Usage de méthodes et techniques statistiques dans la formation doctorale

Jean-Claude Régnier

CONFÉRENCE UNIVERSITAIRE DE SUISSE OCCIDENTALE

Journée D2
11 mars 2022

Études doctorales
Sciences de l'éducation

Démarches de recherche quantitatives en sciences de l'éducation : raisonnement statistique, initiation, apports

Usage de méthodes et techniques statistiques dans la formation doctorale.

Réflexion sur les apports méthodologiques et épistémologiques du raisonnement statistique dans les recherches du domaine de l'éducation et sur les précautions à prendre

Jean-Claude Régnier

icar
UMR 5191
Interactions, Corpus, Apprentissages, Représentations

UNIVERSITÉ DE LYON

LABEX ASLAN

CONFÉRENCE UNIVERSITAIRE
DE SUISSE OCCIDENTALE

Journée D2
11 mars 2022

Études doctorales
Sciences de l'éducation

Démarches de recherche quantitatives en sciences de l'éducation : raisonnement statistique, initiation, apports

---

Usage de méthodes et techniques statistiques dans la formation doctorale.

Réflexion sur les apports méthodologiques et épistémologiques du raisonnement statistique dans les recherches du domaine de l'éducation et sur les précautions à prendre

Jean-Claude Régnier
Professeur des universités émérite
Jean-Claude.Regnier@univ-lyon2.fr
Membre du laboratoire UMR 5191 ICAR - Université Lumière Lyon 2
Professeur invité National Research Tomsk State University – Tomsk – Sibérie
Pages persos : http://www.icar.cnrs.fr/membre/jcregnier/

Sommaire

Index des figures

Index des tableaux

1 Introduction

L'objectif de cette communication est de partager, dans un esprit de coopération, notre réflexion centrée sur le raisonnement statistique et la formation de l'esprit statistique. Nous considérons que *l'esprit statistique naît lorsqu'on prend conscience de l'existence de fluctuation d'échantillonnage* et que sa formation qui passe par une formation en statistique minimale, requiert un certain renoncement à l'usage systématique de l'idée de vérité pour chercher à maîtriser celle de vraisemblance et de plausibilité (Régnier, 1998f). Cette formation tire bénéfice d'être organisée, en particulier, autour d'activités fondamentales que sont la modélisation statistique, l'analyse statistique et l'interprétation statistique sans oublier le recours à la simulation. Dans des articles déjà anciens de (Régnier, 2002, 2005a, 2005b) nous avons tenté d'aborder cette question et d'apporter quelques éléments d'éclaircissement.

En tenant d'adopter une posture de praticien-chercheur [en formation] réflexif, nous pouvons nous interroger sur la place de la statistique dans la formation en sciences humaines et sociales.

En relation à notre pratique, nous nous intéressons davantage à la place de la statistique dans la formation en sciences de l'éducation. Nous avons entrevu pour la discipline *statistique*, au moins cinq positions :

- Discipline de base,
- Discipline de service, discipline-outil,
- Discipline d'ouverture,
- Discipline-objet de la didactique de la statistique,
- Discipline-objet de la recherche en statistique dans son application à la recherche en sciences de l'éducation,

Nous pouvons tenter d'identifier ce qu'induit la prise en considération de chacune de ces positions dans une formation doctorale non spécialisée en statistique. Toutefois il ressort que la place la plus habituelle à considérer ici est celle de *Discipline de service, discipline-outil*.

Notre communication tentera de rester guidée par la prise de conscience et l'identification des apports méthodologiques et épistémologiques de la mise en œuvre de méthodes statistiques basées sur le raisonnement statistique dans les recherches du domaine de l'éducation

2 Un détour par les questions récurrentes sur les dimensions qualitatives et quantitatives

Commençons par un petit détour sur l'opposition récurrente entre **quantitatif** et **qualitatif,** objet d'un inépuisable débat dans le domaine de la recherche en sciences de l'éducation, alors qu'il s'agit de deux versants complémentaires et indissociables… Force est de constater dans le discours de nombre de chercheurs avancés et repris par conséquent par les chercheurs en formation parmi lesquels nous considérons les doctorants et les doctorantes, une insistance à situer dans quelle perspective exclusive qualitative ou quantitative ils se placent. En nous basant sur plus de 35 ans d'expérience d'enseignement de la statistique au niveau universitaire et d'encadrement de travaux de recherche, nous avons pu constater quelques associations fortes au moins dans les domaines des sciences humaines et sociales. Réaliser une enquête par questionnaires relève immédiatement d'une perspective quantitative tandis qu'une enquête par entretiens évoque une perspective qualitative. Et à cela, sont associées des représentations liées au rapport avec les mathématiques. Lors de diverses enquêtes par questionnaires que nous avons conduites auprès d'étudiants en sciences de l'éducation, à la question portant sur l'évocation de la statistique, le mot *mathématique* figurait parmi les occurrences les plus fréquentes. À cela, il faut ajouter le fréquent rapport négatif que ces étudiants entretiennent avec le domaine des mathématiques, ce qui conduit à cet enchainement maintes fois constaté :

Approche quantitative = enquête par questionnaire = traitement statistique = usage des mathématiques qui est un domaine dans lequel je me sens incompétent et qui plus est, que je n'aime pas !

C'est dans cette représentation que s'enracine en partie le choix d'une approche prétendument exclusivement qualitative.

2.1 Quantum ? Qualis ? Quot ? Quid ?

Si nous revenons à un des éléments du noyau moteur de la recherche scientifique, à savoir : le regard porté sur le monde pour le comprendre ou l'expliquer, il nous conduit nécessaire à le questionner. Il nous semble qu'à côté des interrogatifs fondamentaux : pourquoi ? pour quoi ? comment ? dans quelles mesures ? … nous devons recourir *a minima* à quatre interrogatifs, selon notre orientation épistémologique, qui vont porter sur des données construites valides, pertinentes et fiables sur lesquelles vont s'enraciner l'argumentation des raisonnements qui fondent les réponses aux questions :

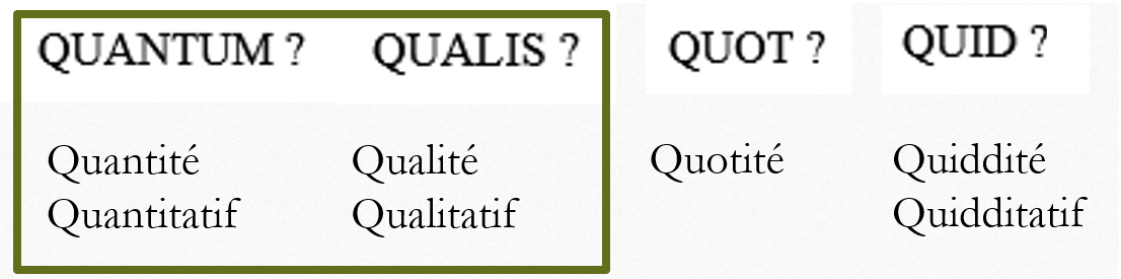


Figure 1 : Quatre questions fondamentales pour explorer et exploiter les données dans la recherche scientifique

Examinons les interrogatifs que la recherche dans le domaine de l'éducation ne semble pas avoir retenus : quot ? quid ?

La **quotité** concerne le nombre. Selon Antoine-Augustin Cournot (1801 – 1877) qui a contribué au XIX^{éme} au développement de la pensée statistique (Cournot, 1843, 1984), il s'agit « *PHILOS.* Nombre cardinal qui correspond à un ensemble d'objets constituant chacun une unité naturelle`` (Foulq. -St-Jean 1962): ... on blesse à la fois le sens philosophique et les analogies de la langue, lorsqu'on applique aux nombres purs, aux nombres qui désignent des collections d'objets vraiment individuels, la dénomination de quantités, en les qualifiant de *quantités discrètes* ou *discontinues*. Le marchand qui livre cent pieds d'arbres, vingt chevaux, ne livre pas des quantités, mais des nombres ou des **quotités**. Cournot, *Fond. connaiss.*, 1851, p. 288. » (Cournot, 1851) Dans notre langage actuel, cela semble ne concerner que les nombres entiers.

La **quiddité** concerne alors au sens philosophique « *PHILOS.* Essence d'une chose, ce qui fait qu'une chose est ce qu'elle est ».

Penchons-nous maintenant sur les deux interrogatifs sur lesquels repose le raisonnement statistique : quantum ? qualis ?

La **quantité** renvoie aussi à l'idée de nombre. Cela renvoie à un « *PHILOS.* [P. oppos. à *qualité*] Ensemble des déterminations mesurables telles que le nombre, la grandeur, le volume; ,,une des catégories fondamentales de la pensée désignant la grandeur abstraction faite de toute qualité et considérée seulement comme mesurable`` (Morf. *Philos.* 1980) »

Tandis que la **qualité** concerne plutôt les « caractéristiques de nature, bonne ou mauvaise, d'une chose ou d'une personne. »

Associés aux substantifs *quantité* et *qualité*, nous trouvons les adjectifs *quantitatifs* et *qualitatifs* qui sont alors utilisés pour qualifier un ensemble de notions du vocabulaire méthodologique.

Dans la figure ci-dessous, nous tentons d'expliciter un réseau notionnel :

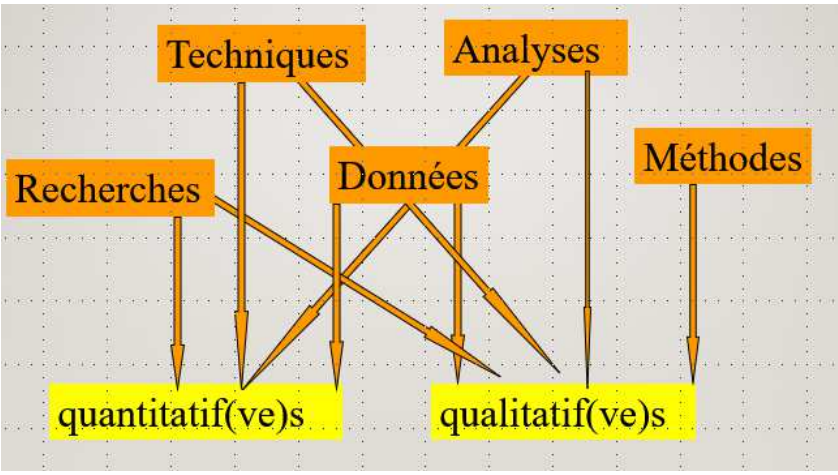


Figure 2 : Mise en œuvre des adjectifs *quantitatif* et *qualitatif*

Le discours mis en œuvre pour évoquer les approches méthodologiques dans les ouvrages spécialisés ou dans les articles scientifiques et les mémoires de recherche font un usage systématique de ces associations. Nous ne développerons pas cette question. Nous nous contentons ici de rapporter les propos de Huberman et Miles (1991 p.21-24) citant Miles (1979) sur le problème crucial de l'analyse qualitative.

> *« La difficulté la plus sérieuse et la plus centrale de l'utilisation de données qualitatives vient du fait que les méthodes d'analyse ne sont pas clairement formulées. Pour les données quantitatives, il existe des conventions précises que le chercheur peut utiliser. Mais l'analyste confronté à une banque de données qualitatives dispose de très peu de garde-fous pour éviter les interprétations hasardeuses, sans parler de la présentation de conclusions douteuses ou fausses à des publics de scientifiques ou de décideurs. Comment pouvons-nous être sûrs qu'une découverte "heureuse", "indéniable", "solide", n'est pas, en fait. erronée ? »*

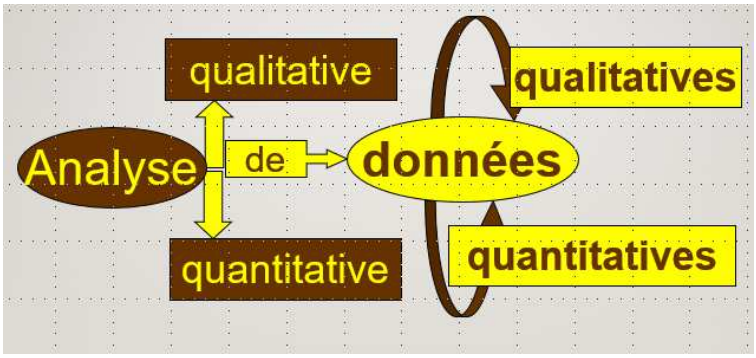


Figure 3 : Mise en œuvre des adjectifs *quantitatif* et *qualitatif*

À ce stade de notre réflexion, adjoindre les adjectifs épithètes *quantitatif* et *qualitatif* au substantif *analyse* et *données* nous semble toutefois acceptable dans les sens suivants :

- une analyse, c'est-à-dire une opération de décomposition, est qualitative s'il s'agit d'identifier la présence ou l'absence, et quantitative s'il s'agit de pondérer cette

présence ou cette absence. Par exemple sur les boites de médicament figure une information relative à la composition qualitative et quantitative en substances actives par unité de prise.

- - une donnée est quantitative si elle est représentée par un nombre (quantité) au sens mathématique, et qualitative, si elle représentée par autre chose qui ne corresponde pas une quantité, cela peut être un numéro dans la langue française. La nature de la donnée qui est issue d'une réalisation d'une variable statistique ou d'un vecteur-variable statistique, est, pour partie, basée sur un choix du chercheur. Si nous nous intéressons à l'âge des individus d'une population donnée, la variable « âge » peut être modélisée sous la forme d'une variable quantitative discrète (nombre d'années) ou continue (tranches d'âge), ou encore d'une variable qualitative ordinale (très jeune, jeune, ancien, très ancien) ou simplement nominale par des catégories non ordonnées.

2.2 Vers une approche mixte articulant la complémentarité du quantitatif et du qualitatif : quanti-qualitatif ou quali-quantitatif

Il existe un courant dans lequel nous nous retrouvons qui soutient la nécessité dans la recherche scientifique du domaine des sciences humaines et sociales, en particulier dans celui des sciences de l'éducation et de la formation, de la prise en compte de la complémentarité du quantitatif et du qualitatif. Voici un exemple.

