Tests inférentiels bivariés

Manuel Gimenes

Partie 2
Chapitre 7 : Tests inférentiels bivariés

Manuel Gimenes
Université de Poitiers, France
manuel.gimenes@univ-poitiers.fr

Résumé : L'objectif de ce chapitre est de présenter les tests inférentiels bivariés (impliquant deux variables) classiquement utilisés dans les recherches expérimentales en sciences du langage. Le test de Student sera tout d'abord abordé, suivi de l'ANOVA et du χ^2 d'indépendance. Pour terminer, le test de corrélation sera expliqué. Pour chaque test, les éléments suivants seront précisés : explication théorique, conditions d'application, alternatives non paramétriques, exemple de script R. Le même jeu de données sera proposé pour toutes les analyses de ce chapitre.

Introduction

L'objectif de ce chapitre est de présenter plusieurs tests inférentiels classiquement utilisés dans les recherches expérimentales en sciences du langage. Les tests présentés ont pour point commun de nécessiter deux variables (ce sont des tests bivariés) mais ils diffèrent quant à la nature des variables impliquées (on parle d'échelles de mesure, voir encadré 1).

Encadré 1 : les échelles de mesure. Les valeurs attribuées à une variable peuvent correspondre à différentes échelles de mesure. On trouve tout d'abord l'échelle nominale : les valeurs sont des catégories, des labels. Par exemple, la couleur des yeux peut être considérée comme une variable nominale : un individu peut avoir les yeux marrons, verts ou bleus. Il existe également l'échelle ordinale : les valeurs sont aussi des catégories mais on peut les ordonner. On peut donner comme exemple les grades de l'armée française : un capitaine a un grade supérieur au lieutenant, lui-même étant supérieur au sergent. Les variables nominales et ordinales peuvent être regroupées sous le nom de variables qualitatives (on parle aussi de variables catégorielles). On trouve ensuite l'échelle d'intervalle : les valeurs sont des nombres et l'intervalle entre deux valeurs successives est considéré comme étant toujours le même. La température en degrés Celsius est un exemple d'échelle d'intervalle. Enfin, il existe l'échelle de rapport : elle est similaire à l'échelle d'intervalle mais possède un zéro absolu. Cela signifie que la valeur zéro correspond à l'absence de ce que la variable mesure (la température dans l'exemple d'échelle ordinale n'a pas un zéro absolu : la valeur zéro correspond à une valeur de température parmi d'autres et non pas à l'absence de température). La taille d'un objet est un exemple d'échelle de rapport : une taille de zéro signifie l'absence de taille. Les variables d'intervalle et de rapport peuvent être regroupées sous le terme de variables quantitatives (on parle aussi de variables numériques). Dans ce chapitre, seule la distinction entre variables qualitatives et quantitatives sera utilisée.

Le test de Student implique une variable dépendante numérique et une variable indépendante catégorielle à deux modalités. Ce test permet de voir si une différence entre deux moyennes est significative. L'ANOVA peut être considérée comme une extension du test de Student et implique une variable numérique et une variable catégorielle pouvant avoir trois modalités ou plus. Le test du χ^2 d'indépendance a pour objectif de voir s'il existe un lien entre deux variables catégorielles. Enfin, l'analyse de corrélation permet de tester si une relation existe entre deux variables numériques. Les tests inférentiels présentés sont des tests paramétriques, c'est-à-dire que les hypothèses nulle et alternative portent sur un paramètre statistique (la moyenne par exemple). Cela a pour conséquence que certaines conditions d'application sont nécessaires pour pouvoir les utiliser. Si ces conditions ne sont pas respectées, il est possible d'utiliser des tests non-paramétriques. Pour chaque test paramétrique présenté dans ce

chapitre, les conditions d'application seront précisées et des alternatives non-paramétriques seront proposées.

Tout au long de ce chapitre, un seul jeu de données sera utilisé. Les données se trouvent dans le fichier nommé « lexique_megalex.csv ». Ces données sont issues du site web lexique.org (New et al., 2004) qui regroupe plusieurs bases de données lexicales. La base de données principale se nomme « Lexique » et fournit pour 142 694 mots de la langue française, diverses informations comme le nombre de lettres, le nombre de phonèmes ou la fréquence. Parmi les autres bases de données proposées, une va nous intéresser particulièrement : il s'agit de « Megalex-visual » (Ferrand et al., 2018) qui fournit des temps de décision lexicale pour 28 466 mots. La tâche de décision lexicale est très utilisée dans les recherches en psycholinguistique. Le principe est le suivant : on présente à des participants des mots un par un sur un écran d'ordinateur. Certains mots sont des mots français (par exemple « maison »), d'autres sont des pseudo-mots (par exemple « maidon »). Si le mot présenté est un mot français, le participant doit appuyer sur la touche « oui » le plus rapidement possible. Si le mot présenté est un pseudo-mot, il doit appuyer sur la touche « non ». Les temps de réponse sont mesurés en millisecondes et sont appelés « temps de décision lexicale ». La base de données « Megalex-visual » regroupe les temps de décision lexicale pour 28 466 mots. L'intérêt du site web lexique.org est qu'il permet d'associer plusieurs bases de données. Ainsi, nous pouvons associer « Lexique » et « Megalex-visual » ce qui fournit un tableau avec 33 811 lignes. Pour chaque mot nous avons ainsi les informations issues des deux bases de données. On peut remarquer qu'il y a plus de lignes que dans la base « Megalex-visual » seule : c'est normal car dans « Lexique », un mot peut correspondre à plusieurs catégories grammaticales (par exemple, le mot « danse » peut être un nom et aussi un verbe). Dans ce cas, il y a deux lignes pour ce mot (une ligne pour chaque catégorie grammaticale). Mais le temps de décision lexicale sera le même pour les deux lignes. Dans ce chapitre, tous les exemples seront basés sur ce jeu de données contenant 33 811 lignes. Pour lire et comprendre le code R décrit dans ce chapitre, vous pouvez vous référer à la partie introduction à R de ce livre (chapitre 5 dans la partie 2).

