- **Autres types de corrélations.** Au chapitre 4, j'ai parlé de deux types de corrélation : Pearson et Spearman. Ces deux méthodes d'évaluation de la corrélation s'appliquent au

cas où vous avez deux variables continues et voulez évaluer la relation entre elles. Qu'en est-il du cas où vos variables sont toutes les deux sur une échelle nominale ? Ou lorsque l'une est nominale et l'autre est continue ? Il existe en fait des méthodes de calcul des corrélations dans de tels cas (par exemple, la corrélation polychorique), et il serait bon de les voir incluses.

- **Plus de détails sur les tailles d'effet.** En général, je pense que le traitement de la taille de l'effet tout au long du livre est un peu plus superficiel qu'il ne devrait l'être. Dans presque tous les cas, j'ai eu tendance à ne choisir qu'une seule mesure de la taille de l'effet (habituellement la plus populaire) et à décrire cela. Cependant, pour presque tous les tests et modèles, il existe de multiples façons de penser la taille de l'effet, et j'aimerais les présenter plus en détail dans l'avenir.

- **Faire face à des suppositions violées.** A plusieurs endroits dans le livre, j'ai parlé de ce que vous pouvez faire lorsque vous constatez que les hypothèses de votre test (ou modèle) sont violées, mais je pense que je devrais en dire plus à ce sujet. En particulier, je pense qu'il aurait été sympa de parler beaucoup plus en détail de la façon dont vous pouvez transformer des variables pour résoudre des problèmes. J'en ai parlé un peu aux sections 6.3 et 6.4, mais la discussion n'est pas assez détaillée, je pense.

- **Termes d'interaction pour la régression.** Au chapitre 14, j'ai parlé du fait que l'on peut avoir des termes d'interaction dans une analyse de variance, et j'ai également souligné que l'analyse de variance peut être interprétée comme une sorte de modèle de régression linéaire. Pourtant, lorsque j'ai parlé de régression au chapitre 12, je n'ai pas du tout parlé des interactions. Cependant, rien ne vous empêche d'inclure des termes d'interaction dans un modèle de régression. C'est juste un peu plus compliqué de comprendre ce qu'une « interaction » signifie réellement quand on parle de l'interaction entre deux prédicteurs continus, et cela peut être fait de plus d'une façon. Malgré tout, j'aurais aimé en parler un peu.

- **Méthode de comparaison planifiée.** Comme je l'ai mentionné au chapitre 14, il n'est pas toujours approprié d'utiliser une correction post hoc comme le HSD de Tukey lors d'une analyse de variance, surtout lorsque vous aviez une série de comparaisons très claire (et limitée) à laquelle vous teniez avant le début. J'aimerais en parler davantage à l'avenir.

- **Méthodes de comparaison multiples.** Même dans le contexte des tests post hoc et des comparaisons multiples, j'aurais aimé parler plus en détail des méthodes et parler des autres méthodes qui existent à part les quelques options que j'ai mentionnées.

Modèles statistiques manquants dans le livre

La statistique est un domaine énorme. Les outils de base que j'ai décrits dans ce livre (tests du chi carré, tests *t*, régression et ANOVA) sont des outils de base qui sont largement utilisés dans l'analyse quotidienne des données, et ils forment le noyau de la plupart des livres de statistiques d'introduction. Cependant, il existe *bien d'autres* outils. Il y a tellement de situations d'analyse de données que ces outils ne couvrent pas, et ce serait formidable de vous donner une idée de ce qu'il en reste, par exemple :

- **Régression non linéaire.** Lorsque nous avons discuté de la régression au chapitre 12, nous avons vu que la régression suppose que la relation entre les prédicteurs et les résultats est linéaire. D'autre part, lorsque nous avons parlé du problème plus simple de la corrélation au chapitre 4, nous avons vu qu'il existe des outils (p. ex., les corrélations de Spearman) qui permettent d'évaluer les relations non linéaires entre les variables. Il existe un certain nombre d'outils statistiques qui peuvent être utilisés pour effectuer une régression non linéaire. Par exemple, certains modèles de régression non linéaire supposent que la relation entre les prédicteurs et les résultats est monotone (p. ex., régression isotonique), tandis que d'autres supposent qu'elle est lisse mais pas nécessairement monotone (p. ex., régression Lowess), tandis que d'autres supposent que la relation a une forme connue qui est non linéaire (p. ex. régression polynomiale).

- **Régression logistique.** Une autre variation de la régression se produit lorsque la variable de résultat est binaire, mais que les prédicteurs sont continus. Supposons, par exemple, que vous enquêtiez sur les médias sociaux et que vous vouliez savoir s'il est possible de prédire si quelqu'un est sur Twitter en fonction de son revenu, de son âge et d'une série d'autres variables. Il s'agit essentiellement d'un modèle de régression, mais vous ne pouvez pas utiliser la régression linéaire standard parce que la variable résultat est binaire (vous êtes sur Twitter ou vous ne l'êtes pas). Comme la variable résultat est binaire, il n'est pas possible que les résidus puissent être distribués normalement. Les statisticiens peuvent appliquer un certain nombre d'outils à cette situation, dont le plus important est la régression logistique.

- **Le modèle linéaire général (GLM).** Le GLM est en fait une famille de modèles qui inclut la régression logistique, la régression linéaire, (certaines) régressions non linéaires, ANOVA et beaucoup d'autres. L'idée de base du GLM est essentiellement la même que celle qui sous-tend les modèles linéaires, mais elle tient compte de l'idée que vos données pourraient ne pas être normalement distribuées et permet des relations non linéaires entre les prédicteurs et les résultats. Il y a beaucoup d'analyses très pratiques que vous pouvez effectuer dans le cadre du GLM, c'est donc une chose très utile à savoir.