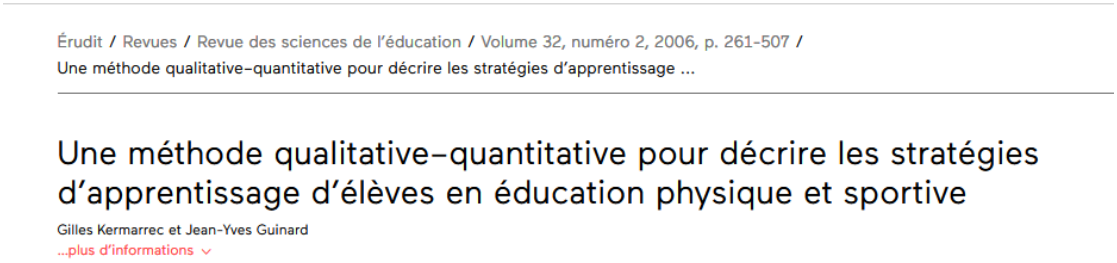


Figure 4 : Source : https://www.erudit.org/fr/revues/rse/2006-v32-n2-rse1456/014575ar/

Dans cette perspective, nous souscrivons à l'argumentation de Jacques Jenny (2014) qui soutient la thèse que nous reprenons ici.

> « […] *non seulement les deux grands «genres méthodologiques» dénommés habituellement «Qualitatif» et «Quantitatif» sont nécessairement complémentaires – banalité largement partagée – mais surtout que* ***leur distinction même est foncièrement artificielle, fallacieuse et par conséquent stérile et contre-productive****. Cette division classique du travail méthodologique procède d'une* ***illusion d'optique****, provoquée par une* ***perception tronquée*** *de ce que sont réellement, concrètement, d'une part les pratiques et méthodes dites «quantitatives» et d'autre part les pratiques et méthodes dites «qualitatives». Déjà,* ***la frontière*** *entre les deux n'est pas toujours placée aux mêmes endroits, selon les points de vue et selon les auteurs. Et que dire des* ***espèces «hybrides»*** *? où classer par exemple les «****statistiques textuelles****» et autres lexicométries, qui font entrer le langage dans le camp des mathématiques ? Ce qui*

> *mériterait **éventuellement** les qualificatifs distincts de Quantitatif et Qualitatif, ce ne sont pas des Méthodes mais ce qu'on pourrait appeler les **«plateformes techniques»** de ces méthodes. Car il est exact qu'on se trouve confronté, dans un cas, à des répartitions numériques et, dans l'autre, à des énoncés langagiers – et que ces deux **matériaux** ont des structures et des contraintes spécifiques telles que leur analyse, leur interprétation, exige la* discipline de spécialités pertinentes et performantes *: disons pour simplifier, respectivement les mathématiques, les statistiques et les sciences du langage, la sociolinguistique. Mais les méthodes, elles, ne peuvent pas se réduire à ces disciplines spécifiques :1) les traitements mathé-statistiques doivent impérativement prendre en compte, intégrer, les significations précises et circonstanciées des catégories de classement des objets dénombrés, ordonnés ou mesurés et, au-delà, les discours dans lesquels ces catégories prennent sens, sous peine de n'être que des* «exercices de calcul», *2) les analyses discursives de corpus textuels ne peuvent négliger ni les «opérateurs de quantification» que contient tout énoncé ni les formes de répartition spatio-temporelle de leurs éléments constitutifs, qui contribuent à leurs significations, sous peine de n'être que des* «exercices de littérature». *»*

Dans une thèse en cours de développement au sein d'une université au Pakistan, nous avons pu identifier, dans l'explicitation méthodologique, le point vue suivant : « *The major concern of this research was to; examine the relationship between principals' conflict handling strategies and incivility of teachers in government degree colleges of the Punjab. This study used mixed methods research approach. Quan-qual design was applied. Teachers working in government degree colleges located in central Punjab were the target population of the study*. »

Elle prend elle-même ses références dans les écrits d'auteurs tel que Lauren H. Bryant (2011) dans *The Structure of Mixed Method Studies in Educational Research: A Content Analysis* ou encore John W. Creswell, Vicki L. Plano Clark (2017) dans *Designing and Conducting Mixed Methods Research.*

3 Mais qu'est-ce que la statistique ?

Nous reprenons ce que nous avons déjà écrit par le passé, en particulier, dans (Régnier, 2005). En nous appuyant sur l'approche de la didactique de la statistique, nous sommes conduit à porter notre attention sur le pôle *savoir* du système didactique qui renvoie à la question récurrente : qu'est-ce que la statistique ? Comme nous tentons de le schématiser (Figures ci-après), nous considérons la statistique comme un domaine scientifique qui se développe dans une tension dialectique entre la *statistique mathématique* et la *statistique appliquée à…* Cette relation dialectique est elle-même en tension dialectique avec les *statistiques* au sens des données construites. Nous avons discuté cette approche dans (Régnier 2002).

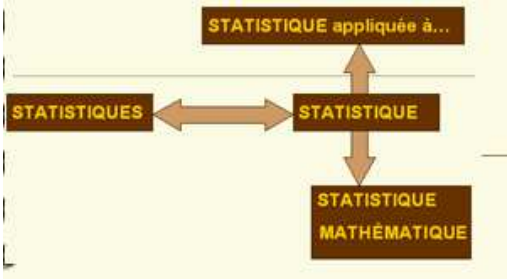


Figure 5 : Statistique (Régnier, 2005)

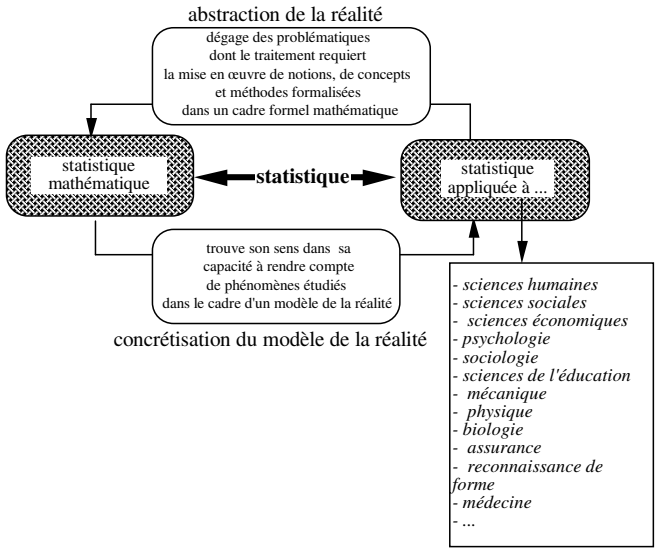


Figure 6 : Statistique (Régnier, 2002)

En ce qui concerne les usages du signifiant *statistique*, il en est un qui lui donne une fonction d'adjectif : analyse, variable, pensée, raisonnement, esprit statistique. En tant que substantif singulier, le terme statistique désigne le domaine scientifique. En ce sens nous parlons de didactique de la statistique comme de la didactique des mathématiques ou de la mathématique. Au pluriel, les statistiques, il désigne les données dont s'occupe la statistique. Ainsi nous considérons que l'expression : la didactique des statistiques ne convient pas pour désigner le cadre théorique pour penser la formation en statistique que nous ne confondrions pas avec une formation aux statistiques : par exemple, que serait alors une formation aux statistiques du chômage ?

Cette distinction est précisée chez des statisticiens de renom. Ainsi Daniel Schwartz considère que les statistiques sont des dénombrements de sujets, d'objets, d'évènements, dans des populations ou des sous-populations et que la statistique est un mode de pensée permettant de recueillir, de traiter et d'interpréter les données qu'on rencontre dans divers domaine. Pour Paul Deheuvels (Acad., 2000) les statistiques englobent la collecte et la gestion des grands fichiers d'observations alors que la statistique mathématique est la science du recueil et de l'interprétation des observations.

À côté des usages précédents, une statistique peut aussi être entendue comme une fonction des données échantillonnales, une variable aléatoire. La moyenne empirique, moyenne échantillonnale d'une variable statistique quantitative est une statistique. Si nous traitons de plusieurs statistiques, le pluriel réapparaît. Dans ce sens parler de didactique des statistiques comme il est possible de le rencontrer dans le domaine de l'éducation statistique, s'apparenterait

alors une expression du type : didactique des variables aléatoires. Mais peut-on parler d'une didactique d'un concept ?

Tout ceci peut ne paraître qu'une querelle de mots et que l'usage indifférencié des substantifs *statistique* et *statistiques* ne pose pas plus de problème que mathématique et mathématiques. Ceci n'est pas notre point de vue. Notre expérience professionnelle d'enseignement de la statistique et les observations systématiques que nous avons menées dans ce cadre, nous conduisent à émettre des conjectures sur l'origine de certains obstacles macrodidactiques dans leur lien avec cette confusion langagière. Des associations trop rapides entre les statistiques sociales par lesquelles l'apprenant se sent manipulé, et le cours de *statistique* confusément entendu comme cours de *statistiques*, sont propices à alimenter des résistances à l'apprentissage surtout quand en plus ces *statistiques* évoquent les *mathématiques*. Nous conjecturons que la confusion terminologique participe aussi d'obstacles épistémologiques à la constitution même du domaine de la didactique de la statistique au sein de la communauté des enseignants de mathématiques voire de celle des enseignants didacticiens.

Singulier ou pluriel, le substantif français utilisé pour la désignation du domaine scientifique, n'en résout pas pour autant la question de sa définition. Un regard historique nous conduit clairement à renoncer à obtenir une définition générale de la statistique. Dans ces conditions il y a tout lieu de croire que ce flou qui accompagne la délimitation du domaine de la statistique participe aussi d'un obstacle épistémologique à la construction de la didactique de la statistique. Nous partageons le point de vue étayé de Guy Brousseau (2004) pour qui l'histoire de la statistique est intimement liée à la conception du monde et à la culture. Sans pour autant tomber dans le piège d'un effet Jourdain, nous pensons que les êtres humains ont une lecture du monde fondée sur un raisonnement-en-acte statistique au sens développé par Gérard Vergnaud (1991) et qu'une part des décisions prises pour lancer, réguler, réorienter ou arrêter leur action s'appuie sur une conceptualisation statistique sous-jacente implicite. Les êtres humains pour vivre au quotidien appliquent des théorèmes-en-acte de statistique intégrant des concepts-en-acte de statistique. Évidemment des erreurs en découlent, par exemple, les confusions entre causalité et association, concomitance, corrélation. Cette confusion se retrouve d'ailleurs chez des individus lettrés ayant atteint des niveaux de conceptualisation élevés dans d'autres domaines scientifiques.

Une telle perspective s'accorde d'une caractérisation de la statistique conçue comme un instrument construit par les êtres humains pour lire et connaître le monde à partir de ses fragments, car ce monde ne peut jamais être appréhendé dans sa totalité spatiale et temporelle.

Au XIX^{ème} siècle, la statistique est une réponse instrumentale et conceptuelle pour distinguer les causes régulières perturbées par les causes fortuites pour connaître un phénomène. Cette position est exprimée en 1836 par A.A. Cournot (1843, 1984). La distinction entre causes régulières et causes fortuites constitue un point d'ancrage du raisonnement statistique qui requiert une certaine façon de lire le monde, un certain mode de pensée, un certain esprit que nous nommons esprit statistique. L'orientation de la conception de Cournot se fonde sur le principe de compensation. Notons que la moyenne (arithmétique), enseignée avant tout comme un algorithme de calcul, constitue un outil mathématique concordant avec ce principe. Il ressort que cette orientation s'inscrit dans une relation étroite entre trois domaines : statistique, mathématiques et probabilités. Au XXI^{ème} siècle, il convient d'y adjoindre le nouveau domaine de l'informatique.

Quand Yves Chevallard (1978) aborde, dans ses travaux pionniers, la question de la didactique de la statistique, il caractérise la problématique de la statistique — dans un sens proche de celui de Cournot — comme celle de la recherche et de la constitution d'une dialectique à caractère scientifique entre régularités et perturbations dans l'analyse des phénomènes marqués par la variabilité. C'est ce dernier concept qui nous semble fondateur de la statistique : la variabilité ! Ce concept est un formidable outil pour lire le monde. Chaque être humain s'y confronte quotidiennement qu'il s'agisse du temps météorologique, des durées de déplacement, etc. Chacun y est sensible et apporte des réponses aux problèmes que posent la variabilité, pour partie avec les concepts quotidiens, et ses concepts-en-acte, ses théorèmes-en-acte. Il doit être central dans la construction du champ de la didactique de la statistique.

Ceci nous amène à une autre définition de la statistique qui peut servir à organiser la transposition didactique dans les situations didactiques : sorte de langage commun, méthode générale reliant divers domaines scientifiques portant sur des ensembles d'individus, de variables et de relations conduisant à des conclusions plutôt vraisemblables et probables que vraies et certaines énonçant des propriétés de groupe valides sur des ensembles parfois mal définis.

Si nous souhaitons davantage insister sur l'orientation décisionnelle, nous dirions que la statistique peut être considérée comme un ensemble de méthodes permettant de prendre des décisions « bonnes » ou au moins « suffisamment bonnes » en situation incertaine. D'un point de vue praxéologique, cette orientation s'appuie sur l'idée générale que la méthode statistique pour étudier un phénomène consiste à associer un modèle aléatoire à ce phénomène, préciser ce modèle par l'observation, utiliser ce modèle pour prendre une décision.

3.1.1 La variabilité, concept fondateur de la statistique, objet du raisonnement statistique

L'observation de situations de la nature nous donne à voir la pertinence du concept de variabilité pour tenter de comprendre celles-ci voire de les expliquer. La photographie ci-contre est celle d'un échantillon de pommes récoltées sur un pommier, le même jour en septembre 2010 au même moment et placées dans ce cageot de bois. Ce cageot a été ensuite stocké dans une cave. Ce que montre la photographie est l'état de l'échantillon de pommes près de 6

Figure 7 : Variabilité pour rendre compte des objets de la nature (Photo de l'auteur)

mois après la cueillette. Nous pourrions construire une variable d'état de maturation des pommes selon 5 modalités ou plus pour réaliser une description statistique de cet échantillon.

Modalités					

Tableau 1 : Modélisation de la variable « état de maturation »

D'autres variables pourraient être construites comme la position spatiale dans le cageot, etc.

3.1.2 Autres concepts fondateurs de la statistique : représentativité et significativité

À côté de la variabilité, deux autres concepts fondateurs de la statistique sont à considérer : représentativité et significativité. La variabilité est la raison épistémologique majeure de la statistique. La représentativité fonde la validité des énoncés produits par l'analyse statistique en tant qu'étude de la variabilité des individus (unités statistiques) sur la population (univers statistique) au regard d'un (des) caractère(s) propre(s) au(x) phénomène(s) étudié(s)à partir de la partie (échantillon). La significativité fonde le degré de confiance du choix dans la décision.