1. Importation du jeu de données et découverte des variables

Commençons par importer les données dans R. La ligne de code suivante permet d'importer les données et de les mettre dans un objet de type data.frame nommé « dat » :

```
dat <- read.csv2("lexique_megalex.csv")
```

La fonction read.csv2 est utilisée ici car dans le fichier de données, le séparateur décimal est la virgule. Si le point avait été choisi comme séparateur décimal, nous aurions pu utiliser la fonction « read.csv ». Nous allons maintenant explorer cet objet « dat » en utilisant la fonction str (qui signifie « structure ») :

```
str(dat)
```

Voici le résultat affiché dans la console R :

```
'data.frame': 33811 obs. of  8 variables:
 $ mot     : chr  "abaissa" "abaissait" "abaissant" "abaissant" ...
 $ cgram   : chr  "VER" "VER" "ADJ" "VER" ...
 $ genre   : chr  "" "" "m" "" ...
```

```
$ nombre   : chr  "" "" "s" "" ...
$ nblettres: int  7 9 9 9 7 7 7 7 8 8 ...
$ nbphons  : int  5 5 5 5 4 4 5 5 5 5 ...
$ nbsyll   : int  3 3 3 3 2 2 3 3 3 3 ...
$ rt       : num  616 684 694 694 614 ...
```

R nous indique que le tableau contient 33 811 lignes et 8 colonnes. Chaque ligne correspond à un mot et les colonnes correspondent à différentes caractéristiques pour chaque mot. La première colonne « mot » indique tout simplement la façon dont s'orthographie le mot. La deuxième colonne « cgram » correspond à la catégorie grammaticale du mot. Pour avoir la liste des différentes catégories grammaticales possibles, il suffit d'écrire :

```
unique(dat$cgram)
```

La fonction unique liste toutes les valeurs strictement différentes dans la colonne. Cela revient donc à demander les différentes modalités de la variable « cgram ». Les résultats nous indiquent les 21 catégories possibles.
La colonne « genre » indique si le mot est masculin ou féminin. Pour certains mots, l'information n'est pas précisée car ce n'est pas pertinent pour certaines catégories grammaticales (par exemple pour les verbes). Il en est de même pour la colonne suivante (« nombre ») qui indique si le mot est singulier ou pluriel. Les trois colonnes suivantes précisent le nombre de lettres, le nombre de phonèmes et le nombre de syllabes. Enfin, la dernière colonne indique le temps de décision lexicale moyen pour chaque mot en millisecondes.
Dans ce jeu de données, certaines variables sont numériques (par exemple le nombre de lettres ou les temps de décision lexicale) et d'autres sont catégorielles (par exemple la catégorie grammaticale ou le genre). Cela va nous permettre d'aborder différents tests inférentiels. Nous allons commencer avec le test de Student.

2. Test de Student

Le test de Student (appelé aussi « test t » ou « t-test ») a été proposé au début du 20ème siècle par William Gosset (il a publié ses travaux sous le pseudonyme « Student »). C'est un test de comparaison de moyennes et son principe est relativement simple : il s'agit de tester si la différence moyenne observée entre deux échantillons est significativement différente de 0. Si c'est le cas, on pourra conclure que la différence entre les deux moyennes est significative. Le test de Student peut être utilisé pour vérifier si la moyenne d'un échantillon diffère d'une norme (t-test de comparaison à une norme), si deux échantillons diffèrent l'un de l'autre (t-test pour échantillons indépendants) ou s'il existe une différence significative entre deux mesures répétées dans un même échantillon (t-test pour échantillons appariés). Dans ce chapitre, nous aborderons uniquement le t-test pour échantillons indépendants (voir encadré 2 pour une distinction entre échantillons indépendants et appariés).

Encadré 2 : échantillons indépendants et appariés. Deux échantillons sont indépendants si les éléments dans les échantillons ne sont pas les mêmes. Les éléments en question peuvent être des mots, des participants ou toute autre chose. Par exemple, pour tester l'effet d'un médicament, il est possible de comparer un groupe expérimental (les participants reçoivent le médicament) et un groupe contrôle (les participants reçoivent un placebo). On dira alors que les deux groupes sont indépendants car ce ne sont pas les mêmes participants dans les deux groupes. A l'inverse, deux échantillons sont appariés si les mêmes éléments sont mesurés dans les deux échantillons. Imaginons que l'on mesure un seul groupe de participants avant et après la prise d'un médicament : chaque participant est mesuré deux

fois et on dira alors que les échantillons (avant et après la prise du médicament) sont appariés. La distinction entre échantillons indépendants et appariés est importante car les tests inférentiels à utiliser peuvent être différents.

La question à laquelle nous tenterons de répondre est la suivante : le temps de décision lexicale (variable « rt ») est-il significativement différent selon que le mot est masculin ou féminin (variable « genre ») ?
Nous allons tout d'abord créer un nouvel objet « dat_nom », dans lequel nous allons sélectionner uniquement les mots appartenant à la catégorie grammaticale « NOM » :

```
dat_nom <- dat[dat$cgram == "NOM", ]
```

Cette ligne de code crée l'objet « dat_nom » et lui attribue une partie de l'objet « dat ». Plus précisément, les crochets après « dat » permettent d'indiquer les lignes et les colonnes que l'on souhaite garder. On peut remarquer qu'il y a une virgule entre les crochets : tout ce qui est placé à gauche de la virgule concerne les lignes, tout ce qui est à droite de la virgule concerne les colonnes. Rien n'est écrit à droite de la virgule ce qui veut dire que nous gardons toutes les colonnes de l'objet « dat ». En revanche, il y a du code juste avant la virgule, ce qui signifie que nous souhaitons filtrer certaines lignes. Plus exactement, nous indiquons que nous souhaitons garder uniquement les lignes pour lesquelles les mots sont des noms.
Nous vérifions ensuite le nombre de modalités pour la variable « genre » :

```
unique(dat_nom$genre)
```

On peut voir qu'il y a trois modalités : « féminin », « masculin » et « vide ». La modalité « vide » est utilisée pour les mots qui peuvent être à la fois féminin et masculin. Par exemple, on peut dire « un acrobate » et « une acrobate ». Le mot « acrobate » a donc la modalité « vide » dans la colonne « genre ». Nous allons exclure les mots pour lesquels le genre n'est pas précisé et par conséquent nous allons garder uniquement les mots qui sont soit masculin soit féminin :

```
dat_nom <- dat_nom[dat_nom$genre == "f" | dat_nom$genre == "m",  ]
```

Nous pouvons vérifier en lançant de nouveau la ligne de code suivante :

```
unique(dat_nom$genre)
```

Il y a bien seulement deux modalités : « féminin » et « masculin ».
Enfin, nous sélectionnons uniquement les mots au singulier :

```
dat_nom <- dat_nom[dat_nom$nombre == "s", ]
```

Le tableau « dat_nom » contient maintenant 6729 mots. Tout est prêt pour pouvoir faire le test de Student en utilisant la fonction t.test :

```
t.test(dat_nom$rt ~ dat_nom$genre)
```

Les résultats du test de Student sont présentés ci-dessous :

```
        Welch Two Sample t-test

data:  dat_nom$rt by dat_nom$genre
```

```
t = 3.9219, df = 6465.4, p-value = 8.878e-05
alternative hypothesis: true difference in means between group f and group m is not equal to 0
95 percent confidence interval:
 2.873690 8.617585
sample estimates:
mean in group f mean in group m
      595.0162       589.2705
```