- **Analyse de survie.** Au chapitre 2, j'ai parlé de « l'attrition différentielle », la tendance des gens à quitter l'étude de façon non aléatoire. À l'époque, j'en parlais comme d'un problème méthodologique potentiel, mais il y a beaucoup de situations où l'attrition différentielle est en fait ce qui vous intéresse. Supposons, par exemple, que vous souhaitiez savoir combien de temps les gens jouent à différents types de jeux informatiques au cours d'une même session. Les gens ont-ils tendance à jouer à des jeux RTS (stratégie en temps réel) pour des durées plus longues que les jeux FPS (tir à la première personne) ? Vous pourriez concevoir votre étude comme ceci. Les gens viennent au laboratoire et peuvent jouer aussi longtemps ou aussi peu qu'ils le souhaitent. Une fois qu'ils ont terminé, vous enregistrez le temps qu'ils ont passé à jouer. Cependant, en raison de restrictions éthiques, supposons que vous ne puissiez pas les laisser jouer plus de deux heures. Beaucoup de gens arrêteront de jouer avant la limite de deux heures, donc vous savez exactement combien de temps ils ont joué. Mais certaines personnes se heurteront à la limite de deux heures, et vous ne savez

donc pas combien de temps elles auraient continué à jouer si vous aviez été en mesure de continuer l'étude. En conséquence, vos données sont systématiquement *censurées* : vous manquez toutes les très longues périodes. Comment analysez-vous judicieusement ces données ? C'est le problème que résout l'analyse de survie. Il est spécialement conçu pour faire face à cette situation, où il vous manque systématiquement un « côté » des données parce que l'étude a pris fin. Il est très largement utilisé dans la recherche en santé et, dans ce contexte, il est souvent utilisé littéralement pour analyser la survie. Par exemple, vous pouvez suivre des personnes atteintes d'un type particulier de cancer, certaines qui ont reçu le traitement A et d'autres qui ont reçu le traitement B, mais vous n'avez de financement que pour les suivre pendant 5 ans. À la fin de la période d'étude, certaines personnes sont vivantes, d'autres non. Dans ce contexte, l'analyse de survie est utile pour déterminer quel traitement est le plus efficace et vous informer du risque de décès auquel les gens font face au fil du temps.

- **Modèles mixtes.** Les ANOVA pour mesures répétées sont souvent utilisées dans des situations où vous avez des observations regroupées au sein d'unités expérimentales. Un bon exemple de ceci est lorsque vous suivez des individus à plusieurs reprises dans le temps. Disons que vous suivez le bonheur dans le temps, pour deux personnes. Le bonheur d'Aaron commence à 10 ans, puis descend à 8 ans, puis à 6 ans. Le bonheur de Belinda commence à 6 ans, puis monte à 8 et ensuite à 10 ans. Ces deux personnes ont le même niveau « global » de bonheur (la moyenne sur les trois points dans le temps est de 8), une analyse ANOVA pour mesures répétées traiterait Aaron et Belinda de la même manière. Mais c'est clairement faux. Le bonheur d'Aaron diminue, tandis que celui de Belinda augmente. Si vous voulez analyser de façon optimale les données d'une expérience où les gens peuvent changer au fil du temps, vous avez besoin d'un outil plus puissant que les ANOVA pour mesures répétées. Les outils que les gens utilisent pour résoudre ce problème sont appelés modèles « mixtes », parce qu'ils sont conçus pour apprendre à connaître les unités expérimentales individuelles (par exemple, le bonheur des individus dans le temps) ainsi que les effets globaux (par exemple, l'effet de l'argent sur le bonheur dans le temps). Les ANOVA pour mesures répétées est peut-être l'exemple le plus simple d'un modèle mixte, mais il y a beaucoup de choses que vous pouvez faire avec des modèles mixtes que vous ne pouvez pas faire avec des mesures répétées ANOVA.

- **Mise à l'échelle multidimensionnelle.** L'analyse factorielle est un exemple de modèle « d'apprentissage non supervisé ». Cela signifie que, contrairement à la plupart des outils « d'apprentissage supervisé » que j'ai mentionnés, vous ne pouvez pas diviser vos variables en prédicteurs et en résultats. La régression est un apprentissage supervisé alors que l'analyse factorielle est un apprentissage non supervisé. Ce n'est cependant pas le seul type de modèle d'apprentissage non supervisé. Par exemple, dans l'analyse factorielle, on s'intéresse à l'analyse des corrélations entre les variables. Cependant, il existe de nombreuses situations où vous êtes réellement intéressé à analyser les similitudes ou les dissemblances entre des objets, des objets ou des personnes. Il existe un certain nombre d'outils que vous pouvez utiliser dans cette situation, dont le plus connu est la mise à l'échelle multidimensionnelle (MDS). Dans

MDS, l'idée est de trouver une représentation « géométrique » de vos objets. Chaque élément est « tracé » comme un point dans un certain espace, et la distance entre deux points est une mesure de la dissemblance de ces éléments.

- **Regroupement.** Un autre exemple de modèle d'apprentissage non supervisé est le regroupement (également appelé classification), dans lequel vous voulez organiser tous vos éléments en groupes significatifs, de sorte que des éléments similaires soient affectés aux mêmes groupes. Beaucoup de regroupements ne sont pas supervisés, ce qui signifie que vous ne savez rien de ce que sont les groupes, vous n'avez qu'à deviner. Il y a d'autres situations de « regroupement supervisé » où il faut prédire l'appartenance à un groupe à partir d'autres variables, et ces appartenances à un groupe sont en fait observables. La régression logistique est un bon exemple d'un outil qui fonctionne de cette façon. Cependant, lorsque vous ne connaissez pas vraiment les appartenances des groupes, vous devez utiliser différents outils (par exemple, *k-means* clustering). Il y a même des situations où vous voulez faire quelque chose que l'on appelle « regroupement semi-supervisé », dans lesquelles vous connaissez l'appartenance à un groupe pour certains éléments mais pas pour d'autres. Comme vous pouvez probablement le deviner, le clustering est un sujet assez important, et assez utile à connaitre.