Figure 8 : Concepts fondateurs de la statistique

Considérant la praxéologie du statisticien, les actions fondatrices de la statistique sont identifiables au travers des verbes suivant.

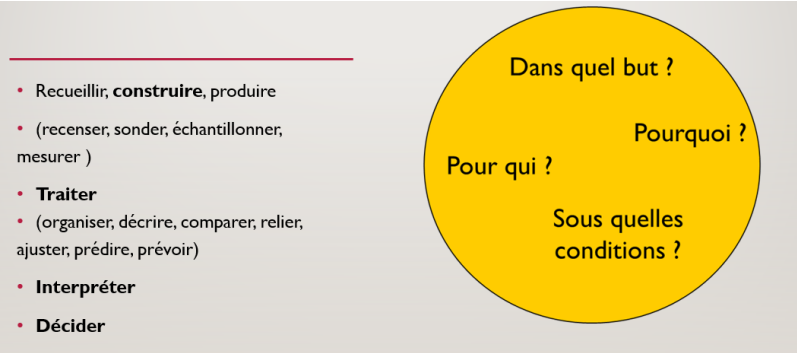


Figure 9 : Actions fondatrices de la statistique - Verbes d'actions de l'approche statistique

Ces actions fondatrices requièrent l'action préalable de modéliser. Nous n'explorerons pas ici le champ sémantique de modèle et modélisation. Le concept de modèle statistique et le processus de modélisation peuvent être abordés à différents niveaux de conceptualisation et selon diverses perspectives et dans différents cadres théoriques convoqués par le chercheur. Il est important d'identifier le cadre théorique dans lequel le chercheur formule les conclusions à partir des résultats des traitements et des analyses statistiques et conduit ses interprétations.

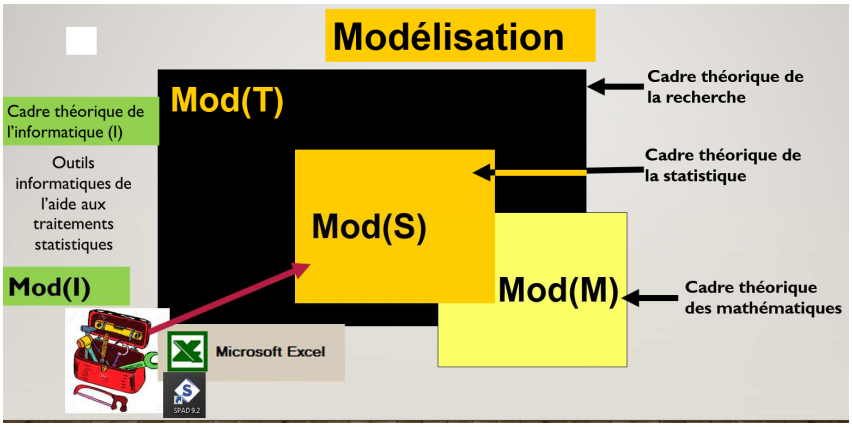


Figure 10 : Différents cadres théoriques de la modélisation

Nous identifions quatre cadres théoriques au sein desquels le chercheur procède à des opérations de modélisation : le cadre théorique constitué par le domaine disciplinaire ou scientifique de référence dans lequel il conduit sa recherche, le cadre théorique de la statistique qui mobilise ceux des mathématiques et de l'informatique. Nous reviendrons plus loin sur cette question dans l'exemple issu d'un article de revue scientifique (Revue de psychologie de l'éducation, 2001)

Pour conclure cette section portant sur *qu'est-ce la statistique ?* nous pourrions dire encore que en considérant l'idée générale selon laquelle les méthodes statistiques pour étudier un phénomène consistent à : 1) associer un modèle aléatoire à ce phénomène ; 2) préciser ce

modèle par l'observation ; 3) utiliser ce modèle pour prendre une décision, la statistique peut être ainsi définie comme un ensemble de méthodes permettant de prendre des décisions « bonnes » ou au moins « suffisamment bonnes » en situation incertaine. Il s'agit là d'une orientation guidée par la prise de décision risquée en situation incertaine.

3.2 Modélisation statistique minimale

Cette modélisation statistique minimale ou plutôt minimaliste requiert d'identifier les unités d'étude, appelées individus statistiques, qui sont regroupées dans un univers d'étude, la population statistique. Dans le langage mathématique, il s'agit d'un ensemble constitué d'éléments. Cet ensemble peut être fini, infini dénombrable ou même infini non dénombrable. La plupart du temps, l'accès à

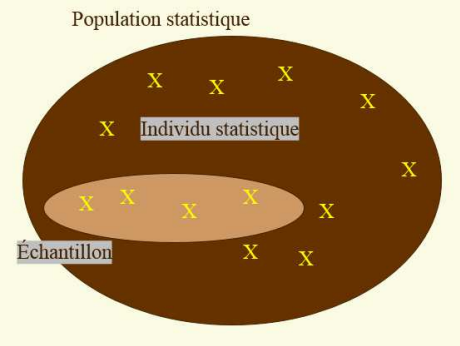


Figure 11 : Univers d'étude

l'univers d'étude dans toute son étendue est impossible ou même seulement non pertinent. Il convient alors de se contenter d'une partie de cet univers, appelée échantillon. La qualité de l'information repose essentiellement sur la qualité de l'échantillon de représenter au mieux le tout sachant que par construction des unités d'étude seront absentes logiquement de cette partie. L'opération qui consiste à accéder à la totalité des unités d'étude est le recensement, et celle qui consiste en la construction d'une partie, est le sondage. Comme nous le schématisons dans la figure ci-dessous, le sondage peut être fondé sur des méthodes aléatoires, c'est-à-dire que la présence d'une unité d'étude dans l'échantillon est le résultat du hasard, ou de méthodes dites non-aléatoires ou empiriques. Dans la procédure d'extraction d'un échantillon, deux situations se présentent qui consistent à remettre ou non une unité d'étude dans l'univers d'étude après son tirage. Dans le cas d'un tirage sans remise, l'échantillon est alors un sous-ensemble, au sens mathématique, de l'univers. Dans celui du tirage avec remise, l'échantillon n'est plus une partie au sens précédent, il s'agit alors d'une suite ordonnée d'unités d'étude. Au sens mathématique, si la taille de l'échantillon vaut n, alors l'échantillon est un n-uplet, élément de P^n produit cartésien de l'univers P. Mais pour des raisons pratiques de transposition, nous continuerons à parler d'échantillon comme une partie du tout !

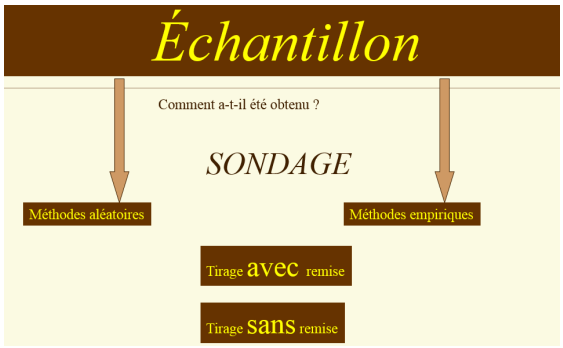


Figure 12 : Construction d'un échantillon

La question de la représentativité de l'échantillon a été très débattue à la fin du XIXème et au début du XXème au sein de la communauté internationale des statisticiens. Le débat reste récurrent mais le consensus actuel se maintient autour de l'idée que la meilleure façon d'assurer la représentativité d'un échantillon est de recourir au hasard. Ainsi un échantillon représentatif est un échantillon construit par des méthodes aléatoires. Ainsi défini, nous constatons que la représentativité est davantage caractérisée par le mode de construction que par sa taille. La taille de l'échantillon intervient davantage dans la précision des résultats des analyses statistiques. Elle joue un rôle dans la significativité.

Ayant établi les unités d'étude et l'univers d'étude, il nous faut alors recourir aux concepts-outils qui permettent d'étudier la variabilité. Il s'agit alors de considérer la variable statistique, c'est-à-dire la mise en correspondance, au sens mathématique, de chaque unité d'étude de l'univers d'étude avec un élément d'un ensemble de valeurs ou de modalités caractérisant ces unités d'étude. Nous avons schématisé cette opération dans la figure (gauche) ci-dessous.

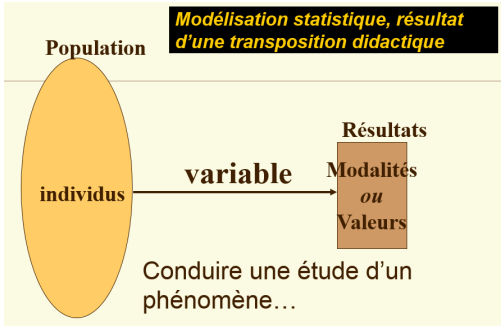


Figure 13 : Modèle statistique minimaliste

Le modèle ci-contre est une représentation sémiotique qui n'a pas de vertu opératoire. Il convient alors de recourir à une représentation en tableau dans lequel la colonne 1 représente les individus et la colonne 2 les résultats de la correspondance

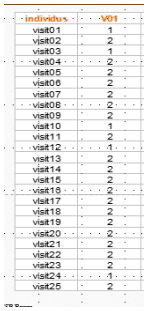


Figure 14 : Tableau de la série statistique

Il reste alors à procéder à une classification des variables statistiques. Nous représentons dans la figure ci-dessous les catégories identifiées.

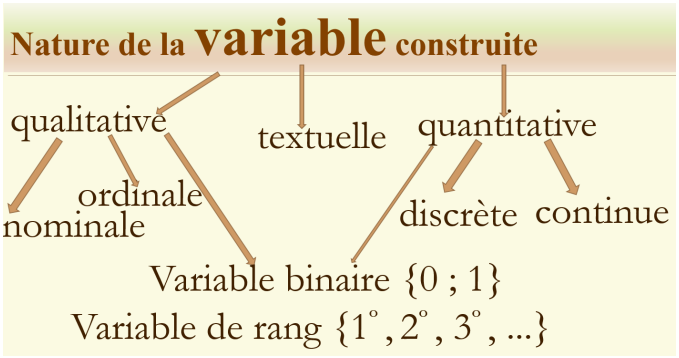


Figure 15 : Différentes catégories de variables

3.3 Raisonnement statistique, réflexion sur les variables statistiques et le problème des biais dans les enquêtes

Pour illustrer notre réflexion et compte tenu de la période de la pandémie Covid19 que nous venons vivre et qui a donné l'occasion de mettre en avant le rôle de l'épidémiologie fondée sur le domaine de la biostatistique, nous allons prendre l'exemple des recherches dans le domaine de la santé. Celles-ci sont basées sur des variables statistiques, nommées d'indicateurs de santé qui peuvent être mesurées directement et qui permettent de décrire l'état de santé des individus d'une communauté. Ils sont utilisés pour la mise au point d'indices plus complexes établis selon des formules spécifiques. La construction des données épidémiologiques comme celles de tout autre domaine est susceptible d'être biaisée.

Un biais dans une enquête épidémiologique[1], mais dans tout autre domaine aussi, désigne tout effet qui altère la représentativité des résultats. Il est caractérisé par une erreur systématique sur la représentation d'un effet. Il entraîne que la mesure de la relation maladie-facteur d'exposition, sur la population étudiée n'est pas égale à la mesure de cette relation dans la population-cible (univers d'étude). Il existe de nombreuses sources de biais que l'on peut schématiquement regrouper en 3 groupes.

3.3.1 Les biais de sélection

Ce type de biais réfère à une distorsion dans l'estimation d'un effet résultant de la façon par laquelle les sujets de la population étudiée ont été sélectionnés. Il peut concerner :

- les erreurs sur le choix des groupes à comparer dans tous les types d'enquêtes,
- les erreurs dans le choix du mode d'échantillonnage dans les enquêtes cas-témoins ou dans les études transversales,
- le biais de Berkson dans les études cas-témoins lorsque les cas et les témoins sont issus d'une population hospitalière non représentative de la population générale,

[1] D'après *Biostatistique Clinique, Épidémiologie et Essais cliniques*. ©Faculté de Médecine Necker -Enfants Malades 2002

- les sujets perdus de vue dans les études de cohortes et les sujets qui ne répondent pas aux demandes des enquêteurs, répétées au fil du temps, et indispensables au suivi d'une cohorte,
- la survie sélective dans les enquêtes cas-témoins ou dans les études transversales. La survie sélective concerne les profils différentiels de mortalité des cas et des témoins qui pourraient être à tort ignorés dans une étude rétrospective ou dans une étude transversale,
- les biais de détection dans les études cas-témoins, quand la procédure d'identification de la maladie varie avec l'exposition. Un exemple est illustré par la relation entre le cancer de l'endomètre et l'utilisation des œstrogènes. Ceux-ci peuvent entraîner des ménométrorragies qui poussent les femmes à consulter et augmenter ainsi la possibilité de diagnostiquer un cancer de l'endomètre chez les sujets exposées par rapport aux non-exposées.

Ainsi il existe de très nombreuses sources de biais de sélection dont il faut tenir compte à l'élaboration du protocole des enquêtes, et particulièrement dans les enquêtes cas-témoins.

3.3.2 Les biais de classification

Ils concernent une distorsion ou erreur systématique dans l'estimation d'un effet quand la mesure de la condition d'exposition ou de la maladie est systématiquement impropre. Ce biais représente des erreurs d'information qui peuvent conduire à une classification impropre des sujets aussi bien sur la maladie que sur le facteur d'exposition. Les origines les plus importantes sont liées à :

- l'utilisation d'appareils de mesure défectueux ou improprement réglés qui conduisent à des erreurs systématiques de classification,
- des critères diagnostiques impropres pour définir la maladie,
- des omissions ou des imprécisions sur des données enregistrées dans le passé,
- une surveillance inégale des sujets exposés et des sujets non exposés dans les études de cohortes.

Il faut prévenir ces biais de classification pour obtenir une meilleure estimation de la relation maladie-facteur d'exposition.

3.3.3 Les facteurs de confusion

La relation entre 2 variables peut être affectée par une troisième variable. Un facteur de confusion représente une variable qui est associée aussi bien au facteur étiologique vrai qu'à la maladie. Par exemple l'âge est un facteur de confusion dans l'étude de la relation entre tabagisme et cancer bronchique.

Nous avons vu précédemment qu'il fallait tenir compte des biais potentiels de sélection ou de classification au moment de l'élaboration du protocole d'une enquête. Il est illusoire, voire impossible, d'en tenir compte au moment de l'analyse.