La fonction t.test prend deux arguments : d'abord la variable qu'on veut expliquer (ici le temps de décision lexicale) et ensuite l'autre variable (ici le genre). Dit autrement, on veut expliquer le temps de décision lexicale en fonction du genre.
Puisque le test de Student est un test inférentiel, une hypothèse nulle est testée. Dans notre exemple, l'hypothèse nulle est la suivante : les deux moyennes (temps de réponse moyen pour les noms masculins et temps de réponse moyen pour les noms féminins) sont égales dans la population parente. Les résultats du test nous indiquent les deux moyennes dans nos échantillons : le temps de réponse moyen est de 595 ms pour les noms féminins et de 589 ms pour les noms masculins. Les deux moyennes sont différentes dans nos échantillons mais cette différence est-elle due au hasard ou bien au genre des mots ? le test de Student nous permet de répondre à cette question. Les résultats du test indiquent que la différence est significative : $t(6465.4) = 3.92$, $p < .001$. La valeur de la *p-value* est tellement petite que R nous indique sa valeur en écriture scientifique. Ainsi, 8.878e-05 signifie que la *p-value* est égale à 0.00008878. Donc la *p-value* est bien inférieure au seuil .05, ce qui permet de rejeter l'hypothèse nulle et dire que la différence entre les deux moyennes est significative. Le temps de réponse est significativement plus long pour les noms féminins que pour les noms masculins.

Si on regarde attentivement, dans les faits les moyennes sont quand même très proches (6 ms de différence seulement). Le fait que la différence soit significative vient du fait que nous avons un échantillon très important. En effet, plus l'effectif est important et plus une différence a de chances d'être significative. C'est pourquoi il faut distinguer la significativité statistique (ce dont nous venons de parler avec la *p-value*) et la significativité pratique. Cette dernière est importante car même si un effet est statistiquement significatif, il peut être est très faible d'un point de vue pratique (comme c'est le cas ici). Dans ce cas, il faut faire attention à ne pas surestimer les résultats. Une manière d'estimer la significativité pratique est de rapporter une taille d'effet : des indicateurs existent pour les différents tests et permettent de dire si un effet est plus ou moins fort (quel que soit son niveau de significativité statistique). Pour le test de Student, l'indicateur de taille d'effet le plus souvent rapporté est le d de Cohen. Pour le calculer, nous devons installer un nouveau package nommé *effectsize* :

```
install.packages("effectsize")
```

Il faut ensuite charger le package pour pouvoir l'utiliser :

```
library(effectsize)
```

Maintenant que le package est chargé, nous pouvons utiliser la fonction cohens_d contenue dans ce package pour calculer la taille d'effet :

```
cohens_d(dat_nom$rt[dat_nom$genre == "f"], dat_nom$rt[dat_nom$genre == "m"])
```

La fonction cohens_d prend deux arguments : la liste des temps de décision lexicale de tous les mots féminins et la liste des temps de décision lexicale de tous les mots masculins. Les résultats indiquent que le d de Cohen dans notre exemple est égal à 0.10. Comment interpréter cette valeur ? Tout d'abord, il faut savoir que plus la valeur est élevée, plus l'effet est fort. Si d'autres études publiées (et s'intéressant à la même problématique) ont déjà calculé un d de Cohen, alors on peut les comparer avec la valeur obtenue. Si nous avons des difficultés pour interpréter la taille d'effet, nous pouvons utiliser une autre solution. En effet, Cohen (1988) a proposé des seuils pour aider à l'interprétation : un d de Cohen dont la valeur est supérieure à 0.2 indique un petit effet. Une valeur supérieure à 0.5 indique un effet moyen et une valeur supérieure à 0.8 indique un effet fort. Dans notre exemple, la valeur est inférieure à 0.2, donc nous ne pouvons même pas dire que nous avons un petit effet. En fait, nous avons ici un effet négligeable. Ainsi, même si la différence est significative statistiquement, elle ne veut pas dire grand-chose d'un point de vue pratique.

Pour pouvoir réaliser un test de Student, il faut respecter certaines conditions d'application. Tout d'abord, la distribution des valeurs dans chaque groupe doit suivre une loi normale. Il existe plusieurs méthodes pour vérifier cette condition de normalité. Nous allons simplement représenter graphiquement, sous forme d'histogramme, la distribution des valeurs dans les deux groupes, grâce à la fonction hist :

```
hist(dat_nom$rt[dat_nom$genre == "f"])
hist(dat_nom$rt[dat_nom$genre == "m"])
```

Les figures 7.1 et 7.2 présentent respectivement les temps de décision lexicale pour les noms féminins et masculins.

Figure 7.1
Histogramme des temps de décision lexicale pour les noms féminins

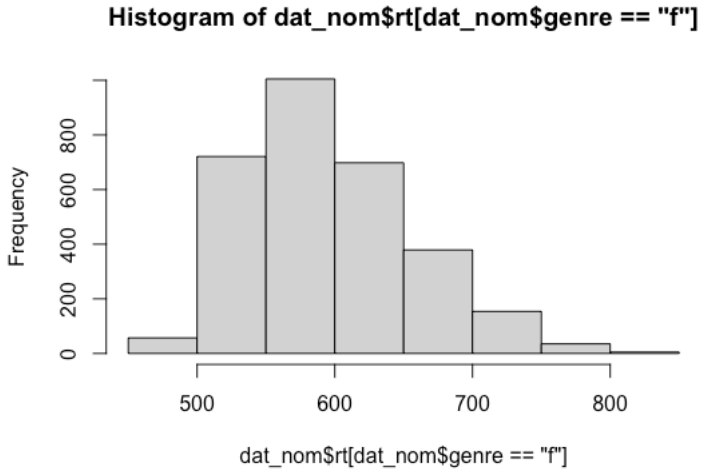
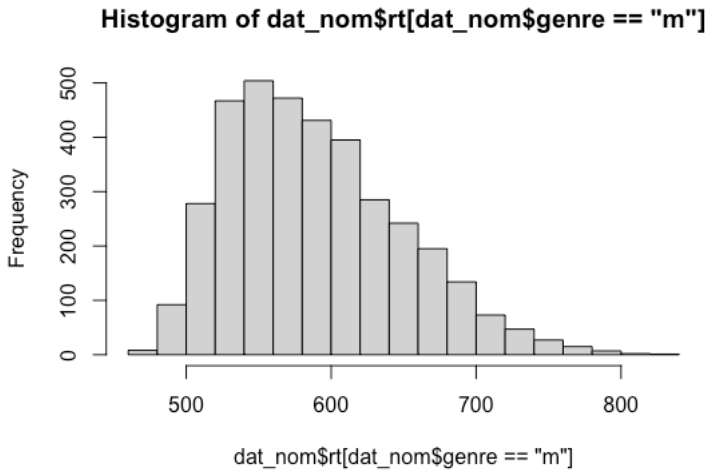