- **Modèles causaux.** Une chose dont je n'ai pas beaucoup parlé dans ce livre est la façon dont vous pouvez utiliser la modélisation statistique pour en apprendre davantage sur les relations causales entre les variables. Par exemple, considérez les trois variables suivantes qui peuvent être intéressantes lorsque vous pensez à la façon dont quelqu'un est mort dans un peloton d'exécution. Nous pourrions vouloir mesurer si un ordre d'exécution a été donné (variable A), si un tireur d'élite a tiré ou non avec son arme (variable B), et si la personne a été touchée ou non par une balle (variable C). Ces trois variables sont toutes corrélées les unes aux autres (c.-à-d. il existe une corrélation entre les armes à feu utilisées et les personnes qui sont frappées par des balles), mais nous voulons en parler de façon plus précise que simplement parler de corrélations. Nous voulons parler de causalité. Nous voulons pouvoir dire que l'ordre d'exécution (A) fait tirer le tireur d'élite (B), ce qui fait que quelqu'un se fait tirer dessus (C). Nous pouvons l'exprimer par une notation de flèche dirigée : nous l'écrivons comme $A \rightarrow B \rightarrow C$. Cette « chaîne causale » est une explication fondamentalement différente des événements que celle dans laquelle le tireur tire d'abord, ce qui provoque le tir $B \rightarrow C$, et ensuite amène le bourreau à émettre « rétroactivement » l'ordre d'exécution, $B \rightarrow A$. Ce modèle « effet commun » dit que A et C sont tous deux causés par B. Vous pouvez voir pourquoi ceux-ci sont différents. Dans le premier modèle causal, si nous avions réussi à empêcher le bourreau d'émettre l'ordre (en intervenant pour changer A), il n'y aurait pas eu de fusillade. Dans le deuxième modèle, le tir se serait produit de toute façon parce que le tireur *ne suivait pas* l'ordre d'exécution. Il existe une abondante documentation statistique sur la façon de comprendre les relations causales entre les variables, et un certain nombre d'outils différents existent pour vous aider à tester différentes hypothèses causales sur vos données. Le plus largement utilisé de ces outils (du moins en psychologie) est la modélisation des équations structurelles (SEM), et à un moment donné, j'aimerais étendre le livre pour en parler.

Bien sûr, même cette liste est incomplète. Je n'ai pas mentionné l'analyse des séries chronologiques, la théorie de la réponse aux questions, l'analyse du panier de consommation, les arbres de classification et de régression, ni aucun autre sujet parmi une vaste gamme d'autres. Cependant, la liste que j'ai donnée ci-dessus est essentiellement ma liste de souhaits pour ce livre. Bien sûr, cela doublerait la longueur du livre, mais cela signifierait que la portée est devenue assez large pour couvrir la plupart des choses que les chercheurs appliqués en psychologie devraient utiliser.

Autres façons de faire des inférences

Une autre raison pour laquelle ce livre est incomplet est qu'il se concentre assez fortement sur une vision très étroite et démodée de la façon dont les statistiques inférentielles devraient être faites. Au chapitre 8, j'ai parlé un peu de l'idée d'estimateurs non biaisés, de distributions d'échantillonnage, etc. Au chapitre 9, j'ai parlé de la théorie des tests de signification des hypothèses nulles et des valeurs *p*. Ces idées existent depuis le début du XXe siècle, et les outils dont j'ai parlé dans le livre s'appuient beaucoup sur les idées théoriques de l'époque. Je me suis senti obligé de m'en tenir à ces sujets parce que la grande majorité de l'analyse des données scientifiques repose également sur ces idées. Cependant, la théorie des statistiques ne se limite pas à ces sujets et, bien que tout le monde devrait les connaître en raison de leur importance pratique, à bien des égards, ces idées ne représentent pas les meilleures pratiques pour l'analyse contemporaine des données. L'une des choses dont je suis particulièrement heureux, c'est que j'ai été capable d'aller un peu plus loin. Le chapitre 15 présente maintenant la perspective bayésienne avec un volume raisonnable de détails, mais le livre dans son ensemble est encore assez fortement orienté vers l'orthodoxie fréquentiste. En outre, il existe un certain nombre d'autres méthodes d'inférence qui méritent d'être mentionnées :

- **Bootstrapping.** Tout au long du livre, chaque fois que j'ai introduit un test d'hypothèse, j'ai eu une forte tendance à faire des affirmations comme « la distribution d'échantillonnage pour BLAH est une distribution *t*» ou quelque chose comme ça. Dans certains cas, j'ai même tenté de justifier cette affirmation. Par exemple, lorsque j'ai parlé des tests χ^2 au chapitre 10, j'ai fait référence à la relation connue entre les distributions normales et les distributions χ^2 (voir chapitre 7) pour expliquer comment nous en arrivons à supposer que la distribution d'échantillonnage de la statistique d'ajustement est χ^2. Cependant, il est également vrai qu'un grand nombre de ces distributions d'échantillonnage sont, eh bien, erronées. Le test χ^2 en est un bon exemple. Elle est basée sur une hypothèse concernant la distribution de vos données, une hypothèse que l'on sait fausse pour des échantillons de petite taille ! Au début du XXe siècle, on ne pouvait pas faire grand-chose contre cette situation. Les statisticiens avaient développé des modèles mathématiques qui disaient que « selon les hypothèses BLAH au sujet des données, la distribution d'échantillonnage est approximativement BLAH », et c'était à peu près le mieux que vous pouviez faire. Souvent, ils n'avaient même pas ça. Il existe de nombreuses situations d'analyse de données pour lesquelles personne n'a trouvé de solution mathématique pour les distributions d'échantillonnage dont vous avez besoin. Ainsi, jusqu'à la fin du XXe siècle, les tests correspondants n'existaient pas ou ne fonctionnaient pas. Cependant, les ordinateurs