Pour les facteurs de confusion plusieurs méthodes permettent de les prendre en compte soit lors de l'élaboration du protocole, soit au moment de l'analyse. À l'élaboration du protocole il s'agit de la stratification.

La stratification consiste dans un échantillon de malades et de témoins à former des classes de sujets par rapport aux facteurs de confusion. L'âge et le sexe sont deux facteurs de confusion pour l'étude de la relation entre tabac et cancer broncho-pulmonaire. On classera les sujets par tranche d'âge et par sexe.

L'appariement consiste à neutraliser les facteurs de confusion en groupant les sujets de telle sorte que ceux d'un même groupe partagent le(s) même(s) facteur(s) de confusion. Dans l'exemple précédent, chaque cas sera apparié avec un (ou plusieurs) témoin(s) de même âge et de même sexe. L'appariement n'est qu'une modalité particulière de stratification. Lors de l'analyse on peut utiliser des techniques statistiques dites d'ajustement. L'ajustement est un procédé qui vise à éliminer d'une comparaison de série d'observations le lien entre un effet et une ou plusieurs causes autres que celles qui sont le sujet propre de l'étude. La prévention des biais ou leur recherche lors de l'analyse constituent des étapes importantes dans une enquête afin de juger de la causalité entre un ou plusieurs facteurs de risque et une maladie.

3.3.4 L'identification des biais

Afin d'identifier des biais potentiels, il est bon de se poser quelques questions :

- la population de l'étude a-t-elle été bien définie ?
- est-ce que la population étudiée représente de manière adaptée la population cible ? (la population cible est la population pour laquelle on souhaitera généraliser les résultats de l'étude).
- les définitions de la maladie et de l'exposition sont-elles claires ?
- la définition des cas est-elle précise ?
- quels sont les critères d'inclusion et d'exclusion ?
- les contrôles représentent-ils de manière adéquate la population dont sont issus les cas ?
- l'identification ou la sélection des cas ou des contrôles a-t-elle pu être influencée le statut d'exposition ?
- les cohortes sont-elles similaires à l'exclusion du statut de l'exposition ?
- les mesures sont-elles aussi objectives que possibles ?
- l'étude est-elle réalisée le plus en aveugle possible ?
- le suivi est-il adapté ?
- le suivi est-il identique pour toutes les cohortes ?
- l'analyse est-elle appropriée ?
- l'interprétation qui en est faite est-elle étayée par les résultats ?

Il est clair que ce questionnement s’applique à l’ensemble des travaux de recherches dont l’argumentation se fonde sur une approche et raisonnement statistique.

3.4 Modélisation avancée dans le cadre théorique de la statistique mathématique

II s'agit ici de se placer à un niveau supérieur de conceptualisation tel que nous pourrions attendre d'un statisticien professionnel, par exemple. Voilà comment Jean-Pierre Lecoutre et Philippe Tassi abordent la question du modèle statistique « Le *problème général de la statistique inférentielle classique peut être représenté de la façon suivante : possédant un constat expérimental x appartenant à U, ensemble des résultats possibles d'une expérience, le statisticien suppose que U est muni d'une tribu d'événements B, et donc du couple (U, B) peut être probabilisé par une loi de probabilité P qui caractérise le phénomène étudié. En général, la loi P appartient à une famille de lois de probabilité indicée par une famille de paramètres* $P_\theta\ \theta \epsilon \Theta$ » (Lecoutre, Tassi, 1987, p.7)

Ici, nous ne développerons pas notre propos à ce niveau de conceptualisation.

4 Esprit critique, esprit statistique et raisonnement statistique

Nous reprenons le point de vue développé sur le site EDUSCOL de Ministère de l'éducation nationale français selon lequel « *L'esprit critique est à la fois un état d'esprit et un ensemble de pratiques qui se nourrissent mutuellement. Il n'est jamais acquis, il est une exigence, toujours à actualiser. Il naît et se renforce par des pratiques, dans un progrès continuel : on ne peut jamais prétendre le posséder parfaitement et en tous domaines, mais on doit toujours chercher à l'accroître.* » (Eduscol octobre 2016). L'esprit statistique est une dimension de l'esprit critique. Sophie Mazet (2016) y apporte un point de vue qui s'accorde tout à fait avec le nôtre en relation à l'esprit statistique quand elle met l'accent sur le fait que « *la question du point de vue (échantillonnage, suppression des données pertinentes) et la réflexion sur les outils statistiques, la construction des sondages notamment, constituent un moment essentiel de la construction intellectuelle. Cette réflexion permet d'éviter les paralogismes, (i.e. un raisonnement faux perçu rigoureux par le producteur comme le récepteur de bonne foi tous les deux, en raison du fait qu'il s'appuierait sur des statistiques).* »

Nous nous appuyons sur l'idée que « *l'esprit statistique naît lorsqu'on prend conscience de l'existence de fluctuation d'échantillonnage* » comme stipulé dans les textes d'orientation de la réforme des programmes de mathématiques au lycée en 2000 en France. La formation de l'esprit statistique passe par une formation en statistique qui requiert un certain renoncement à l'usage systématique de l'idée de vérité pour chercher à maîtriser celle de plausibilité.

4.1 Opérations logiques à l'œuvre dans le raisonnement statistique

Au cours de nos expériences d'enseignement de la statistique, nous avons fréquemment été confronté à la question suivante : comment interpréter en statistique ? comment apprend-on à interpréter en statistique ? et pour nous-même comment peut-on enseigner l'interprétation statistique ?

À ce jour encore, il semble que l'interprétation statistique ne peut être enseignée qu'au travers d'une pratique réflexive portant à la fois sur le processus et sur le résultat, en organisant des situations de réalisation d'interprétations statistiques soumises à des débats entre apprenants. Le point de vue exprimé par Brigitte Escoffier et Jérôme Pagès (Escoffier & Pagès 1990 p.217-218) nous fournit un appui intéressant

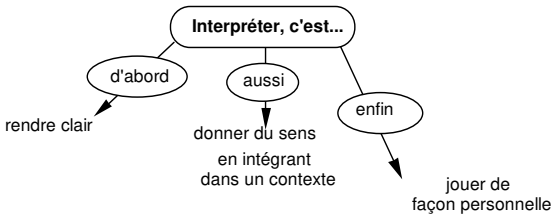


Figure 16 : Interprétation statistique selon Escoffier et Pagès

Cette interprétation statistique est alors soumise aux contraintes de trois opérations logiques que nous avons désignées par induction, déduction et éduction. Il nous semble que le développement de *l'esprit statistique* consiste en leur acquisition et en leur *manipulation consciente* pour conduire un *raisonnement statistique* et étayer l'activité d'*interprétation statistique*. Nous entendons par éduction, l'opération par laquelle « une cause efficiente, agissant sur une matière, y fait apparaître une forme déterminée » en nous inspirant de Lalande (1926, 1991 p.266)

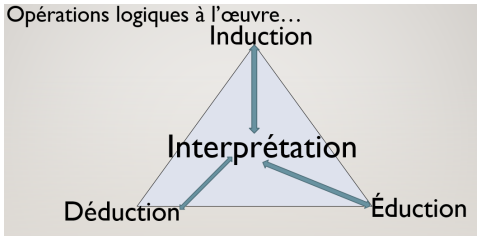


Figure 17 : Interprétation sous la contrainte des opérations logiques

Dans nos travaux publiés (Régnier, 2000), nous avons tenté, à l'aide du schéma ci-dessous, de situer ce questionnement sur l'interprétation statistique dans l'articulation générale des situations d'enseignement et d'apprentissage au sein du parcours de formation en statistique vers un niveau d'éducation statistique qui s'élève à chaque étape.

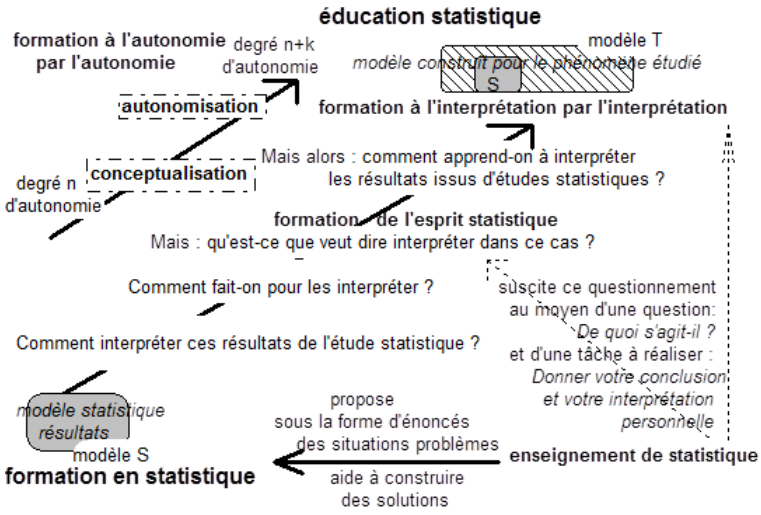


Figure 18 : Schématisation de la place de la question de l'interprétation soulevée par les étudiants dans le parcours de formation et du rôle de l'enseignement. (Régnier, 2000 p.139)

4.2 Le raisonnement statistique au service de la curiosité et du développement de la connaissance

La recherche scientifique est mue par le désir humain de connaître le monde par curiosité mais aussi par nécessité pour vivre ou survivre. Les problèmes surgissent quand les êtres humains sont confrontés, dans des contextes et des circonstances données, à des situations portant, par exemple, sur des propriétés d'objets ou de phénomènes qui sont certaines mais inconnues pour eux. Prenons en exemple celui de la durée de vie d'un appareil ménager. Il s'agit là d'une préoccupation des services de contrôle de fiabilité et de qualité dans le domaine de la fabrication. Nous pouvons considérer que cette durée de vie jusqu'à la première panne est certaine pour chaque objet mais totalement inconnue pour l'observateur (acheteur, utilisateur, vendeur ou fabricant). Une approche serait de faire fonctionner tout le lot fabriqué à un moment jusqu'à la première panne. Sur le plan économique, cette approche est peu intéressante car elle conduit le fabricant à conserver l'ensemble de sa production et donc de ne rien vendre, c'est-à-dire, cesser son activité pour cause de faillite ! Une autre démarche qui prend appui sur le raisonnement statistique conduit à abandonner la situation (certain, inconnu) pour se placer dans la situation (incertain, connu). Pour reprendre nos propos tenus au cours des sections précédentes, il s'agit de formuler des énoncés de propriétés plus plausibles, vraisemblables et probables que vrais et certains. Dans notre exemple, cela signifie avoir recours à un échantillonnage aléatoire extrait du lot d'appareils ménagers et de procéder à l'expérience de contrôle.

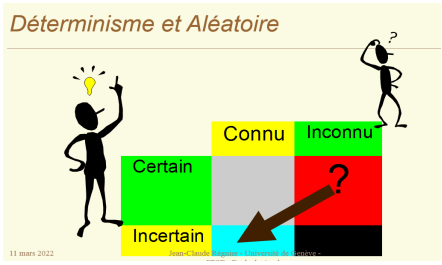


Figure 19 : Passage de l'inconnu certain au connu incertain.

C'est sur cette perspective que se fondent les procédures d'estimation statistique et celles de tests statistiques d'hypothèse. Dans l'exemple simplifié que nous avons pris ici, il s'agirait de procéder à une estimation statistique de la durée moyenne avant la première panne. C'est sur cette information que le fabricant va établir la durée de la garantie !

4.3 Raisonnement statistique à l'œuvre dans l'estimation statistique

L'estimation statistique consiste, à l'aide d'un estimateur, à fournir des valeurs possibles et probables à des caractéristiques certaines mais inconnues d'une population, à partir d'un échantillon représentatif issu de celle-ci. Dans l'idéal, pour que puissent être appliquées les outils mathématiques de la statistique, il convient que cet échantillon respecte les conditions d'un échantillonnage aléatoire simple. Dans la figure ci-dessous, nous avons schématisé la mise en œuvre d'une procédure d'estimation statistique de la proportion π certaine mais inconnue d'un caractère dans une population d'étude. Ce schéma vise à montrer comment l'information acquise à partir du calcul de la proportion p_i sur l'échantillon E_i comme réalisation de l'estimateur nommé proportion empirique ou échantillonnale va dépendre de ce même échantillon E_i

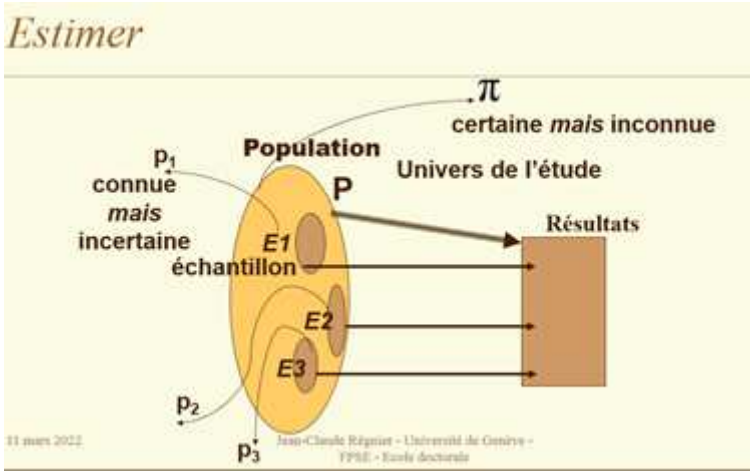


Figure 20 : Procédure d'estimation statistique d'une proportion

Cette procédure conduit à fournir une valeur plausible mais non certaine de la proportion π sur la population d'étude. Il ne s'agit pas d'une valeur approchée au sens déterministe de

l'approximation en mathématiques, comme par exemple 1.414 est une valeur décimale approchée à 1/1000 par défaut du nombre réel algébrique $\sqrt{2}$. Il s'agit d'une valeur possible qui dépend du choix de l'échantillon. La justification est fournie par une théorie, celle de l'échantillonnage, qui relève du cadre théorique de la statistique mathématique. L'idée vient du fait que si nous considérons l'ensemble de tous les échantillons E_i possibles de taille n sur la population de taille N, et si nous calculons la proportion p_i alors nous constaterions que nombre de ces valeurs sont proches de la valeur certaine π. Cette justification passe par des modèles mathématiques et des relations algébriques que nous n'abordons pas ici. Il n'est pas même requis pour un usager non spécialiste de maîtriser ces fondements mathématiques pour un emploi raisonné. Ce niveau de conceptualisation n'est pas nécessaire si on admet les propriétés mathématiques en faisant confiance aux statisticiens qui traitent de la statistique-objet.