Figure 7.2
Histogramme des temps de décision lexicale pour les noms masculins



Sur les deux histogrammes, on peut voir une asymétrie positive des distributions : les valeurs s'étalent plus sur la moitié droite de la distribution (les temps les plus longs) que sur la moitié gauche (les temps les plus courts). La forme de ces deux distributions indique que la condition de normalité n'est pas respectée. Une solution possible est de faire un test non-paramétrique à la place du test de Student. En effet, le test de Student est un test paramétrique basé sur la distribution normale, ce qui n'est pas le cas pour un test non-paramétrique. Dans le cas d'une comparaison entre deux groupes indépendants, nous pouvons utiliser le test de Mann-Whitney (aussi appelé « test de la somme des rangs de Wilcoxon » ou « test de Wilcoxon »). La ligne de code suivante permet de faire cela :

```
wilcox.test(dat_nom$rt ~ dat_nom$genre)
```

La *p-value* est inférieure à .05, la conclusion est donc la même qu'avec le test de Student. L'autre condition d'application à respecter quand on fait un test de Student est l'homogénéité des variances. Le principe est le suivant : la variance dans chacun des deux groupes doit être homogène. Pour vérifier cette condition d'application, le test le plus souvent utilisé est le test de Levene. Il existe une fonction dans le package *car* permettant de faire ce test. Commençons par installer et charger ce package :

```
install.packages("car")
library(car)
```

Nous pouvons maintenant utiliser la fonction leveneTest et lancer le test :

```
leveneTest(dat_nom$rt ~ dat_nom$genre)
```

L'hypothèse nulle testée est la suivante : les variances dans les différents groupes sont homogènes dans la population parente. La *p-value* du test de Levene est égal à 0.45, on ne peut donc pas rejeter l'hypothèse nulle : cela signifie que les variances sont homogènes et que la condition d'homogénéité des variances est respectée. Si la condition n'avait pas été respectée (et que par conséquent le test de Levene avait été significatif), une solution aurait été de réaliser le test de Student mais avec la correction de Welch. En réalité, quand on réalise un test de Student dans R, la correction de Welch est appliquée par défaut (c'est d'ailleurs indiqué dans les résultats) suite à des travaux recommandant de faire ainsi (Delacre et al.,

2017). Au final, nous n'avons pas trop à nous soucier de cette condition d'homogénéité des variances puisque la correction de Welch est appliquée par défaut dans R.

Un autre point à connaître quand on fait un test de Student concerne l'hypothèse alternative du chercheur. Dans notre exemple, plusieurs hypothèses alternatives sont possibles : nous pouvons faire l'hypothèse que les deux moyennes (temps de décision lexicale pour les mots féminins et pour les mots masculins) sont différentes (sans préciser le sens de cette différence). Mais nous pouvons aussi faire l'hypothèse que le temps moyen est plus long pour les mots féminins que pour les mots masculins. Enfin, nous pouvons faire l'hypothèse que le temps moyen est plus long pour les mots masculins que pour les mots féminins. Si on fait l'hypothèse d'une différence sans préciser le sens, on parle d'une hypothèse bilatérale. Si on précise le sens, on parle d'une hypothèse unilatérale (cf. Chapitre 6 dans la partie 2). Le choix de l'hypothèse alternative peut-être précisé quand on fait un test de Student dans R, et ce choix a une influence sur la valeur de la *p-value* : sur un même jeu de données, la *p-value* sera plus faible avec une hypothèse unilatérale par rapport à une hypothèse bilatérale. Pour préciser l'hypothèse alternative, il existe un argument dans la fonction t.test nommé « alternative ». La valeur par défaut attribuée à cet argument est « two.sided », ce qui signifie « bilatéral ». Par défaut, quand on fait un test de Student dans R, l'hypothèse alternative est bilatérale. Si nous exécutons la ligne de code suivante, nous obtiendrons exactement les mêmes résultats que pour le test de Student pour lequel nous n'avions pas précisé l'hypothèse alternative :

```
t.test(dat_nom$rt ~ dat_nom$genre, alternative = "two.sided")
```

Nous pouvons modifier la valeur de l'argument « alternative ». Si notre hypothèse alternative stipule que les mots féminins ont en moyenne un temps de décision lexicale plus long que les mots masculins, nous pouvons écrire la ligne de code suivante :

```
t.test(dat_nom$rt ~ dat_nom$genre, alternative = "greater")
```

Le test est toujours significatif, mais la *p-value* est plus faible que précédemment. Si notre hypothèse alternative stipulait que le temps moyen était plus court pour les mots féminins que pour les mots masculins, nous aurions pu utiliser la ligne de code suivante :

```
t.test(dat_nom$rt ~ dat_nom$genre, alternative = "less")
```

Les moyennes étant à l'inverse de cette hypothèse, la *p-value* est égal à 1, indiquant que nous ne pouvons pas rejeter l'hypothèse nulle. Pour savoir s'il faut indiquer « greater » ou « less », il faut savoir dans quel ordre R a classé les deux modalités de la variable « genre ». Le principe est très simple : R classe les modalités par ordre alphabétique. Les deux modalités étant « f » pour les mots féminins et « m » pour les mots masculins, la première modalité est « f » et la seconde est « m ».

Comme nous venons de le voir, le test de Student permet de tester la différence entre deux moyennes au maximum. Si nous nous intéressons aux différences entre trois moyennes ou plus, nous pouvons utiliser un autre test inférentiel : l'ANOVA.

3. ANOVA

L'ANOVA (qui signifie « ANalysis Of VAriance ») a été formalisée par Ronald Fischer au début du 20ème siècle. Elle permet d'étudier l'effet d'une variable catégorielle à trois

modalités ou plus sur une variable numérique. C'est en quelque sorte une extension du test de Student permettant de comparer trois moyennes ou plus. Dans le jeu de données, nous allons nous intéresser à la variable « cgram ». Nous avons vu précédemment que cette variable a 21 modalités différentes. Nous allons filtrer cette variable et garder uniquement 3 modalités : les noms, les adjectifs et les verbes. Pour cela, nous allons créer un nouvel objet « dat_cgram » :

```
dat_cgram <- dat[dat$cgram == "NOM" | dat$cgram == "ADJ" | dat$cgram == "VER", ]
```

Nous pouvons vérifier que dans ce nouveau tableau, la variable « cgram » a bien 3 modalités seulement :

```
unique(dat_cgram$cgram)
```