ont changé tout cela maintenant. Il y a beaucoup de d'astuces sophistiquées, et certaines moins sophistiqués, que vous pouvez utiliser pour les contourner. Le plus simple d'entre eux est le bootstrapping, et dans sa forme la plus simple c'est incroyablement simple. Ce que vous faites, c'est simuler les résultats de vos expériences à maintes et maintes reprises, en supposant que l'hypothèse nulle est vraie et (b) la distribution inconnue de la population ressemble en fait à celle de vos données brutes. En d'autres termes, au lieu de supposer que les données sont (par exemple) distribuées normalement, supposez simplement que la population ressemble à votre échantillon, puis utilisez des ordinateurs pour simuler la distribution d'échantillonnage pour votre statistique de test si cette hypothèse tient. Bien qu'il repose sur une hypothèse quelque peu douteuse (c.-à-d. que la distribution de la population est la même que celle de l'échantillon !), le bootstrapping est une méthode rapide et facile qui fonctionne remarquablement bien dans la pratique pour de nombreux problèmes d'analyse de données.

- **Validation croisée.** Une question qui surgit de temps en temps dans mes cours de statistiques, habituellement par un étudiant qui essaie d'être provocateur, est « Pourquoi nous soucions-nous des statistiques inférentielles ? Pourquoi ne pas simplement décrire votre échantillon ? » La réponse à la question est généralement la suivante : « Parce que notre véritable intérêt en tant que scientifiques n'est pas l'échantillon spécifique que nous avons observé dans le *passé*, nous voulons faire des prédictions sur les données que nous pourrions observer » à l'*avenir* ». Un grand nombre des problèmes liés à l'inférence statistique découlent du fait que nous nous attendons toujours à ce que l'avenir soit semblable au passé, mais un peu différent. Ou, plus généralement, les nouvelles données ne seront pas tout à fait les mêmes que les anciennes. Ce que nous faisons, dans bien des situations, c'est d'essayer de dériver des règles mathématiques qui nous aident à tirer les inférences qui sont les plus susceptibles d'être correctes pour de nouvelles données, plutôt que de choisir les énoncés qui décrivent le mieux les anciennes données. Par exemple, compte tenu de deux modèles A et B et d'un ensemble de données X que vous avez recueilli aujourd'hui, essayez de choisir le modèle qui décrira le mieux un nouvel ensemble de données Y que vous allez recueillir demain. Parfois, il est pratique de simuler le processus, et c'est ce que fait la validation croisée. Ce que vous faites est de diviser votre ensemble de données en deux sous-ensembles, X1 et X2. Utiliser le sous-ensemble X1 pour former le modèle (par exemple, estimer les coefficients de régression), mais évaluer ensuite la performance du modèle sur X2. Cela vous donne une mesure de la qualité de la *généralisation* du modèle d'un ancien ensemble de données à un nouvel ensemble, et c'est souvent une meilleure mesure de la qualité de votre modèle que si vous l'ajustez simplement à l'ensemble complet de données X.

- **Statistiques robustes.** La vie est désordonnée, et rien ne fonctionne vraiment comme prévu. C'est tout aussi vrai pour les statistiques que pour n'importe quoi d'autre, et lorsque nous essayons d'analyser des données, nous sommes souvent confrontés à toutes sortes de problèmes dans lesquels les données sont tout simplement plus confuses qu'elles ne sont censées l'être. Les variables qui sont censées être distribuées normalement ne sont pas distribuées normalement, les relations qui sont censées être

linéaires ne le sont pas et certaines des observations de votre ensemble de données sont presque certainement de la camelote (c.-à-d. qu'elles ne mesurent pas ce à quoi elles sont censées servir). Tout ce désordre est ignoré dans la plupart des théories statistiques que j'ai développées dans ce livre. Cependant, ignorer un problème ne le résout pas toujours. Parfois, il n'y a pas de mal à ignorer la pagaille, car certains types d'outils statistiques sont « robustes », c'est-à-dire que si les données ne satisfont pas vos hypothèses théoriques, elles fonctionnent tout de même assez bien. D'autres types d'outils statistiques ne sont pas robustes, et même des écarts mineurs par rapport aux hypothèses théoriques entraînent leur rupture. Les statistiques robustes sont une branche des statistiques concernées par cette question, et elles abordent des choses comme le « point de rupture » d'une statistique. En d'autres termes, dans quelle mesure vos données doivent-elles être imparfaites avant que l'on ne puisse faire confiance aux statistiques ? J'en ai parlé à certains endroits. La moyenne n'est pas un estimateur robuste de la tendance centrale d'une variable, mais la médiane l'est. Par exemple, supposons que je vous dise que mes cinq meilleurs amis ont 34, 39, 31, 43 et 4003 ans. Quel âge pensez-vous qu'ils ont en moyenne ? Autrement dit, qu'est-ce que la vraie population signifie ici ? Si vous utilisez la moyenne de l'échantillon comme estimateur de la moyenne de la population, vous obtenez une réponse de 830 ans. Si vous utilisez la médiane de l'échantillon comme estimateur de la moyenne de la population, vous obtenez une réponse de 39 ans. Remarquez que, même si vous faites « techniquement » la mauvaise chose dans le second cas (en utilisant la médiane pour estimer la moyenne !), vous obtenez en fait une meilleure réponse. Le problème ici, c'est que l'une des observations est clairement, évidemment, erronée. Je n'ai pas d'ami âgé de 4003 ans. C'est probablement une faute de frappe, je voulais probablement taper 43. Mais si j'avais tapé 53 au lieu de 43, ou 34 au lieu de 43 ? Pourriez-vous savoir si c'était une faute de frappe ou non ? Parfois, les erreurs dans les données sont subtiles, donc vous ne pouvez pas les détecter simplement en observant l'échantillon, mais ce sont quand même des erreurs qui contaminent vos données, et elles affectent toujours vos conclusions. De statistiques robustes s'intéressent à la façon dont vous pouvez faire des déductions *sûres,* même lorsque vous êtes confronté à une contamination que vous ne connaissez pas. C'est plutôt cool.