La vérification directe qui est théoriquement possible, s'avère rapidement impossible si on tient compte du nombre d'échantillons possibles. À titre d'exemple, considérant une population d'étude P de taille N, c'est-à-dire card(P)=N, le nombre d'échantillons de taille n sans remise, c'est-à-dire correspondant exactement à un sous-ensemble de P, est obtenu par la formule suivante :

$$C_N^n = \binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

Nous rappelons la signification de la notation $N! = 1\text{x}2\text{x}\ldots\text{x}N$ c'est-à-dire une manière condensée de décrire le produit des nombres entiers de 1 à N.

En ce qui concerne le nombre d'échantillons avec remise, c'est-à-dire le nombre de n-uplets possibles que nous pouvons obtenir par tirage au sein de la population P est obtenu par la formule suivante : N^n

Si la population est finie, le cardinal de l'ensemble des échantillons reste fini mais rapidement très grand. Si la population est infinie, alors le nombre d'échantillons l'est aussi.

À titre d'illustration, imaginons une population de petite taille N=20. En ce qui concernent les échantillons sans remise de taille n=10, nous en dénombrons $C_{20}^{10} = \frac{20!}{10!10!} = 184756$. Pour les échantillons avec remise $N^n = 20^{10} = 20000000000$. Même dans ces conditions d'effectifs réduits, la prise en considération de tous les échantillons pour y réaliser les calculs reste pratiquement impossible.

L'estimation statistique peut donc être abordée comme l'exemple précédent de l'estimation d'une proportion π certaine mais inconnue sur la population d'étude ; il s'agit de réaliser une estimation ponctuelle. Mais elle peut être aussi abordée sous forme dite estimation par intervalle

ou estimation ensembliste, ou encore, dans le langage courant, fourchette d'estimation. Sous certaines conditions, il est alors possible de construire un ensemble, souvent un intervalle dont on connait la probabilité, appelée la confiance, de l'évènement : l'intervalle proposée contient la valeur certaine inconnue. $Prob\{\ [a;b] \ni \pi\} = 1 - \alpha$

Il ressort donc de l'approche fournie par l'estimation statistique tant ponctuelle que par intervalle de confiance que les résultats obtenus doivent toujours être interprétés sous l'angle de la plausibilité et non celui de la certitude. C'est en ce sens qu'est mobilisé le raisonnement statistique.

4.4 Raisonnement statistique à l'œuvre dans les tests d'hypothèse

Dans les travaux de recherche, le chercheur est conduit dans certaines situations à formuler des hypothèses en tant que réponses anticipées à la question centrale de sa problématique. Une catégorie de ces hypothèses réunit celles dont les énoncés admettent une traduction dans le cadre théorique de la statistique et correspondent à un modèle statistique. Par exemple, dans une étude sur la question du genre en mathématiques, nous pouvons nous intéresser à une comparaison des performances dans la résolution de problèmes. Une croyance ancienne colporte que les cerveaux des femmes ne sont pas structurés comme ceux des hommes pour l'acquisition des connaissances et le développement des compétences en mathématiques. Sur le site[2] suisse *Apprendre en ligne. Ressources pour les enseignants et les élèves du secondaire II. Statistiques interactives concernant la Suisse*, un article déposé en 2007 et modifié en 2010, est consacré à la thématique Femmes et mathématiques : différences génétiques ou stéréotypes sociaux ? Il est encore courant en France d'entendre des déclarations affirmant que les hommes sont plutôt *scientifiques* alors que les femmes, plutôt *littéraires* ! Sur la page du site, il est rappelé que « *une étude historiographique sur les femmes et les mathématiques montre que la participation des femmes à l'activité mathématique est étroitement liée à leur rôle et leur position dans la société* » Nous pouvons donc concevoir une question centrale sous la forme : *Les femmes et les hommes sont-ils inégaux face aux mathématiques ?*

En réponse à cette question, nous pouvons énoncer les hypothèses, c'est-à-dire, des réponses hypothétiques possibles. Un premier énoncé hypothétique que nous désignons par H_0 peut être formulé ainsi : il n'existe pas de différence entre les performances des femmes et celles des hommes dans la résolution d'un ensemble de problèmes de mathématiques. Nous pouvons

[2] https://owl-ge.ch/prospective/article/femmes-et-mathematiques-differences-genetiques-ou-stereotypes-sociaux

ensuite formuler des énoncés alternatifs que nous désignerons par H_1, H_1', H_1'' par simple négation de H_0. Nous obtenons ainsi :

- H_1 Il existe une différence entre les performances des femmes et celles des hommes dans la résolution d'un ensemble de problèmes de mathématiques
- H_1' Il existe une différence et les performances des femmes sont supérieures à celles des hommes dans la résolution d'un ensemble de problèmes de mathématiques
- H_1'' Il existe une différence et les performances des femmes sont inférieures à celles des hommes dans la résolution d'un ensemble de problèmes de mathématiques

Il s'agit alors de choisir entre l'hypothèse H_0 et l'une des trois hypothèses alternatives quelle est la plus plausible au vu des données construites. Celles-ci peuvent être construites à partir des traces écrites des réponses fournies à l'ensemble des problèmes de mathématiques par les individus d'un échantillon construit par le chercheur. Remarquons que l'ensemble des problèmes de mathématiques est lui-même une construction échantillonnale raisonnée au sein du domaine des mathématiques.

Revenons au raisonnement statistique mobilisé dans ce processus de prise de décision. Il s'agit là d'une prise de décision en situation incertaine, puisque cela concerne un univers d'étude, c'est-à-dire, l'ensemble des femmes et des hommes, dont nous pouvons percevoir ici le flou de la frontière, alors que l'information est réduite aux performances sur un échantillon d'individus soumis à une épreuve de résolution de problèmes de mathématiques, eux-mêmes dépendant du choix et de la nature de ces problèmes dans leurs contenus mathématique et extra-mathématique.

Le pivot de notre raisonnement statistique est fixé sur l'hypothèse H_0. Ainsi logiquement, cet énoncé est vrai ou faux mais cette propriété nous est inconnue. Le chercheur sera alors conduit à considérer l'hypothèse H_0 soit comme vraisemblable soit comme faux-semblable. Le tableau ci-dessous (Tableau 2) montre les conséquences de cette prise de décision auxquelles il est impossible de se soustraire. L'approche statistique apporte un cadre théorique permettant de développer des outils d'aide à cette prise de décision et au contrôle des risques encourus.

		« Situation de nature » inconnue	
		Ho vraie H1 fausse	H1 vraie Ho fausse
Décision du chercheur	Rejeter Ho Conserver H1	Erreur de première espèce α	Correcte
	Conserver Ho Rejeter H1	Correcte	Erreur de seconde espèce β

Tableau 2 : Décision prise par le chercheur

		« Situation de nature » inconnue	
		Ho vraie H1 fausse	H1 vraie Ho fausse
Décision du chercheur	Rejeter Ho Conserver H1	Prob{Rejeter Ho sachant Ho vraie}= α	Prob{Conserver H1 sachant H1 vraie}= 1–β
	Conserver Ho Rejeter H1	Prob{Conserver Ho sachant Ho vraie}= 1–α	Prob{Rejeter H1 sachant H1 vraie}= β

Tableau 3 : Niveau du risque encouru

C'est en ce sens qu'ont été construits les tests statistiques. Il s'agit de procédures d'aide à la prise de décision relative au choix préférentiel, fondé sur la plausibilité, d'une des deux hypothèses à partir des informations obtenues sur un échantillon supposé représentatif de l'univers d'étude.

Cette procédure s'organise ainsi :

- Formuler les hypothèses H_0 *et l'une des hypothèses alternatives* H_1
- Déterminer ou construire une variable de décision D
- Déterminer la forme de la région critique en fonction de l'énoncé de H_1
- Déterminer la région critique en fonction du niveau de risque α de 1ère espèce
- Calculer quand cela est possible la puissance 1-β où β est le niveau de risque de 2ème espèce
- Calculer la valeur empirique de la variable de décision D
- Conclure entre le rejet ou le non-rejet de H_0

Dans la pratique, il est d'usage de fixer *a priori* la valeur α du niveau de risque de 1ère espèce, en général α=0.05 ou 0.01, ce qui revient à privilégier l'hypothèse H_0, historiquement appelée hypothèse nulle et ceci pour diverses raisons parmi lesquelles, reprenant le point de vue de Gilbert Saporta (1990, p.320) nous pourrions citer :

- puisque le chercheur ne veut pas abandonner trop souvent l'hypothèse H_0, elle doit être solidement établie et non contredite jusqu'alors ;
- l'hypothèse H_0 est celle à laquelle le chercheur tient particulièrement même pour des raisons subjectives ;
- l'hypothèse H_0 correspond à une hypothèse de prudence. Par exemple, concernant l'innocuité d'un vaccin, il est prudent de partir d'un point de vue défavorable à ce nouveau vaccin ;
- l'hypothèse H_0 est la seule que le chercheur peut formuler avec précision et caractérisations mathématiques.

Observons que α étant fixé *a priori,* le niveau β de l'erreur de 2nde espèce peut être calculé mais à condition de connaître les lois de probabilités sous l'hypothèse alternative H_1 Il faut par ailleurs savoir que les deux niveaux de risque α et β varient en sens inverse : diminuer α conduit à augmenter β !

Il existe à ce jour un nombre impressionnant de tests statistiques d'hypothèses. Le chercheur doit alors choisir avec précaution le test statistique adapté au problème d'étude auquel il se confronte afin de ne pas s'exposer à un 3ème type d'erreur, celui de recourir à une procédure non pertinente !

Dans un article (Régnier, 1998b) nous avions abordé cette question de la prise de décision risquée en situation incertaine en la mettant en œuvre dans une situation didactique visant l'acquisition du raisonnement statistique.

5 Raisonnement statistique et compétences en statistique pour la compréhension de l'usage des concepts statistiques dans des articles courants ou scientifiques…

Nous allons examiner ici quelques exemples où le raisonnement statistique est sollicité avec pertinence. Par le passé, nous avons publié une analyse d'un article de journal quotidien portant sur le thème de accidents de la route (Régnier, 1998a). Il s'agissait de voir comment un article de presse régionale que nombre de lecteurs parcourent en diagonale, s'avère requérir un traitement parallèle écrit explicitant le modèle mathématique sous-jacent pour parvenir à la compréhension et au contrôle de la validité de ce qui est énoncé.

5.1 Petit retour sur le vocabulaire de la statistique

Il convient de prendre garde aux ambiguïtés du vocabulaire. L'exemple de la variance illustre cette nécessaire attention. Variance, ce peut être le nom d'un sous-vêtement féminin tel que les usagers du TGV peuvent le constater sur des panneaux publicitaires affichant des objets de la mode à ne pas confondre avec le mode en statistique, sur les quais de gare en France

Figure 21 : Point de vue sémantique sur la variance dans la mode comparé à celui de la statistique (extrait d'une photographie d'un panneau publicitaire dans une gare SNCF -France)

Pour une variable statistique quantitative discrète finie, notée X, la variance ou fluctuation, notée V(X) ou σ_X^2 est la moyenne (arithmétique) des carrés des écarts des valeurs à la moyenne (arithmétique), notée m_X *ou* $\bar{X}$ de cette variable sur un univers d'étude: population de taille N ou échantillon de taille n. Sous forme symbolique dans le cas d'un échantillon, nous pouvons représenter la moyenne et la variance par des formules algébriques à partir de la distribution statistique des k couples (x_i, n_i) présentée dans un tableau statistique :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{i=k} n_i\, x_i$$

$$V(X) = \frac{1}{n}\sum_{i=1}^{i=k} n_i (x_i - \bar{X})^2$$

En statistique, nous pouvons remarquer que tant la moyenne que la variance sont définies par des algorithmes qui permettent d'obtenir respectivement ces valeurs sur un univers d'étude pour des variables statistiques dont les propriétés mathématiques rendent possible la réalisation de cet algorithme de calcul.

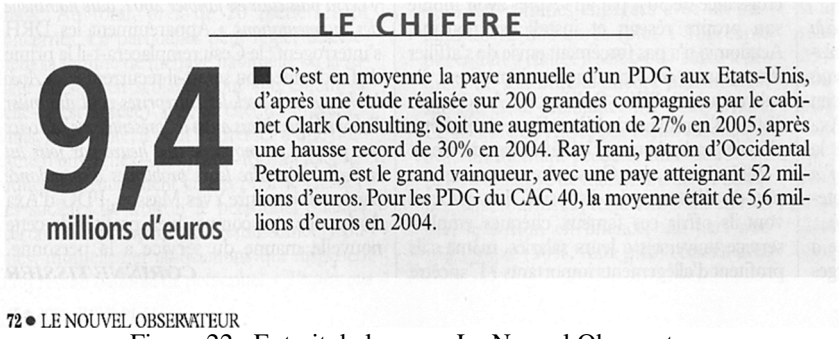
L'origine épistémologique de ces deux concepts est à chercher en partie dans la recherche d'optimisation. La moyenne arithmétique est une solution du problème de la détermination

d'une ou plusieurs valeurs rendant minimum la somme des écarts entre chaque valeur de réalisation de la variable quantitative statistique que nous limitons au cas discrète et finie dans notre propos, et cette ou ces valeurs. Sous forme symbolique, notre problème est modélisé sous l'équation suivante. Notant θ cette inconnue, l'équation devient $f(\theta) = \sum_{i=1}^{i=k} n_i(x_i - \theta) = 0$
Sa résolution n'offre aucune difficulté et conduit au résultat suivant :
$f(\theta) = \left(\sum_{i=1}^{i=k} n_i\, x_i\right) - \left(\sum_{i=1}^{i=k} n_i\, \theta\right) = \left(\sum_{i=1}^{i=k} n_i\, x_i\right) - \left(\sum_{i=1}^{i=k} n_i\right)\theta = 0$ d'où $\theta = \frac{1}{n}\sum_{i=1}^{i=k} n_i x_i$
Nous pouvons ainsi reconnaître l'origine de l'algorithme de calcul de la moyenne arithmétique qui est l'unique valeur réalisant l'équation $f(\theta) = 0$.