L'objectif est de voir si cette variable à 3 modalités a un effet sur les temps de décision lexicale. Dit autrement : est-ce que le temps varie selon que ce mot est un nom, un adjectif ou un verbe ? Pour répondre à cette question, nous créons un objet qui contiendra les résultats de l'ANOVA :

```
model_anova <- aov(rt ~ cgram, data = dat_cgram)
```

Le fait de créer un objet (plutôt que d'afficher directement les résultats de l'ANOVA) va permettre d'utiliser cet objet dans les autres lignes de code de notre ANOVA et éviter ainsi d'écrire à chaque fois le nom des variables et le nom du jeu de données. Pour afficher les résultats de l'ANOVA, il suffit d'appliquer la fonction summary sur notre objet « model_anova » :

```
summary(model_anova)
```

Voici les résultats de l'ANOVA :

```
             Df    Sum Sq Mean Sq F value Pr(>F)
cgram         2    562160  281080    89.7 <2e-16 ***
Residuals   33194 104012548    3133
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le test F rapporté est un test inférentiel ce qui signifie qu'une hypothèse nulle a été testée. Dans notre exemple, l'hypothèse nulle testée est la suivante : les 3 moyennes (temps moyen pour les noms, les adjectifs et les verbes) sont égales dans la population parente. Les résultats du test F sont les suivants : $F(562160, 104012548) = 89.7$, $p < .001$. Au seuil .05, nous pouvons rejeter l'hypothèse nulle : la catégorie grammaticale a un effet significatif sur le temps de décision lexicale. Dit autrement, il y a au moins une différence significative entre 2 moyennes parmi les 3. Mais le test F ne nous dit pas quelle moyenne diffère significativement de quelle autre. C'est pourquoi en général, après avoir calculé le test F, l'analyse se poursuit pour savoir où se situent les différences significatives entre les moyennes. Avant de faire cela, commençons par afficher les valeurs des 3 moyennes. Nous pouvons utiliser la fonction tapply :

```
tapply(dat_cgram$rt, dat_cgram$cgram, mean)
```

La fonction tapply prend trois arguments : la variable sur laquelle nous souhaitons calculer les moyennes (variable « rt »), la variable spécifiant les différents groupes (variable « cgram ») et l'indicateur que l'on souhaite calculer (ici la moyennne, en anglais « mean »). Les résultats montrent que le temps moyen est très proche dans les trois catégories grammaticales : 590 ms pour les adjectifs, 589 ms pour les noms et 597 ms pour les verbes. Mais comme dit précédemment, nous ne savons pas si les différences entre ces trois moyennes sont dues au hasard (plus précisément si c'est dû à des erreurs d'échantillonnage, cf. Chapitre 6 dans la partie 2) ou si c'est vraiment dû à la variable « cgram ». Le test F nous indique qu'il y a au moins une différence significative. Peut-être que toutes les différences sont significatives mais pour l'instant nous ne le savons pas. Pour répondre à cette question, nous pouvons réaliser des tests post-hoc en utilisant la fonction TukeyHSD :

```
TukeyHSD(model_anova)
```

Voici les résultats des tests post-hoc :

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = rt ~ cgram, data = dat_cgram)

$cgram
            diff       lwr       upr     p adj
NOM-ADJ -0.850377 -2.869604  1.168850 0.5850704
VER-ADJ  7.706468  5.778160  9.634776 0.0000000
VER-NOM  8.556845  6.930361 10.183330 0.0000000
```

Les résultats affichent trois lignes. Chaque ligne teste la différence entre 2 moyennes parmi les 3. En regardant les *p-value*s (dans la dernière colonne à droite), nous pouvons voir que la différence entre les noms et les adjectifs n'est pas significative (p = .59). En revanche, la différence entre les verbes et les adjectifs est significative, de même que la différence entre les verbes et les noms. Encore une fois, étant donné que nous avons un effectif très important de mots dans notre échantillon, ces différences statistiquement significatives n'ont pas forcément une significativité pratique. Pour vérifier cela, nous pouvons calculer une taille d'effet. Un indicateur de taille d'effet classiquement utilisé dans l'ANOVA est eta carré. Une fonction dans le package *effectsize* permet de calculer cette taille d'effet avec la fonction eta_squared :

```
eta_squared(model_anova)
```

Les résultats indiquent que eta carré est égal à 0.00538. Nous pouvons mettre cette valeur en pourcentage et l'interpréter de la manière suivante : la catégorie grammaticale d'un mot explique 0.5% de la variation des temps de décision lexicale. D'un point de vue pratique, l'effet est très faible. Cohen (1988) a proposé des seuils permettant d'interpréter une valeur d'eta carré : une valeur supérieure à 0.01 indique un petit effet, une valeur supérieure à 0.06 indique un effet moyen, et une valeur supérieure à 0.14 indique un effet fort. Dans notre exemple, on peut dire encore une fois que l'effet est négligeable. Comme pour l'exemple du test de Student, l'effet est statistiquement significatif mais pas d'un point de vue pratique.

Pour faire une ANOVA, il est nécessaire de respecter certaines conditions d'application. Les conditions sont les mêmes que pour le test de Student : normalité et homogénéité des variances. Pour vérifier la condition de normalité, nous pouvons faire un histogramme pour chaque groupe :

```
hist(dat_cgram$rt[dat_cgram$cgram == "NOM"])
hist(dat_cgram$rt[dat_cgram$cgram == "ADJ"])
hist(dat_cgram$rt[dat_cgram$cgram == "VER"])
```

Les trois histogrammes montrent une déviation par rapport à la normalité, due encore une fois à une asymétrie positive des temps de décision lexicale.
Pour vérifier la condition d'homogénéité des variances, nous pouvons utiliser le test de Levene :

```
leveneTest(dat_cgram$rt ~ dat_cgram$cgram)
```

Les résultats indiquent que le test de Levene est significatif, ce qui veut dire que la condition d'homogénéité des variances n'est pas respectée.
Si les conditions d'application ne sont pas respectées (comme c'est le cas dans notre exemple), il est possible de faire un test non-paramétrique nommé « test de Kruskal-Wallis » en utilisant la fonction kruskal.test :

```
kruskal.test(rt ~ cgram, data = dat_cgram)
```

Les résultats indiquent que le test est significatif et donc la conclusion est la même que pour le test F classique de l'ANOVA : la catégorie grammaticale a un effet significatif sur les temps de décision lexicale. Pour savoir quelles sont les différences significatives parmi les trois paires de moyennes, nous pouvons faire des tests post-hoc. Le test le plus utilisé est le test de Dunn. Une fonction permettant de faire ce test se trouve dans le package *FSA*, nous allons donc installer puis charger ce package :

```
install.packages("FSA")
library(FSA)
```

Nous pouvons maintenant utiliser la fonction dunnTest de la manière suivante :

```
dunnTest(rt ~ cgram, data = dat_cgram)
```

Pour savoir si une différence entre deux moyennes est significative, il suffit de regarder la dernière colonne nommée « P.adj » : les résultats indiquent que la différence entre les adjectifs et les noms n'est pas significative (p = .14) mais les deux autres différences sont significatives.