Sujets divers

- **Données manquantes.** Supposons que vous faites un sondage et que vous vous intéressez à l'exercice et au poids. Vous envoyez des données à quatre personnes. Adam dit qu'il fait beaucoup d'exercice et qu'il n'est pas en surpoids. Briony dit qu'elle fait beaucoup d'exercice et qu'elle n'est pas en surpoids. Carol dit qu'elle ne fait pas d'exercice et qu'elle a de l'embonpoint. Dan dit qu'il ne fait pas d'exercice et refuse de répondre à la question sur son poids. Elaine ne retourne pas le questionnaire. Vous avez maintenant un problème de données manquantes. Il manque une enquête entière, et une question d'une autre, Que faites-vous pour cela ? Ignorer les données manquantes n'est pas, en général, une solution sûre. Réfléchissons à l'enquête de Dan. Tout d'abord, remarquez que, d'après mes autres réponses, je ressemble plus à Carol (aucun de nous ne fait d'exercice) qu'à Adam ou Briony. Si vous deviez deviner mon poids, vous diriez que je suis plus proche d'elle que d'eux. Vous pourriez peut-être corriger le fait qu'Adam et moi sommes des hommes et que Briony et Carol sont des

femmes. Le nom statistique de ce type de supposition est « imputation ». Il est difficile de procéder à l'imputation en toute sécurité, mais c'est important, surtout lorsque les données manquantes font défaut de façon systématique. Étant donné que les personnes en surpoids sont souvent poussées à se sentir mal par rapport à leur poids (souvent grâce à des campagnes de santé publique), nous avons en fait des raisons de soupçonner que les personnes qui ne répondent pas sont plus susceptibles d'être en surpoids que les personnes qui répondent. Imputer un poids à Dan signifie que le nombre de personnes en surpoids dans l'échantillon passera probablement de 1 sur 3 (si on ignore Dan) à 2 sur 4 (si on impute le poids de Dan). Il est clair que c'est important. Mais le faire raisonnablement est plus compliqué qu'il n'y paraît. Tout à l'heure, je vous ai suggéré de me traiter comme Carol, puisque nous avons donné la même réponse à la question de l'exercice. Mais ce n'est pas tout à fait juste. Il y a une différence systématique entre nous. Elle a répondu à la question, et je ne l'ai pas fait. Étant donné les pressions sociales auxquelles font face les personnes en surpoids, n'est-il pas probable que je sois *plus* obèse que Carol ? Et bien sûr, c'est toujours ignorer le fait qu'il n'est pas raisonnable de m'imputer un poids unique, comme si vous connaissiez réellement mon poids. Au lieu de cela, vous devez imputer une série de suppositions plausibles (appelées imputation multiple), afin de saisir le fait que vous êtes plus incertain au sujet de mon poids que vous ne l'êtes de celui de Carol. Et ne parlons pas du problème posé par le fait qu'Elaine n'a pas envoyé le sondage. Comme vous pouvez probablement le deviner, le traitement des données manquantes est un sujet de plus en plus important. En fait, on m'a dit qu'un grand nombre de revues dans certains domaines n'acceptent pas les études pour lesquelles il manque des données, à moins qu'un système d'imputation multiple raisonnable soit suivi.

- **Analyse de puissance.** Au chapitre 9, j'ai discuté du concept de puissance (c.-à-d., dans quelle mesure êtes-vous pouvez de détecter un effet s'il existe réellement) et j'ai fait référence à l'analyse de la puissance, un ensemble d'outils qui sont utiles pour évaluer la puissance dont dispose votre étude. L'analyse de puissance peut être utile pour planifier une étude (p. ex. pour déterminer la taille de l'échantillon dont vous aurez probablement besoin), mais elle joue également un rôle utile dans l'analyse des données que vous avez déjà recueillies. Supposons, par exemple, que vous obteniez un résultat significatif et que vous ayez une estimation de la taille de votre effet. Vous pouvez utiliser cette information pour estimer la puissance réelle de votre étude. C'est un peu utile, surtout si votre taille d'effet n'est pas grande. Supposons, par exemple, que vous rejetiez l'hypothèse nulle à $p < .05$, mais vous utilisez l'analyse de puissance pour déterminer que votre puissance estimée n'était que de .08. Le résultat significatif signifie que, si l'hypothèse nulle était en fait vraie, il y avait 5 % de chances d'obtenir des données comme celle-ci. Mais la faible puissance signifie que, même si l'hypothèse nulle est fausse et que la taille de l'effet était aussi petite qu'elle en a l'air, il n'y avait que 8% de chances d'obtenir des données comme les vôtres. Cela suggère que vous devez être assez prudent, parce que le hasard semble avoir joué un grand rôle dans vos résultats, d'une façon ou d'une autre !

- **Analyse des données à l'aide de modèles inspirés de la théorie.** À plusieurs endroits dans ce livre, j'ai mentionné les données sur le temps de réponse (RT), où l'on