Si nous nous intéressons maintenant à la variance, il s'agit alors de trouver une solution à un problème de détermination d'une ou plusieurs valeurs rendant minimum la somme des carrés des écarts entre chaque valeur de réalisation de la variable et cette valeur. La fonction dont on cherche le minimum est alors ainsi formulée : $g(\theta) = \sum_{i=1}^{i=k} n_i(x_i - \theta)^2$ En étudiant sa variation parabolique, on constate que cette fonction atteint son minimum quand θ est égale à la moyenne arithmétique. La variance est alors la valeur $V(X) = \frac{1}{n} g(\bar{X})$

5.2 La moyenne a-t-elle un sens ? Usage de la moyenne dans la presse

Considérons cet extrait de l'hebdomadaire français *Le Nouvelle Observateur*. Abordant le concept de moyenne dans un cours de statistique en licence et master de Sciences de l'éducation et de la formation, il est ressorti une discussion sur le sens même de la moyenne appuyée sur des arguments propres à laisser dubitatif. Parmi les arguments, nous pourrions citer cet exemple caricatural relatif à la température moyenne : imaginons une main dans un réfrigérateur et l'autre dans un four, nous pourrions conclure que nous jouirions d'une température moyenne fort agréable !
Nous avons présenté cet extrait ! Nous avons alors observé plusieurs réactions émotionnelles desquelles nous avons déduit que la moyenne pouvait avoir un sens.


LE CHIFFRE

9,4
millions d'euros

■ C'est en moyenne la paye annuelle d'un PDG aux Etats-Unis, d'après une étude réalisée sur 200 grandes compagnies par le cabinet Clark Consulting. Soit une augmentation de 27% en 2005, après une hausse record de 30% en 2004. Ray Irani, patron d'Occidental Petroleum, est le grand vainqueur, avec une paye atteignant 52 millions d'euros. Pour les PDG du CAC 40, la moyenne était de 5,6 millions d'euros en 2004.

72 ● LE NOUVEL OBSERVATEUR

Figure 22 : Extrait de la revue Le Nouvel Observateur

Recevoir 9,4 millions d'euros par année ne laisse personne indifférent quand bien même il ne s'agit que d'une paye moyenne ! Il est aussi loisible de constater que ce court article mobilise à la fois des connaissances et des compétences en statistique et en mathématiques. Nous pouvons ainsi voir l'importance de certaines informations par leur absence, par exemple : la valeur minimale des émoluments, la fluctuation autour de la moyenne fournie par la variance.

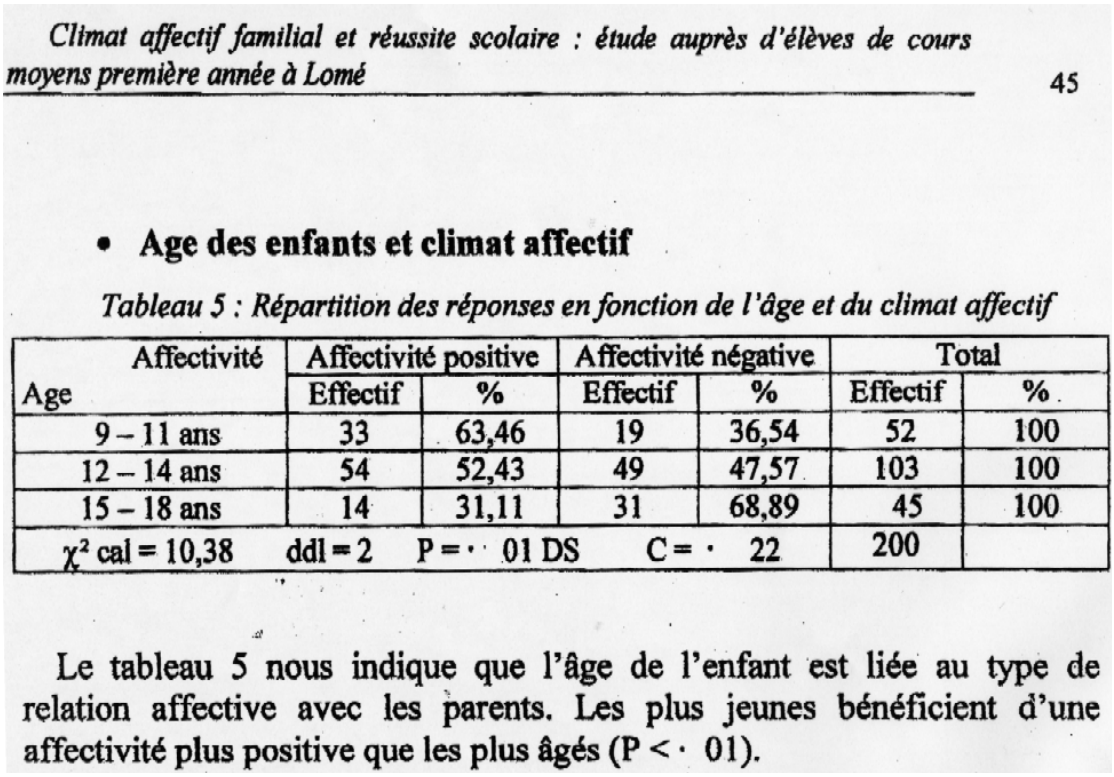
5.3 Lecture-compréhension d'un article scientifique

Nous allons prendre l'exemple extrait d'un article paru en 2001 dans la Revue de Psychologie de l'éducation qui est une publication de l'Université François Rabelais de Tours (France). Cet article écrit par Yanakou Koffiwai Gbati porte sur la thématique du climat affectif familial et réussite scolaire. Il s'agit d'une étude réalisée auprès d'élèves de cours moyen première année à Lomé (Togo).

Nous rapportons le résumé :

> *« Cet article est une première tentative de recherche de liens entre l'affectivité familiale et la réussite scolaire des élèves dans un milieu africain, en particulier, à Lomé (Togo). L'étude a porté sur des élèves des classes de CM1 à partir d'un questionnaire sur les relations parents-enfants. Il ressort des investigations que le résultat scolaire est lié au climat affectif dans lequel vit l'enfant : le climat affectif positif va de pair avec le bon rendement scolaire même dans des conditions matérielles difficiles d'existence. Néanmoins, dans un climat affectif défavorable, les résultats des garçons sont supérieurs à ceux des filles. »*

Ici nous nous restreignons à un des traitements, celui du croisement entre deux variables statistiques : la variable Age et la variable Climat affectif. Nous reproduisons l'extrait tel quel.



Climat affectif familial et réussite scolaire : étude auprès d'élèves de cours moyens première année à Lomé 45

- **Age des enfants et climat affectif**

Tableau 5 : Répartition des réponses en fonction de l'âge et du climat affectif

Affectivité / Age	Affectivité positive		Affectivité négative		Total	
	Effectif	%	Effectif	%	Effectif	%
9 – 11 ans	33	63,46	19	36,54	52	100
12 – 14 ans	54	52,43	49	47,57	103	100
15 – 18 ans	14	31,11	31	68,89	45	100
χ^2 cal = 10,38	ddl = 2	P = · · 01 DS		C = · 22	200	

Le tableau 5 nous indique que l'âge de l'enfant est liée au type de relation affective avec les parents. Les plus jeunes bénéficient d'une affectivité plus positive que les plus âgés (P < · 01).

Figure 23 : extrait de l'article de Yanakou Koffiwai Gbati (2001)

Bien plus que dans l'article de la Revue *Le Nouvel Observateur*, la lecture-compréhension de cet extrait exige un niveau élevé de connaissances et de compétences dans le domaine de la statistique. Le raisonnement s'appuie sur la mise en œuvre d'un test statistique d'hypothèse : il s'agit du test du χ^2 d'indépendance, ainsi nommé dans le jargon méthodologique. Le tableau (Tableau 5 de l'article) rendant compte des données sur lesquelles le raisonnement prend appui est un tableau complexe qui en intègre deux avec les informations propres à la démarche adoptée.

Nous procédons à la décomposition de ce tableau complexe. Un premier tableau est à considérer, appelé tableau de contingence, qui est le tableau statistique conjoint de la variable **Age**, considérée comme une variable qualitative ordinale à trois modalités : les tranches d'âge 9-11ans, 12-14ans et 15-18ans, avec la variable qualitative nominale **Climat affectif** qui comporte deux modalités : affectivité positive, affectivité négative.

Affectivité / Age	Affectivité positive	Affectivité négative	Total
9 – 11 ans	33	19	52
12 – 14 ans	54	49	103
15 – 18 ans	14	31	45
Total	101	99	200

Tableau 4 : Tableau de contingence n=200

Celui-ci est un tableau 2x3 car il comporte trois lignes correspondant aux trois modalités de la variable V1=Age et deux colonnes correspondant aux deux modalités de la variable V2=Climat affectif.

Un second tableau est à extraire et à compléter. Il s'agit du tableau des profils-lignes, c'est-à-dire des fréquences conditionnelles donnant les proportions respectives des individus ayant exprimé une affectivité positive et une affectivité négative dans chaque tranche d'âge.

Affectivité Age	Affectivité positive	Affectivité négative	Total (%)
9 – 11 ans	$\frac{33}{52} \approx 63{,}46\%$	$\frac{19}{52} \approx 36.54\%$	100
12 – 14 ans	$\frac{54}{103} \approx 52{,}43\%$	$\frac{49}{103} \approx 47{,}57\%$	100
15 – 18 ans	$\frac{14}{45} \approx 31{,}11\%$	$\frac{31}{45} \approx 68{,}89\%$	100
Total %	$\frac{101}{200} = 50{,}5\%$	$\frac{99}{200} = 49{,}5\%$	100

Tableau 5 : Tableau des profils-lignes (fréquences conditionnelles) n=200

Nous pouvons changer de registre sémiotique de représentation en utilisant une représentation graphique

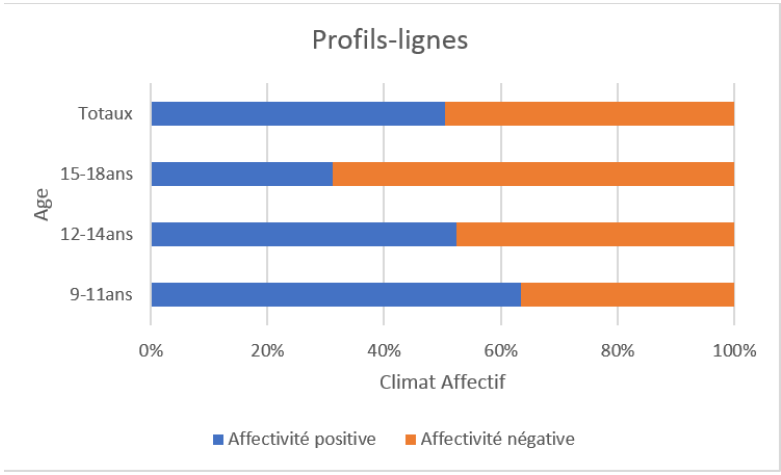


Figure 24 : Représentation graphique des profils-lignes contingents

Il s'agit alors de comparer les barres bleues entre elles correspondant la modalité affectivité positive. Nous pouvons observer l'ordre d'importance de l'expression de l'affectivité positive décroissant en fonction de l'âge. Dans l'idée de prolonger cette propriété au-delà de l'échantillon, il convient de recourir à des outils statistiques d'aide à la prise de décision adaptés afin de juger du caractère significatif ou non des différences des profils-lignes.

Sous l'hypothèse Ho d'une absence de différence entre les groupes d'âge ou d'homogénéité des groupes par rapport à l'affectivité ou encore d'indépendance (statistique) de la variable V1=âge et de la variable V2=Climat affectif, nous devrions avoir cette représentation graphique.

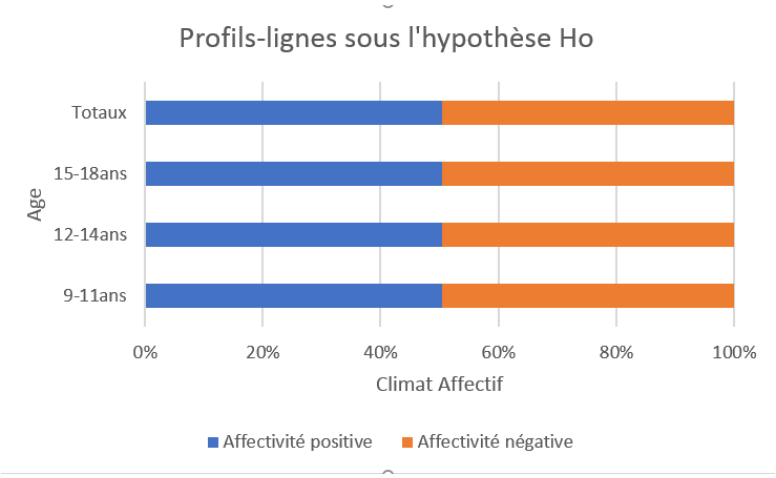


Figure 25 : Représentation graphique des profils-lignes sous l'hypothèse Ho d'absence de lien entre les deux variables V1 et V2

Cela conduirait alors à obtenir un tableau de profils-lignes théoriques suivant :

Affectivité / Age	Affectivité positive	Affectivité négative	Total (%)
9 – 11 ans	$\frac{T_{11}}{52} \approx 50.5\%$	$\frac{T_{12}}{52} \approx 49.5\%$	100
12 – 14 ans	$\frac{T_{21}}{103} \approx 50.5\%$	$\frac{T_{22}}{103} \approx 49.5\%$	100
15 – 18 ans	$\frac{T_{31}}{45} \approx 50.5\%$	$\frac{T_{32}}{45} \approx 49.5\%$	100
Total %	$\frac{101}{200} = 50{,}5\%$	$\frac{99}{200} = 49{,}5\%$	100

Tableau 6 : Tableau des profils-lignes (fréquences conditionnelles) sous l'hypothèse Ho de l'indépendance des deux variables (n=200)

Nous pouvons en déduire les relations d'égalité suivantes relatives aux proportions :

$$\frac{T_{11}}{52} + \frac{T_{21}}{103} + \frac{T_{31}}{45} = \frac{T_{11} + T_{21} + T_{31}}{52 + 103 + 45} = \frac{101}{200}$$

Puis en réalisant les calculs suivants, nous obtenons le tableau d'effectifs théoriques T_{ij}. Par exemple, l'effectif théorique T_{11} est obtenu par $T_{11} = \frac{52x101}{200} = 26.26$

Affectivité / Age	Affectivité positive	Affectivité négative	Total
9 – 11 ans	$T_{11} \approx 26.26$	$T_{12} \approx 25.74$	52
12 – 14 ans	$T_{21} \approx 52.015$	$T_{22} \approx 59.985$	103
15 – 18 ans	$T_{31} \approx 22.725$	$T_{32} \approx 22.275$	45
Total	101	99	200

Tableau 7 : Tableau des effectifs théoriques sous l'hypothèse Ho (n=200)

Si nous comparons le tableau des effectifs théoriques sous l'hypothèse Ho avec le tableau de contingence, nous obtenons la description suivante en considérant qu'il y a tendance à l'attraction entre deux modalités quand l'effectif observé est strictement supérieur à l'effectif théorique et tendance à la répulsion entre deux modalités dans le cas contraire, strictement

inférieur. L'indépendance de deux modalités correspond à l'égalité de l'effectif théorique et de l'effectif contingent.