Seule l'ANOVA pour échantillons indépendants a été présentée dans ce chapitre. Pour comprendre le principe de l'ANOVA pour échantillons appariés et son application dans R, vous pouvez vous référer au livre de Field et al. (2012).

4. Test du χ^2

Le test du χ^2 (prononcer « khi deux » ou « khi carré ») a été proposé par Karl Pearson en 1900. Il existe plusieurs sortes de χ^2, nous aborderons uniquement le test du χ^2

d'indépendance (pour plus d'informations sur les différents tests de χ^2, voir Howell et al., 2008).
Le test du χ^2 d'indépendance est un test non-paramétrique permettant de voir s'il existe un lien entre deux variables catégorielles. Avec notre jeu de données, nous pouvons par exemple répondre à la question suivante : y-a-t-il un lien entre la catégorie grammaticale et le genre des mots ? Plus précisément, les proportions de mots masculins et féminins sont-elles les mêmes pour les noms et pour les adjectifs ? Pour répondre à cette question, nous pouvons faire un χ^2 d'indépendance. On parle d'indépendance car nous voulons voir si les deux variables (catégorie grammaticale et genre) sont dépendantes ou indépendantes. Si elles sont dépendantes, cela veut dire que la proportion de mots féminins est différente pour les noms et les adjectifs.
Nous allons commencer par créer un objet « dat_chi2 » qui contiendra uniquement les mots correspondant aux catégories grammaticales « nom » et « adjectif » et uniquement les mots masculins et féminins. Les deux lignes de code suivantes permettront de faire cela :

```
dat_chi2 <- dat[dat$cgram == "NOM" | dat$cgram == "ADJ", ]
dat_chi2 <- dat_chi2[dat_chi2$genre != "", ]
```

On peut vérifier en affichant les différentes modalités de chaque variable :

```
unique(dat_chi2$cgram)
unique(dat_chi2$genre)
```

Nous pouvons maintenant demander à R de nous afficher les effectifs résultant du croisement des deux variables « cgram » et « genre », grâce à la fonction table :

```
table(dat_chi2$cgram, dat_chi2$genre)
```

On peut voir qu'il y a 2121 adjectifs féminins, 3477 adjectifs masculins, 4726 noms féminins et 6176 noms masculins. Étant donné qu'il y a plus de noms que d'adjectifs, il est plus intéressant de demander à afficher les proportions, grâce à la fonction prop.table :

```
prop.table(table(dat_chi2$cgram, dat_chi2$genre), margin = 1)*100
```

Quelques mots sur cette ligne de code : la fonction prop.table prend comme argument principal le tableau des effectifs (table(dat_ chi2$cgram, dat_chi2$genre)). Un autre argument, « margin », est précisé : si la valeur 1 est attribuée à cet argument, alors les proportions seront calculées par ligne (c'est-à-dire que la somme des proportions sera égale à 1 pour chaque ligne). Si la valeur 2 est attribuée, alors les proportions seront calculées par colonne. Enfin, la fonction prop.table est multipliée par 100 afin d'avoir les résultats en pourcentage, ce qui est généralement plus facile à lire. Les résultats montrent ainsi qu'il y a 37.9% d'adjectifs féminins et 43.3% de noms féminins. Dans notre échantillon, la proportion de mots féminins est plus importante pour les noms que pour les adjectifs. Mais cette différence est-elle due au hasard ou bien au fait que les mots appartiennent à deux catégories grammaticales différentes ? Pour répondre à cette question, il faut faire un test inférentiel : le test du χ^2 d'indépendance. La ligne de code suivante permet de lancer l'analyse :

```
chisq.test(table(dat_chi2$cgram, dat_chi2$genre))
```

L'hypothèse nulle testée est que les deux variables (catégorie grammaticale et genre) sont indépendantes. Dit autrement, l'hypothèse nulle stipule que les deux proportions (pourcentage

d'adjectifs féminins et pourcentage de noms féminins) sont égales dans la population parente et que la différence dans l'échantillon est due au hasard. Les résultats sont les suivants : chi2(1) = 45.22, p < .001. Au seuil .05, on peut rejeter l'hypothèse nulle et conclure que les deux variables sont dépendantes : la proportion de noms féminins est significativement plus importante que la proportion d'adjectifs féminins.

La taille d'effet la plus souvent rapportée quand on fait un test de χ^2 d'indépendance est le V de Cramer. Pour calculer cette taille d'effet, nous pouvons utiliser la fonction cramers_v du package « effectsize » :

```
cramers_v(table(dat_chi2$cgram, dat_chi2$genre))
```

Les résultats indiquent que la valeur du V de Cramer est égale à 0.05. Les seuils proposés par Cohen (1988) sont les suivants : si la valeur est supérieure à 0.10, on peut parler de petit effet. Si la valeur est supérieure à 0.30, on peut parler d'effet moyen. Au-dessus de 0.50, on parle d'effet fort. Encore une fois, l'effet dans notre exemple est négligeable.

Comme pour les autres tests, certaines conditions d'application doivent être respectées pour pouvoir faire un test de χ^2 d'indépendance. Tout d'abord, les deux variables doivent être catégorielles (avec 2 modalités ou plus). De plus, les variables ne doivent pas avoir de mesures répétées. Par exemple, si une des deux variables est du type « pré-test / post-test », avec les mêmes participants dans les deux conditions, vous ne pourrez pas faire un test de χ^2 d'indépendance. Enfin, les effectifs théoriques doivent être supérieurs à 5 dans la majorité des conditions (80% selon Yates et al., 1999). Pour calculer l'effectif théorique d'une condition, il suffit de multiplier l'effectif total de la ligne contenant cette condition par l'effectif total de la colonne contenant cette condition, puis de diviser par l'effectif total du tableau. Pour être plus clair, prenons le tableau des effectifs que nous avions affiché avec la ligne de code suivante :

```
table(dat_chi2$cgram, dat_chi2$genre)
```

L'effectif réel pour la condition « adjectif féminin » est 2121. Pour calculer l'effectif théorique de cette condition, il suffit de calculer la somme de la ligne « adjectif » (2121 + 3477 = 5598), de calculer la somme de la colonne « féminin » (2121 + 4726 = 6847). On fait ensuite le produit : 5598 * 6847 = 38 329 506. Et on divise ce produit par l'effectif total du tableau 38 329 506 / (2121 + 3477 + 4726 + 6176) = 2323. L'effectif théorique pour la condition « adjectif féminin » est de 2323, ce qui largement au-dessus de 5. En suivant la même procédure pour les trois autres conditions, nous pourrions voir que cette condition d'application est respectée puisque tous les effectifs théoriques sont supérieurs à 5.