enregistre le temps qu'il faut à quelqu'un pour faire quelque chose (p. ex., prendre une décision simple). J'ai mentionné que les données de la TR sont presque invariablement non-normales et faussées de façon positive. De plus, il y a un compromis connu sous le nom de compromis sur la précision de la vitesse : si vous essayez de prendre des décisions trop rapidement (RT faible), vous risquez de prendre de moins bonnes décisions (précision plus faible). Donc, si vous mesurez à la fois l'exactitude des décisions d'un participant et sa RT, vous constaterez probablement que vitesse et précision sont liées. Il y a plus que cela, bien sûr, parce que certaines personnes prennent de meilleures décisions que d'autres, quelle que soit la vitesse à laquelle elles vont. De plus, la vitesse dépend à la fois des processus cognitifs (c.-à-d. le temps passé à penser) et des processus physiologiques (p. ex. à quelle vitesse pouvez-vous bouger vos muscles). Il semble que l'analyse de ces données sera un processus compliqué. Et c'est effectivement le cas, mais l'une des choses que l'on trouve en fouillant dans la littérature psychologique, c'est qu'il existe déjà des modèles mathématiques (appelés « modèles d'échantillonnage séquentiel ») qui décrivent comment les gens prennent des décisions simples, et ces modèles prennent en compte un grand nombre des facteurs que j'ai mentionnés ci-dessus. Vous ne trouverez aucun de ces modèles inspirés de la théorie dans un manuel de statistiques standard. Les manuels de statistiques standard décrivent des outils standard, des outils qui pourraient être appliqués de manière significative dans un grand nombre de disciplines différentes, et pas seulement en psychologie. L'ANOVA est un exemple d'outil standard qui s'applique aussi bien à la psychologie qu'à la pharmacologie. Les modèles d'échantillonnage séquentiel ne le sont pas, ils sont plus ou moins spécifiques à la psychologie. Cela ne les rend pas moins puissants. En fait, si vous analysez des données où les gens doivent faire des choix rapidement, vous devriez vraiment utiliser des modèles d'échantillonnage séquentiels pour analyser les données. L'utilisation de l'analyse de variance, de la régression ou de tout autre méthode ne fonctionnera pas aussi bien, car les hypothèses théoriques qui les sous-tendent ne correspondent pas bien à vos données. En revanche, les modèles d'échantillonnage séquentiel ont été explicitement conçus pour analyser ce type spécifique de données, et leurs hypothèses théoriques sont *extrêmement* bien adaptées aux données.

Apprendre les bases, et les apprendre avec Jamovi

Bien, c'était une longue liste. Et même cette liste est largement incomplète. Il y a vraiment *beaucoup* de grandes idées statistiques que je n'ai pas couvertes dans ce livre. Il peut sembler assez déprimant de terminer un manuel de près de 500 pages pour se faire dire que ce n'est que le début, surtout quand on commence à soupçonner que la moitié des choses qu'on vous a enseignées sont erronées. Par exemple, il y a beaucoup de gens sur le terrain qui s'opposeraient fortement à l'utilisation du modèle classique ANOVA, mais j'y ai consacré deux chapitres entiers ! L'analyse de variance standard peut être attaquée d'un point de vue bayésien, ou du point de vue des statistiques robustes, ou même parce que « c'est tout simplement faux » (les gens utilisent très souvent ANOVA alors qu'ils devraient utiliser des modèles mixtes). Alors pourquoi l'apprendre ?

Selon moi, il y a deux arguments clés. Premièrement, il y a l'argument du pur pragmatisme. ANOVA est largement utilisé, à tort ou à raison. Si vous voulez comprendre la littérature scientifique, vous devez comprendre l'ANOVA. Et deuxièmement, il y a l'argument de la « connaissance incrémentale ». De la même façon qu'il était pratique d'avoir vu l'ANOVA à un facteur avant d'essayer d'apprendre l'ANOVA factorielle, comprendre l'ANOVA est utile pour comprendre des outils plus avancés, car beaucoup de ces outils prolongent ou modifient d'une certaine façon l'installation ANOVA fondamentale. Par exemple, bien que les modèles mixtes soient beaucoup plus utiles que l'analyse de variance et la régression, je n'ai jamais entendu parler de quelqu'un qui apprend comment fonctionnent les modèles mixtes sans avoir d'abord travaillé sur l'analyse de variance et la régression. Il faut apprendre à ramper avant de pouvoir gravir une montagne.

En fait, j'aimerais pousser ce point un peu plus loin. Une chose que j'ai souvent faite dans ce livre, c'est de parler des principes fondamentaux. J'ai passé beaucoup de temps sur la théorie des probabilités. J'ai parlé de la théorie de l'estimation et des tests d'hypothèse plus en détail que nécessaire. Pourquoi ai-je fait tout ça ? En y repensant, vous pourriez me demander si j'avais vraiment *besoin* de passer tout ce temps à parler de ce qu'est une distribution de probabilités, ou pourquoi il y avait même une section sur la densité de probabilité. Si le but du livre était de vous apprendre à faire un *test t* ou une ANOVA, est-ce que tout cela était vraiment nécessaire ? Tout ça n'était qu'une énorme perte de temps pour tout le monde ?

La réponse, j'espère que vous serez d'accord, est non. Le but d'une introduction en statistique *n'est pas* d'enseigner l'ANOVA. Ce n'est pas non plus enseigner les *tests t*, les régressions, les histogrammes ou les valeurs *p*. L'objectif est de vous mettre sur la voie qui vous mènera à devenir un analyste de données compétent. Et pour devenir un analyste de données compétent, vous devez être capable de faire plus que l'ANOVA, plus que des tests *t*, des régressions et des histogrammes. Vous devez être capable de *penser correctement* aux données. Vous devez être en mesure d'apprendre les modèles statistiques plus avancés dont j'ai parlé dans la dernière section et de comprendre la théorie sur laquelle ils sont fondés. Et vous devez avoir accès à un logiciel qui vous permettra d'utiliser ces outils avancés. Et c'est là que, à mon avis du moins, tout le temps supplémentaire que j'ai passé sur les fondamentaux est payant. Si vous comprenez la théorie des probabilités, il vous sera beaucoup plus facile de passer des analyses fréquentistes aux analyses bayésiennes.