Affectivité Age	Affectivité positive	Affectivité négative	Total
9 – 11 ans	+ (attraction)	-(répulsion)	52
12 – 14 ans	+(attraction)	-(répulsion)	103
15 – 18 ans	-(répulsion)	+(attraction))	45
Total	101	99	200

Tableau 8 : Tableau des écarts entre contingence et effectifs théoriques sous l'hypothèse Ho

Le test du χ^2d'indépendance de Pearson repose sur la variable de décision D(χ^2) construite à partir d'une mesure de la distance, nommée distance du χ^2, entre le tableau de contingence et le tableau des effectifs théoriques sous l'hypothèse Ho d'indépendance statistique des deux variables.

$$D(\chi^2) = \sum_{i=1,j=1}^{i=3,j=2} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

Dans le contexte de l'article, cela donne :

$$D(\chi^2) = \frac{(33-26{,}26)^2}{26{,}26} + \frac{(54-52{,}015)^2}{52{,}015} + \frac{(14-22{,}725)^2}{22{,}725} + \frac{(19-25{,}74)^2}{25{,}74} + \frac{(49-59{,}985)^2}{59{,}975} + \frac{(31-22{,}275)^2}{22{,}275} \approx 10{,}415$$

Cette information figure dans la marge inférieure du tableau 5 de l'article sous la forme $\chi^2_{Cal[culé]} = 10{,}38$. Il s'agit de la réalisation de la variable de décision $D(\chi^2)$. L'écart entre 10.415 et 10,38 provient des approximations numériques utilisées. Ici cela n'a pas d'incidence mais il faut être vigilant sur la précision numérique mathématique qui peut conduire à des divergence de conclusion. Il est admis que la variable $D(\chi^2)$ est une variable aléatoire dont la distribution de probabilité suit approximativement celle de la variable de Pearson de type III appelée couramment variable du χ^2 (Khi-2) de degré de liberté ddl=(nombre de lignes-1)(nombre de colonnes-1)=(3-1)(2-1)=2. Le nombre de lignes est déterminé par le nombre de modalités de la variable V1 tandis que celui des colonnes, l'est par celui des modalités de la variable V2. L'information relative au degré de liberté figure dans la marge inférieur du tableau 5 de l'article sous la ddl=2.

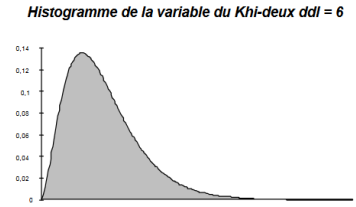
Dans la conception de Neyman-Pearson relative aux tests statistiques d'hypothèse, nous sommes dans la situation de confronter deux énoncés hypothétiques :

Ho = Les deux variables V1=âge et V2=climat affectif sont indépendantes

H_1= il existe un lien de nature statistique entre ces deux variables.

Il s'agit de retenir la plus plausible au sens des procédures de test statistique dont nous avons exposé le raisonnement dans les sections précédentes. Il nous faut alors choisir un niveau de risque du premier type usuellement noté α. Nous prenons α=0.01 en nous référant à l'information dans le tableau 5 formulée par P=.01. Il s'agit alors de trouver le fractile k_α d'ordre α de la variable du χ^2 avec ddl=2, appelé valeur critique. C'est la solution de l'équation : $Prob\{\chi^2_{ddl=2} > k_\alpha\} \leq \alpha = 0.01$. Pour résoudre cette équation, nous pouvons nous reporter à une table de valeurs critiques.

Nous pouvons lire ci-dessous $k_\alpha = 9.2104$



ddl	α = 0,01	α = 0,05	α = 0,10
1	6,6349	3,8415	2,7055
2	9,2104	5,9915	4,6052
3	11,3449	7,8147	6,2514
4	13,2767	9,4877	7,7794
5	15,0863	11,0705	9,2363
6	16,8119	12,5916	10,6446
7	18,4753	14,0671	12,0170

Tableau 9 : Extrait d'une table de valeurs critiques de la variable de Pearson de type III

Ou encore recourir aux fonctions programmées sous le logiciel Tableur Excel 2016.

Figure 26 : Fonction sous le tableur Excel

Dit autrement cela signifie que sous l'hypothèse Ho la variable de décision $D(\chi^2)$ est considérée comme une variable du χ^2 avec ddl=2 et par conséquent la zone critique au niveau de risque α=0.01, appelée encore zone de rejet de l'hypothèse Ho, est la demi-droite [9.21 ; +∞[Dit encore autrement cela signifie que la probabilité d'obtenir un tableau de contingence distant d'une valeur de plus 9.21 sous l'hypothèse Ho est moindre que 0.01. Au prix de ce niveau de risque, nous rejetons l'hypothèse d'indépendance entre les deux variables Age et Climat affectif et considérons qu'il est plus plausible d'accepter l'hypothèse H_1 d'existence d'un lien de nature statistique. C'est dans ce sens que figure l'information DS (différence significative) dans la marge du tableau 5 et que la conclusion est ainsi formulée : « *L'âge de l'enfant est lié au type de relation affective avec les parents* ». Toutefois il nous semblerait plus judicieux de considérer la dépendance du type de relations affectives avec les parents en fonction de l'âge V2=f(V1) que l'autre sens. Enfin il convient d'interpréter la signification de ce lien dans le cadre théorique de la psychologie.

5.3.1 Raisonnement statistique et coefficients d'association en tant que mesure d'intensité de liaison entre deux variables qualitatives

Une autre information est fournie dans la marge du tableau 5 de l'article. C=.22. Il s'agit du coefficient de Karl Pearson $C = \sqrt{\frac{\chi^2_{cal}}{n+\chi^2_{cal}}}$ qui donne bien la valeur indiquée $\sqrt{\frac{10,38}{200+10,38}} \approx 0,2222$. Notons qu'il existe d'autres mesures d'association dérivées de la mesure χ^2 . Ainsi nous trouvons le coefficient de Tschuprow : $0 \leq T = \sqrt{\frac{\chi^2}{n\sqrt{(p-1)(q-1)}}} \leq 1$ où p=nombre de lignes=nombre de modalités de la variable V1, q=nombre de colonnes=nombre de modalités de la variable V2 et n=effectif total de l'échantillon :

$$T = \sqrt{\frac{10.38}{200\sqrt{2}}} \cong 0,191569$$

Ou encore le coefficient de Cramer : $0 \leq V = \sqrt{\frac{\chi^2}{nInf\{p-1;q-1\}}} \leq 1$ où p=nombre de lignes=nombre de modalités de la variable V1, q=nombre de colonnes=nombre de modalités de la variable V2 et n=effectif total de l'échantillon

$$V = \sqrt{\frac{10.38}{200(1)}} \cong 0,2278157$$

Ces deux coefficients qui constituent une mesure de l'intensité de la liaison entre les variables, rendent possibles la comparaison entre deux tableaux de contingence qui n'ont pas les mêmes dimensions ni les mêmes effectifs totaux. Dit autrement, imaginons une seconde étude réalisée avec la variable V1 « âge » modélisée selon 4 tranches d'âge et la variable V2 « Affectivité » modélisée selon 3 modalités sur un échantillon de taille différente de la première étude, par exemple : 400. Imaginons qu'elle conduise au tableau de contingence suivant :

				Variable 2
Variable 1	*Affectivité positive*	*Affectivité neutre*	*Affectivité négative*	Totaux
9-11ans	58	16	30	104
12-14ans	92	32	82	206
15-16ans	14	8	22	44
17-18ans	14	8	24	46
Totaux	178	64	158	400

Tableau 10 : Tableau de contingence de la seconde étude (imaginaire)

Sous l'hypothèse Ho la variable de décision $D(\chi^2)$ est considérée comme une variable du χ^2 avec ddl=6 et par conséquent la zone critique au niveau de risque α=0.01, appelée encore zone de rejet de l'hypothèse Ho, est la demi-droite [16.81 ; +∞[. La valeur empirique est alors $\chi^2_{Cal[culé]} = 12,93$ qui ne se situe pas dans la zone de rejet. Nous conservons donc l'hypothèse

d'indépendance entre ces deux variables en prenant un risque de seconde espèce de niveau β inconnu toutefois. Si nous acceptons d'élever le niveau de risque de 1ère espèce en choisissant α=0.05, la zone de rejet devient alors [12.59 ; +∞[. Nous constatons que cette fois la valeur empirique de la variable de décision $\chi^2_{Cal[culé]} = 12{,}93$ se situe dans cette zone de rejet.

Ici nous voyons comment les options du chercheur ont une incidence sur les conclusions et les interprétations des résultats de son étude.

Nous ne pouvons cependant comparer cette étude à la précédente. Toutefois nous pourrions comparer le niveau d'intensité de la liaison par le coefficient de Tschuprow :

$$T = \sqrt{\frac{12.93}{400\sqrt{6}}} \cong 0{,}1149$$

Dans la conception de Fisher, cela reviendrait à calculer la p-value, c'est-à-dire la probabilité d'obtenir une valeur supérieure ou égale à $\chi^2_{Cal[culé]} = 10{,}38$. A l'aide de la fonction suivante :

Figure 27 : Fonction sous le tableur Excel

nous obtenons $Prob\{\chi^2_{ddl=2} > 10.38\} \leq 0.00558$. Cette valeur est bien moindre que P=0.01.

5.3.2 Raisonnement statistique et effet de la taille de l'échantillon dans le test du χ^2 de Karl Pearson

Il est important de prendre en considération l'effet de la taille de l'échantillon dans l'interprétation des résultats d'une étude. Nous allons montrer comment la variable de décision $D(\chi^2)$ est sensible à la taille de l'échantillon n. Supposons qu'une autre étude soit menée sur un échantillon de n=600, c'est-à-dire n= 3x200 et que nous obtenions les données suivantes :

Affectivité / Age	Affectivité positive	Affectivité négative	Total
9 – 11 ans	99	57	156
12 – 14 ans	162	147	309
15 – 18 ans	42	93	135
Total	303	297	600

Tableau 11 : Tableau de contingence n=600

Si nous examinons le tableau des profils-ligne, il ressort immédiatement que les fréquences conditionnelles sont identiques à celles qui figurent dans le tableau (Tableau 5) des profils-ligne de l'étude portant sur un échantillon de taille n=200.

			Variable 2
Variable 1	*Affectivité positive*	*Affectivité négative*	Totaux
9-11ans	0,634615385	0,365384615	1
12-14ans	0,524271845	0,475728155	1
15-18ans	0,311111111	0,688888889	1
Profil marginal	0,505	0,495	1

Tableau 12 : Tableau des profils-ligne n=600

Reprenant le raisonnement suivit dans le cas de l'étude portant sur un échantillon de taille n=200, nous pouvons en déduire les relations d'égalité suivantes relatives aux proportions :

$$\frac{T_{11}}{156}+\frac{T_{21}}{309}+\frac{T_{31}}{135}=\frac{T_{11}+T_{21}+T_{31}}{156+309+135}=\frac{303}{600}$$

Nous pouvons alors constater que les effectifs observés dans l'étude avec l'échantillon de taille n=200 étant le triple de ceux avec l'échantillon de taille n=600, les effectifs théoriques sont alors dans le même rapport. En conséquence nous obtenons un effet multiplicateur sur la réalisation de la variable de décision $D(\chi^2)$

$$\chi^2_{Cal[culé]}(n=600)=3\chi^2_{Cal[culé]}(n=200).$$

Tout se passe comme si ce test statistique d'hypothèse avait tendance à conduire au rejet de l'hypothèse nulle d'indépendance de deux variables qualitatives pourvu que la taille de l'échantillon soit suffisamment grande ! C'est une des critiques adressées à cette procédure.

5.3.3 Raisonnement statistique et annulation de l'effet de la taille de l'échantillon dans le test du χ^2 de Karl Pearson : coefficient φ^2

Pour annuler l'effet de la taille de l'échantillon et pouvoir ainsi comparer des études de liaison entre deux variables qualitatives conduites sur des échantillons de taille différente, le coefficient φ^2 a été proposé. Il est obtenu par la formule suivante :

$$\varphi^2=\frac{\chi^2}{n}$$

5.4 Raisonnement statistique et liaisons entre deux variables quantitatives ou entre une variable quantitative et une variable qualitative

Nous avons précédemment traité le cas où les deux variables sont qualitatives. Considérons d'abord le cas où les deux variables sont quantitatives puis celui où l'une est quantitative et l'autre est qualitative. Pour exposer les notions mises en jeu nous allons utiliser un exemple simulé. Pour cela nous rapportons le tableau statistique des résultats de quatre variables sur une population de taille n=25. V1, V2 et V3 sont des variables quantitatives discrètes et V4, une variable qualitative.

individus	V1	V2	V3	V4	individus	V1	V2	V3	V4
w1	15	0	0	C	w14	18	9	9	B
w2	15	0	9	C	w15	18	9	9	B
w3	15	0	9	C	w16	12	6	0	A
w4	18	9	9	A	w17	15	0	6	C
w5	18	9	0	B	w18	12	6	6	A
w6	18	9	0	B	w19	15	0	0	C
w7	15	0	9	C	w20	18	9	6	A
w8	18	9	0	B	w21	15	0	0	C
w9	15	0	0	C	w22	15	0	9	C
w10	12	6	0	A	w23	15	0	6	C
w11	12	6	9	A	w24	18	9	0	B
w12	18	9	6	B	w25	18	9	6	B
w13	12	6	9	B					

Tableau 13 : Tableau de séries statistiques (simulées) de variables quantitatives sur un échantillon (n=25)

De ce tableau ci-dessus des séries statistiques, nous pouvons extraire les informations utiles pour déterminer les distributions conjointes de fréquence des couples de variables (V1,V2), (V1,V3) et (V1,V4) et les distributions marginales de fréquence de chaque variable.