5. Test de Corrélation

Le coefficient de corrélation a été développé par Karl Pearson à partir d'une idée proposée par Galton dans les années 1880, lui-même inspiré par les travaux d'Auguste Bravais en 1844. Le coefficient de corrélation (que l'on nomme également le r de Bravais-Pearson) permet de voir s'il existe un lien entre deux variables numériques. La valeur de cet indicateur est toujours comprise entre -1 et +1. Plus la valeur est proche de -1 ou de 1 et plus la relation entre les deux variables est forte. Le signe du r indique le sens de la relation. Par exemple, la distance entre deux points est corrélée positivement au temps de trajet en voiture : lorsque la distance augmente, le temps de trajet augmente. A l'inverse, la vitesse de déplacement de la

voiture est corrélée négativement au temps de trajet : lorsque la vitesse de déplacement augmente, le temps de trajet diminue.
Dans notre jeu de données, nous pouvons par exemple voir s'il existe une relation entre le nombre de lettres dans un mot et le nombre de phonèmes. Intuitivement, cela paraît très probable et nous pouvons même faire l'hypothèse que cette relation est positive : plus le nombre de phonèmes augmente et plus le nombre de lettres augmente.
Pour calculer ce coefficient de corrélation, nous allons utiliser la fonction cor :

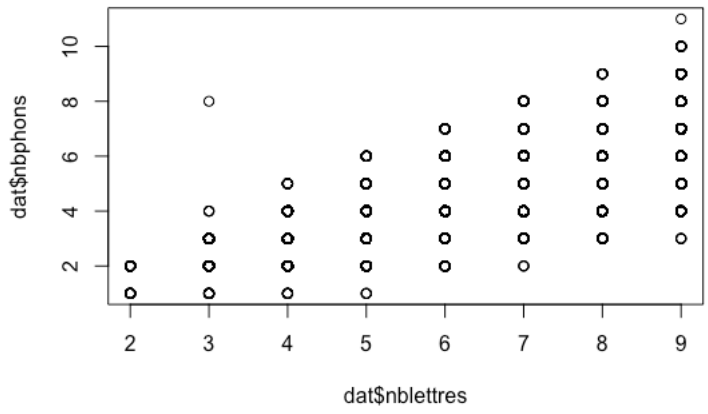
```
cor(dat$nblettres, dat$nbphons)
```

R nous indique que le r est égal à 0.73. Comme on pouvait l'imaginer, la relation est relativement forte et positive. Il est conseillé de toujours représenter la relation sous forme de graphique, cela permet par exemple de repérer certains points atypiques (qui peuvent parfois fortement influencer la valeur du coefficient de corrélation). Pour cela, nous allons utiliser la fonction « plot » :

```
plot(dat$nblettres, dat$nbphons)
```

Le nuage de points est représenté sur la figure 7.7.

Figure 7.7
Nuage de points illustrant la relation entre le nombre de lettres et le nombre de phonèmes



Le nuage de points montre bien une relation positive entre les deux variables. On peut également remarquer un point atypique : c'est un mot de 3 lettres mais qui possède 8 phonèmes, ce qui semble étrange au premier abord. Pour savoir quel est ce mot, nous pouvons utiliser la ligne de code suivante :

```
dat$mot[dat$nblettres == 3 & dat$nbphons == 8]
```

R nous indique qu'il s'agit du mot « etc ». Si nous souhaitions exclure ce mot des analyses, le fait qu'il soit une forme abrégée de la locution « et cetera » serait une bonne justification. Mais pour ne pas perdre de temps, nous allons garder ce mot dans le jeu de données.
Comme nous l'avons vu, il y a une corrélation positive entre les deux variables dans notre échantillon. Mais cette corrélation existe-t-elle dans la population parente ? Pour répondre à cette question, nous allons faire un test inférentiel. Il s'agit en fait d'un test de Student.

Comme nous l'avons vu précédemment, le test de Student permet de tester si deux moyennes sont significativement différentes. Mais il permet aussi de tester la significativité d'un coefficient de corrélation. L'hypothèse nulle testée dans ce cas est la suivante : le coefficient de corrélation est égal à 0 dans la population parente. Pour tester la significativité du coefficient de corrélation, il faut utiliser la fonction cor.test :

```
cor.test(dat$nblettres, dat$nbphons)
```

Les résultats du test de corrélation sont les suivants :

```
        Pearson's product-moment correlation

data:  dat$nblettres and dat$nbphons
t = 195.57, df = 33809, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7235267 0.7335297
sample estimates:
    cor
0.728567
```

Les résultats sont les suivants : $t(33809) = 195.57$, $p < .001$. Au seuil .05, nous pouvons rejeter H0 et dire que la corrélation est significative. Le coefficient de corrélation obtenu dans notre échantillon (0.73, la valeur indiquée sur la dernière ligne de la sortie R) n'est pas dû au hasard mais bien au fait qu'il y a un lien entre les deux variables.
Il est possible de spécifier l'hypothèse alternative quand on teste la significativité du coefficient de corrélation. L'argument à spécifier est « alternative ». Par défaut, cet argument a la valeur « two.sided », ce qui signifie que l'hypothèse alternative est bilatérale. La ligne de code suivante donne les mêmes résultats que précédemment :

```
cor.test(dat$nblettres, dat$nbphons, alternative = "two.sided")
```

Mais dans notre exemple, nous pouvons intuitivement faire l'hypothèse que la corrélation entre le nombre de lettres et le nombre de phonèmes est positive. Nous pouvons donc tester la significativité du coefficient de corrélation avec une hypothèse unilatérale. La valeur « greater » signifie que nous faisons l'hypothèse que la corrélation est positive et la valeur « less » signifie que la corrélation est négative. Nous allons donc attribuer la valeur « greater » :

```
cor.test(dat$nblettres, dat$nbphons, alternative = "greater")
```