Bref, je pense que l'*extensibilité* est le gros avantage d'apprendre les statistiques de cette façon. Pour un livre qui ne couvre que les bases mêmes de l'analyse des données, ce livre a une énorme surcharge en termes d'apprentissage de la théorie des probabilités et ainsi de suite. Il y a beaucoup d'autres choses qu'il vous pousse à apprendre en plus des analyses spécifiques que le livre couvre. Donc, si votre but avait été d'apprendre à exécuter une ANOVA en un minimum de temps, eh bien, ce livre n'était pas un bon choix. Mais comme je l'ai dit, je ne pense pas que ce soit votre but. Je pense que vous voulez apprendre à analyser les données. Et si c'est vraiment votre objectif, vous voulez vous assurer que les compétences que vous apprenez dans votre cours d'introduction aux statistiques sont naturellement et proprement transférables aux modèles plus complexes dont vous avez besoin dans l'analyse de données du monde réel. Vous voulez vous assurer d'apprendre à utiliser les mêmes outils que les vrais analystes de données, de sorte que vous puissiez

apprendre à faire ce qu'ils font. Et bien, d'accord, vous êtes un débutant pour le moment (ou vous l'étiez quand vous avez commencé ce livre), mais cela ne veut pas dire qu'on devrait vous donner une version édulcorée, une version où je ne vous raconte rien de la densité de probabilité ou une version où je vous parle de ce qui constitue le cauchemar d'une ANOVA factorielle aux plans non équilibrés. Et cela ne signifie pas qu'il faille vous donner des jouets pour bébés au lieu d'outils d'analyse de données appropriés. Les débutants ne sont pas muets, ils manquent simplement de connaissances. Ce dont vous avez besoin, c'est de ne pas avoir à vous cacher les complexités de l'analyse des données réèlles. Ce dont vous avez besoin, ce sont les compétences et les outils qui vous permettront de gérer ces complexités lorsqu'elles vous tendront inévitablement une embuscade dans le monde réel.

Et ce que j'espère, c'est que ce livre, ou le livre fini que cela deviendra un jour, pourra vous aider à le faire.

Note de l'auteur - Je l'ai déjà mentionné auparavant, mais je vais rapidement le mentionner à nouveau. Cette liste de références est épouvantablement incomplète. Ne présumez pas que ce sont les seules sources sur lesquelles j'ai compté. La version finale de ce livre aura *beaucoup* plus de références. Et si vous voyez quelque chose d'intelligent dans ce livre qui ne semble pas avoir une référence, je peux vous promettre que l'idée était celle de quelqu'un d'autre. Il s'agit d'un manuel d'introduction : *aucune des idées n'est originale*. J'assumerai la responsabilité de toutes les erreurs, mais je ne peux m'attribuer le mérite d'aucune des bonnes choses. Tout ce qu'il y a d'intelligent dans ce livre vient de quelqu'un d'autre, et ils méritent tous d'être reconnus pour leur excellent travail. Je n'ai pas encore eu l'occasion de le leur rendre.

Adair, John G. 1984. "The Hawthorne Effect: A Reconsideration of the Methodological Artifact." *Journal of Applied Psychology* 69 (2): 334–45. doi:10.1037/0021-9010.69.2.334.

Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

———. 2013. *Categorical Data Analysis*. Third. Wiley.

Akaike, H. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6): 716–23. doi:10.1109/TAC.1974.1100705.

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. doi:10.1080/00031305.1973.10478966.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187 (4175): 398–404. doi:10.1126/science.187.4175.398.

Box, G. E. P. 1953. "Non-Normality and Tests on Variances." *Biometrika* 40 (3/4): 318–35. doi:10.2307/2333350.

Box, George E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71 (356): 791–99. doi:10.1080/01621459.1976.10480949.

Box, Joan Fisher. 1987. “Guinness, Gosset, Fisher, and Small Samples.” *Statistical Science* 2 (1): 45–52. doi:10.1214/ss/1177013437.

Brown, Morton B., and Alan B. Forsythe. 1974. “Robust Tests for the Equality of Variances.” *Journal of the American Statistical Association* 69 (346): 364–67. doi:10.2307/2285659.

Campbell, Donald Thomas, and Julian Cecil Stanley. 1967. *Experimental and Quasi-Experimental Designs for Research*. 2. print; Reprinted from “Handbook of research on teaching”. Boston: Houghton Mifflin Comp.

Cochran, William G. 1952. “The *X*2 Test of Goodness of Fit.” *The Annals of Mathematical Statistics* 23 (3): 315–45.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates.

Cramer, Harald. 1999. *Mathematical Methods of Statistics*. 19. printing. Princeton Landmarks in Mathematics and Physics. Princeton: Princeton Univ. Press.

Cronbach, Lee J. 1951. “Coefficient Alpha and the Internal Structure of Tests.” *Psychometrika* 16 (3): 297–334. doi:10.1007/BF02310555.

Dunn, Olive Jean. 1961. “Multiple Comparisons Among Means.” *Journal of the American Statistical Association* 56 (293): 52–64. doi:10.1080/01621459.1961.10482090.

Ellis, Paul D. 2010. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press.

Ellman, Michael. 2002. “Soviet Repression Statistics: Some Comments.” *Europe-Asia Studies* 54 (7): 1151–72.

Evans, J. St. B. T., Julie L. Barston, and Paul Pollard. 1983. “On the Conflict Between Logic and Belief in Syllogistic Reasoning.” *Memory & Cognition* 11 (3): 295–306. doi:10.3758/BF03196976.

Everitt, Brian S. 1996. *Making Sense of Statistics in Psychology: A Second-Level Course*. Making Sense of Statistics in Psychology: A Second-Level Course. New York, NY, US: Oxford University Press.

Fisher, R. A. 1922. “On the Interpretation of *X*2 from Contingency Tables, and the Calculation of P.” *Journal of the Royal Statistical Society* 85 (1): 87–94. doi:10.2307/2340521.

Fox, John, and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. 2nd Revised edition. Thousand Oaks, Calif: SAGE Publications, Inc.

Gelman, Andrew, and Eric Loken. 2014. “The Statistical Crisis in Science.” *American Scientist* 102 (November): 460. doi:10.1511/2014.111.460.

Gelman, Andrew, and Hal Stern. 2006. “The Difference Between ‘Significant’ and ‘Not Significant’ Is Not Itself Statistically Significant.” *The American Statistician* 60 (4): 328–31. doi:10.1198/000313006X152649.

Geschwind, Norman. 1972. “Language and the Brain.” *Scientific American* 226 (4): 76–83.