Puisque le test était significatif avec une hypothèse bilatérale, il est forcément aussi significatif avec une hypothèse unilatérale.
Comme nous l'avons vu pour les autres tests, le fait que le coefficient de corrélation soit statistiquement significatif ne nous dit rien quant à la significativité pratique. Il faut donc aussi rapporter une taille d'effet. Dans le cas du coefficient de corrélation, ce n'est vraiment pas compliqué car le coefficient de corrélation est une taille d'effet ! Plus le coefficient est proche de -1 ou de 1 et plus l'effet est fort. Souvent, c'est le contexte dans lequel l'étude est menée qui permet d'interpréter si le coefficient de corrélation représente un effet plus ou

moins fort. Mais Cohen (1988) a proposé des seuils : un coefficient de corrélation supérieur à 0.10 (en valeur absolue) indique un petit effet, une valeur supérieure à 0.30 indique un effet moyen et une valeur supérieure à 0.50 indique un effet fort.
Le test de significativité du coefficient de corrélation implique de respecter certaines conditions d'application. Tout d'abord, la relation entre les deux variables doit être linéaire. Cette condition peut se vérifier simplement sur un graphique représentant le nuage de points (avec la fonction « plot » comme nous l'avons vu) : si le nuage de points a une forme linéaire (ressemblant à une ligne droite), alors la condition est respectée. Si le nuage de points présente une ou plusieurs courbures, alors la condition de linéarité n'est pas respectée. Affichons de nouveau le nuage de points avec la ligne de code suivante :

```
plot(dat$nblettres, dat$nbphons)
```

Le nuage de points ne présente pas de courbures, on peut dire que la relation est linéaire.
Il y a également une autre condition d'application : les deux variables doivent être distribuées normalement. Pour vérifier cette condition de normalité, nous pouvons afficher un histogramme pour chaque variable :

```
hist(dat$nblettres)
hist(dat$nbphons)
```

La figure 7.9 présente l'histogramme de la variable « nblettres » et la figure 7.10 l'histogramme de la variable « nbphons ».

Figure 7.9
Histogramme de la variable « nblettres »

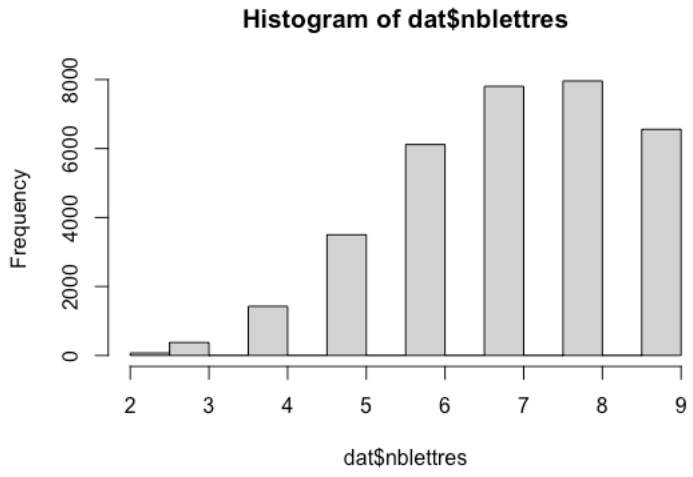
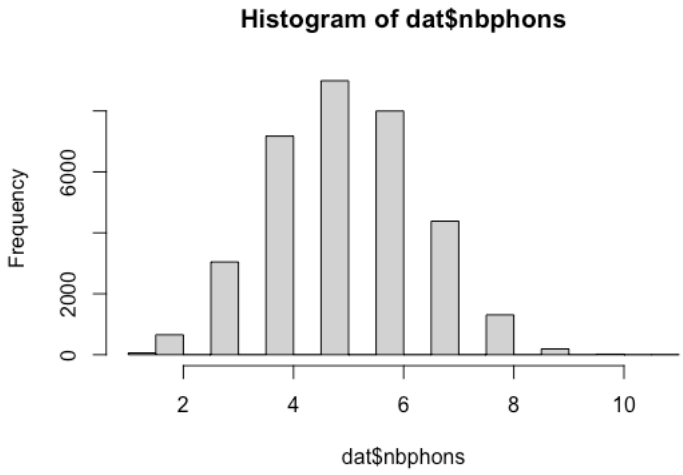


Figure 7.10
Histogramme de la variable « nbphons »



Le graphique montre une distribution plutôt normale pour la variable « nbphons ». En revanche, le graphique pour la variable « nblettres » montre une asymétrie négative. La condition de normalité n'est pas respectée. Une solution possible est d'utiliser un test non-paramétrique. Un test souvent utilisé est le rho de Spearman. Pour lancer ce test, il suffit d'utiliser la fonction cor.test et de spécifier l'argument (method = « spearman ») :

```
cor.test(dat$nblettres, dat$nbphons, method = "spearman")
```

Les résultats indiquent que la corrélation est significative.

Conclusion

Ce chapitre avait pour objectif de présenter quelques tests inférentiels classiquement utilisés lorsque deux variables sont en jeu. Cette présentation est loin d'être exhaustive. Par exemple, pour le test de Student, seule la situation dans laquelle les groupes sont indépendants a été présentée. Mais il existe un test de Student pour groupes appariés (on parle aussi de mesures répétées) : ce test est utile lorsque ce sont les mêmes participants dans les deux conditions (par exemple dans un protocole type « pré-test / post-test »). Dans ce cas, le format du fichier de données est différent, le test et les résultats sont également différents. De même, il existe aussi une ANOVA pour groupes appariés, même si nous ne l'avons pas abordé dans ce chapitre. Pour chaque test, plusieurs méthodes existent pour vérifier chaque condition d'application, chacune avec leurs avantages et leurs inconvénients. Il en est de même pour les tests non-paramétriques. Pour résumer, ce chapitre n'est qu'une brève introduction et le lecteur intéressé pourra trouver de nombreuses ressources dans les livres de statistiques ou sur internet pour en savoir plus (voir section « Pour aller plus loin » dans ce chapitre).

Points à retenir

- Les tests inférentiels bivariés permettent de vérifier s'il existe un lien significatif entre deux variables. Le test bivarié à utiliser dépend des échelles de mesure des variables.
- Le test de Student permet de voir si une variable catégorielle à deux modalités a un effet sur une variable numérique.
- L'ANOVA permet de voir si une variable catégorielle à trois modalités (ou plus) a un effet sur une variable numérique.

- Avec le test du χ^2, il est possible de tester l'existence d'un lien entre deux variables catégorielles.
- Le test de corrélation vérifie le lien entre deux variables numériques.
- Ces tests impliquent de respecter certaines conditions d'application. Si ces contraintes ne sont pas respectées, des alternatives non-paramétriques sont possibles.

Pour aller plus loin

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Howell, D., Bestgen, Y., Yzerbyt, V., & Rogier, M. (2008). *Méthodes statistiques en sciences humaines*. De Boeck.

Navarro, D. (n.d.). Learning statistics with R. https://learningstatisticswithr.com

Bibliographie

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Erlbaum.

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology*, *30*(1).

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., Dufau, S., Mathôt, S., & Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, *50*(3), 1285-1307.

Howell, D., Bestgen, Y., Yzerbyt, V., & Rogier, M. (2008). *Méthodes statistiques en sciences humaines*. De Boeck.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516-524.

Yates, D., Moore, D., & McCabe, G. (1999). *The Practice of Statistics*. Freeman.