Hedges, Larry V. 1981. “Distribution Theory for Glass’s Estimator of Effect Size and Related Estimators.” *Journal of Educational Statistics* 6 (2): 107–28. doi:10.3102/10769986006002107.

Hedges, Larry V., and Ingram Olkin. 2014. *Statistical Methods for Meta-Analysis*. Academic Press.

Hewitt, Anthea K., David R. Foxcroft, and John MacDonald. 2004. “Multitrait-Multimethod Confirmatory Factor Analysis of the Attributional Style Questionnaire.” *Personality and Individual Differences* 37 (7): 1483–91. doi:10.1016/j.paid.2004.02.005.

Hogg, Robert V., and Allen T. Craig. 2005. *Introduction to Mathematical Statistics*. 6th ed. New York: Pearson.

Holm, Sture. 1979. “A Simple Sequentially Rejective Multiple Test Procedure.” *Scandinavian Journal of Statistics* 6 (2): 65–70.

Hothersall, David. 2004. *History of Psychology*. New York: McGraw-Hill.

Hsu, Jason. 1996. *Multiple Comparisons: Theory and Methods*. London: Chapman and Hall/CRC.

Ioannidis, John P A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2 (8): 6.

Jeffreys, Sir Harold. 1998. *The Theory of Probability*. Third Edition. Oxford Classic Texts in the Physical Sciences. Oxford, New York: Oxford University Press.

Kahneman, Daniel, and Amos Tversky. 1973. “On the Psychology of Prediction.” *Psychological Review* 80 (4): 237–51. doi:10.1037/h0034747.

Kass, Robert E., and Adrian E. Raftery. 1995. “Bayes Factors.” *Journal of the American Statistical Association* 90 (430): 773–95. doi:10.2307/2291091.

Keynes, John Maynard. 2009. *A Tract on Monetary Reform*. Place of publication not identified: WWW.Therichestmaninbabylon.Org.

Kruschke, John K. 2011. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press.

Kruskal, William H., and W. Allen Wallis. 1952. “Use of Ranks in One-Criterion Variance Analysis.” *Journal of the American Statistical Association* 47 (260): 583–621. doi:10.2307/2280779.

Kühberger, Anton, Astrid Fritz, and Thomas Scherndl. 2014. “Publication Bias in Psychology: A Diagnosis Based on the Correlation Between Effect Size and Sample Size.” *PLOS ONE* 9 (9): e105825. doi:10.1371/journal.pone.0105825.

Larntz, Kinley. 1978. “Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics.” *Journal of the American Statistical Association* 73 (362): 253–63. doi:10.2307/2286650.

Lee, Michael D., and Eric-Jan Wagenmakers. 2014. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lehmann, Erich L. 2011. *Fisher, Neyman, and the Creation of Classical Statistics*. New York: Springer-Verlag. doi:10.1007/978-1-4419-9500-1.

Levene, H. 1960. “Robust Tests for Equality of Variances.” In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, edited by I. Oklin, Sudhish G. Ghurye, w; Hoeffding, W.G. Madow, and Henry B. Mann, 278–92. Palo Alto, CA: Stanford University Press.

McGrath, Robert E., and Gregory J. Meyer. 2006. “When Effect Sizes Disagree: The Case of R and d.” *Psychological Methods* 11 (4): 386–401. doi:10.1037/1082-989X.11.4.386.

McNEMAR, Q. 1947. “Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages.” *Psychometrika* 12 (2): 153–57. doi:10.1007/bf02295996.

Meehl, Paul E. 1967. “Theory-Testing in Psychology and Physics: A Methodological Paradox.” *Philosophy of Science* 34 (2): 103–15.

Navarro, Daniel Joseph. 2014. *Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners*. Université d’Adelaide,

Navarro, Danielle J, David R Foxcroft, and Thomas J Faulkenberry. 2019. *Learning Statistics with JASP*.

Pearson, Karl. 1900. “X. on the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling.” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50 (302): 157–75. doi:10.1080/14786440009463897.

Peterson, Christopher, and Martin E P Seligman. 1984. “Causal Explanations as a Risk Factor for Depression: Theory and Evidence,” 28.

Pfungst, Oskar. 1911. *Clever Hans (the Horse of Mr. von Osten) a Contribution to Experimental Animal and Human Psychology,* New York, H. Holt and company.

Sahai, Hardeo, and Mohammed I. Ageel. 2000. *The Analysis of Variance: Fixed, Random and Mixed Models*. Birkhäuser Basel. doi:10.1007/978-1-4612-1344-4.

Shaffer, J P. 1995. “Multiple Hypothesis Testing.” *Annual Review of Psychology* 46 (1): 561–84. doi:10.1146/annurev.ps.46.020195.003021.

Shapiro, S. S., and M. B. Wilk. 1965. “An Analysis of Variance Test for Normality (Complete Samples).” *Biometrika* 52 (3/4): 591–611. doi:10.2307/2333709.

Stigler, Sm. 1986. *The History of Statistics of Uncertainty Before 1900*. Reprint. Cambridge, Mass.: Harvard University Press.

Student, A. 1908. “The Probable Error of a Mean.” *Biometrika* 6 (1): 1–25. doi:10.1093/biomet/6.1.1.

Welch, B. L. 1947. “The Generalisation of Student’s Problems When Several Different Population Variances Are Involved.” *Biometrika* 34 (1-2): 28–35. doi:10.1093/biomet/34.1-2.28.

———. 1951. “On the Comparison of Several Mean Values: An Alternative Approach.” *Biometrika* 38 (3/4): 330–36. doi:10.2307/2332579.

Wilkinson, Leland. 2005. *The Grammar of Graphics*. Second. Statistics and Computing. New York: Springer-Verlag. doi:10.1007/0-387-28695-0.

Yates, F. 1934. “Contingency Tables Involving Small Numbers and the *X*2 Test.” *Supplement to the Journal of the Royal Statistical Society* 1 (2): 217–35. doi:10.2307/2983604.