	V2			V3			V4			
V1	0	6	9	0	6	9	A	B	C	effectifs
12	0	5	0	2	1	2	4	1	0	5
15	10	0	0	4	2	4	0	0	10	10
18	0	0	10	4	2	4	2	8	0	10
effectifs	10	5	10	10	5	10	6	9	10	**25**

Tableau 14 : Tableau des distributions conjointes des effectifs de couples (V1,V2), (V1,V3) et (V1,V4)

De là nous calculons les paramètres usuels de V1, V2 et V3 :

variables	moyenne	variance	covariance
V1	m(V1) = 15,6	var(V1) = 5,04	
V2	m(V2) = 4,8	var(V2) = 16,56	cov(V1,V2) = 4,32
V3	m(V3) = 4,8	var(V3) = 16,56	cov(V1,V3) = 0

Tableau 15 : Tableau des moyennes, variances et covariances

Rappelons que la covariance entre deux variables est définie sous forme discursive comme suit :

La covariance Cov(X, Y) entre deux variables statistiques quantitatives X et Y est la moyenne (arithmétique) du produit des écarts à la moyenne de chacune des deux variables.

Et sous forme symbolique par la relation algébrique suivante :

$$Cov(X,Y) = moy[(X - \bar{X})(Y - \bar{Y})]$$

$$Cov(X,Y) = \overline{XY} - \bar{X}\,\bar{Y}$$

Nous pouvons aussi faire apparaître les fréquences conditionnelles par les profils lignes et les profils colonnes :

	V2			V3			V4			
V1	0	6	9	0	6	9	A	B	C	
12	0	1	0	0,4	0,2	0,4	0,8	0,2	0	1
15	1	0	0	0,4	0,2	0,4	0	0	1	1
18	0	0	1	0,4	0,2	0,4	0,2	0,8	0	1
profil moyen	0,4	0,2	0,4	0,4	0,2	0,4	0,24	0,36	0,4	**1**

Tableau 16 : Tableau des profils-lignes

	V2			V3			V4			
V1	0	6	9	0	6	9	A	B	C	profil moyen
12	0	1	0	0,2	0,2	0,2	2/3	1/9	0	0,2
15	1	0	0	0,4	0,4	0,4	0	0	1	0,4
18	0	0	1	0,4	0,4	0,4	1/3	8/9	0	0,4
	1	1	1	1	1	1	1	1	1	1

Tableau 17 : Tableau des profils-colonnes

Rappelons que l'indépendance statistique de deux variables X et Y est définie par la relation suivante : $Pr\{X\ sachant\ Y\} = \Pr\{X\}$ traduisant sous forme symbolique que la fréquence d'apparition d'un résultat quelconque de X ne dépend d'aucune condition relative à un résultat quelconque de Y. De là, nous pouvons déduire une caractéristique algébrique de l'indépendance statistique de X et de Y : $\Pr\{X\ et\ Y\} = \Pr\{X\} \Pr\{Y\}$

Par négation, s'il existe au moins un résultat de X pour lequel $Pr\{X\ sachant\ Y\} \neq \Pr\{X\}$, c'est dire, tel que $\Pr\{X\ et\ Y\} \neq \Pr\{X\} \Pr\{Y\}$, on dit alors que les deux variables sont dépendantes statistiquement. Naturellement ce lien est à analyser de très près car il n'est pas nécessairement causal, cependant si ce lien est explicitable il peut permettre de prévoir le résultat d'une variable connaissant l'autre. C'est le cas si nous pouvons trouver une relation fonctionnelle mathématique explicite à partir de laquelle des calculs sont possibles.

Si nous observons les données de notre exemple, nous pouvons remarquer que V1 et V3 sont deux variables statistiquement indépendantes alors que V1 et V2 ainsi V1 et V4 sont statistiquement dépendantes. L'indépendance de V1 et V3 justifie la nullité de la covariance cov(V1,V3). Mais attention comme nous le verrons dans la section 5.5, la covariance de deux variables peut être nulle, même si les deux variables sont dépendantes. En revanche la non-nullité de la covariance implique l'existence d'une dépendance statistique entre les deux variables statistiques.

5.4.1 Coefficient ρ de corrélation linéaire de Bravais-Pearson et coefficient empirique R_{BP}.

Lorsque nous avons affaire à un couple de variables quantitatives, une première approche de l'étude de la liaison consiste à construire une représentation graphique géométrique dans laquelle les points ont pour coordonnées les couples de résultats (x, y). La forme du **nuage**

statistique suggère des pistes pour donner un sens à une liaison possible entre X et Y. On définit le coefficient ρ de corrélation linéaire de Bravais-Pearson par la relation algébrique[3] suivante :

$$\rho = \frac{\text{cov (X,Y)}}{\sigma(X)\,\sigma(Y)} \quad \text{ou encore par le coefficient de détermination } \rho^2 = \frac{\text{cov}^2(X,Y)}{V(X)\,V(Y)} \quad .$$

On montre que $-1 \leq \rho \leq 1$. Ce coefficient est nul si les deux variables sont indépendantes. En revanche si la nullité de ρ exclut l'existence d'une **relation linéaire** entre X et Y, elle n'exclut pas l'existence d'autres relations et même des relations fonctionnelles comme nous le verrons dans la section 5.5.

Par ailleurs $\rho = \pm 1$ s'il y a une relation linéaire entre les deux variables X et Y. Cette propriété serait suggérée par la forme rectiligne du nuage.

Dans notre exemple $\rho(V1,V2) = \frac{4{,}32}{\sqrt{5{,}04}\sqrt{16{,}56}} \approx 0{,}4728$ et $\rho(V1,V3) = 0$

Le problème est que nous ne possédons la plupart du temps que des données d'échantillon. Il convient alors de définir une *statistique* permettant d'estimer la valeur ρ inconnue ou de prendre une décision dans un test statistique d'hypothèse relatif à ce coefficient de corrélation linéaire de Bravais-Pearson.

Considérons alors un n-échantillon (X_i , Y_i) du couple (X,Y). On définit la variable R_{BP} « coefficient de corrélation linéaire empirique de Bravais-Pearson », qu'on appelle aussi statistique comme nous l'avons vu dans la section 3, dont les réalisations sont calculées sur le n-échantillon, de la manière suivante :

$$R_{BP} = \frac{COV(X_n , Y_n)}{S(X_n)S(Y_n)}$$

qui est obtenu à partir de la variable « **moyenne empirique »** définie par

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \;\; et \; \overline{Y_n} = \frac{1}{n}\sum_{i=1}^{i=n} Y_i$$

et de la variable « **variance empirique sans biais »** (estimateur non biaisé) est définie par :

$$S(X_n)^2 = \frac{1}{n-1}\sum_{i=1}^{i=n} (X_i - \overline{X_n})^2 \;\; et \; S(Y_n)^2 = \frac{1}{n-1}\sum_{i=1}^{i=n} (Y_i - \overline{Y_n})^2$$

et de la variable **« covariance empirique »** est définie par :

$$COV(X_n, Y_n) = \frac{1}{n}\sum_{i=1}^{i=n} (X_i - \overline{X_n})(Y_i - \overline{Y_n})$$

[3] Ce coefficient peut être interprété géométrique comme le cosinus d'un angle dans un espace bien choisi, la variance comme une norme et la covariance comme un produit scalaire.

Nous précisons que ces statistiques ne sont pas autre chose que des **estimateurs** des paramètres inconnus des variables en jeu.

Lorsque le couple (X,Y) est un couple de variables binormal[4], les paramètres de cette statistique R_{BP} sont connus et admettent les valeurs approximatives :

$$E(R_{BP}) \approx \rho - \frac{\rho(1-\rho^2)}{2n} \qquad V(R_{BP}) \approx \frac{(1-\rho^2)^2}{n-1}\left(1+\frac{11\rho^2}{2n}\right)$$

$$\gamma_1 \approx \frac{-6\rho}{\sqrt{n}} \qquad \gamma_2 \approx \frac{6(12\rho^2-1)}{n}$$

5.4.2 Raisonnement statistique dans l'interprétation de la significativité du coefficient de corrélation de Bravais-Pearson

Avant toute étude plus approfondie, il convient de réaliser une représentation graphique du nuage statistique des n points de coordonnées (x_i, y_i). La forme de ce nuage est bonne façon d'orienter l'analyse même si l'interprétation peut être rendue plus difficile par la superposition de points sur le nuage statistique.

Supposons que les n observations proviennent d'une population dans laquelle les deux variables X et Y sont indépendantes. Dans ce cas, la valeur réelle du coefficient de corrélation de Bravais-Pearson est $\rho = 0$ .On peut alors utiliser la distribution de probabilité de la statistique R_{BP} correspondant à cet échantillonnage. Il est établi que si $\rho = 0$ et si le couple (X,Y) est un couple de variables de Laplace-Gauss, l'espérance de R_{BP} vaut $E(R_{BP}) = 0$ et la variance vaut $V(R_{BP}) = \frac{1}{1-n}$. Ajoutons maintenant que la distribution de probabilité de la variable transformée $T(R_{BP}) = \frac{R_{BP}\sqrt{n-2}}{\sqrt{1-R_{BP}^2}}$ est celle de la variable de Student de ddl =n-2.

Dans le cas général où ρ est quelconque dans [-1 ; +1], on peut utiliser la transformée Z de Fisher : $Z = \frac{1}{2}\ln\left(\frac{1+R_{BP}}{1-R_{BP}}\right) \xrightarrow[n\to\infty]{converge\ en\ loi} LG\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right); \frac{1}{\sqrt{n-3}}\right)$

Cette transformation permet traiter le cas général, y compris lorsque le couple (X,Y) n'est pas une variable de Laplace-Gauss de dimension 2 , dès que n est grand (n>30). Cependant le fait de ne pas rejeter l'hypothèse selon laquelle le coefficient de corrélation est nul, n'entraîne pas nécessairement l'indépendance des deux variables. Il n'y a là qu'une présomption

[4] variable de Laplace Gauss à deux dimensions dont la densité est :

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\ \mathbf{exp}\left\{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\left(\frac{y-\mu_2}{\sigma_2}\right)^2-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right)\right\}$$

d'indépendance. La nullité de ρ est une condition nécessaire mais pas suffisante pour l'indépendance. En d'autres mots, l'absence de corrélation linéaire n'implique pas l'indépendance.

5.4.3 Raisonnement statistique et rapports de corrélation

Examinons maintenant une mesure de liaison non symétrique, le rapport de corrélation $\eta^2_{Y|X}$ de Y en X, et le rapport de corrélation $\eta^2_{X|Y}$ de X en Y qui sont définis comme suit :

$$\eta^2_{Y|X} = \frac{V(m(Y|X))}{V(Y)} \qquad \eta^2_{X|Y} = \frac{V(m(X|Y))}{V(X)}$$

$$(0 \leq \eta^2_{Y|X} \leq 1) \qquad (0 \leq \eta^2_{X|Y} \leq 1)$$

Ce rapport $\eta^2_{Y|X}$ est maximal et égal à 1 si la variable Y est **fonctionnellement**[5] liée à la variable X. Il est à noter que ce coefficient est calculable que X soit une variable quantitative ou une variable qualitative. Dans ce dernier cas, nous ne pouvons calculer $\eta^2_{X|Y}$ **.** Ces deux coefficients ne sont pas symétriques.

La statistique $E^2_{Y|X}$« rapport de corrélation empirique » s'obtient de la façon suivante à partir des données issues d'un échantillon de taille n, en considérant que la variable X détermine k catégories respectivement de taille n_j avec j= 1 à k.

$$E^2_{Y|X} = \frac{\frac{1}{n}\sum_{j=1}^{j=k} n_j(\overline{Y_j} - \bar{Y})^2}{S(Y)^2} \; avec \; S(Y)^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}(Y_i - \overline{Y})^2$$

5.4.4 Caractère significatif du rapport de corrélation :

Sous Ho l'hypothèse de nullité de $\eta^2_{Y|X}$, on utilise la variable de décision : $D = \frac{\frac{E^2_{Y|X}}{k-1}}{\frac{1-E^2_{Y/X}}{n-k}}$

qui suit la distribution de probabilité de la variable de Fisher-Snedecor de ddl (k-1;n-k) si les distributions conditionnelles de Y pour chaque valeur ou modalité de X sont celles de la variable de Laplace-Gauss de même espérance μ et de même écart-type σ . Le nombre k correspond au nombre de valeurs ou de modalités de la variable X.

[5] au sens mathématique du terme qui indique l'existence d'un lien univoque de X vers Y: à un résultat x de X ne correspond qu'un et un seul résultat y de Y. Si la réciproque est vraie, nous avons affaire à un lien biunivoque qui caractérise une fonction bijective.

5.5 Raisonnement statistique et relation entre l'indépendance de deux variables et la covariance

Une question importante est à prendre en considération dans le raisonnement statistique :*Quelle information apporte la nullité de la covariance à l'égard de l'indépendance de deux variables quantitatives ? Quel intérêt peuvent avoir une représentation graphique et le rapport de corrélation* $\eta^2_{Y|X}$ *?*

5.5.1 Rappelons tout d'abord quelques définitions et propriétés.

Soient X et Y deux variables statistiques de moyennes m(X) et m(Y), de variances V(X) et V(Y), ces deux variables sont dites indépendantes statistiquement si la réalisation de n'importe quel résultat pour X n'influence d'aucune façon celle d'un résultat quelconque pour Y. Cela peut se traduire par : A étant un événement lié à X et B un événement lié à Y, la fréquence de (A et B) est égale au produit de la fréquence de (A) par la fréquence de (B) ou encore que la fréquence conditionnelle de (B sachant A) est égale à la fréquence de (A). Et la covariance Cov(X, Y) entre deux variables statistiques quantitatives X et Y est la moyenne (arithmétique) du produit des écarts à la moyenne de chacune des deux variables.

Propriété 1 : **Si les deux variables X et Y sont indépendantes statistiquement alors la covariance est nulle.**

Cette propriété résulte immédiatement du fait que dans ce cas la moyenne du produit XY est égale au produit des moyennes respectives de X et de Y. De là nous pouvons déduire logiquement le corollaire suivant :

Propriété 2 : **Si les deux variables X et Y ont une covariance non-nulle alors elles ne sont pas indépendantes statistiquement.**

Nous insistons une fois de plus qu'elles sont dépendantes **statistiquement** et que cette dépendance ne doit pas être étendue abusivement à la dépendance causale sans de multiples précautions.

5.5.2 Rappelons la prudence pour conclure et interpréter une étude appuyée par une démarche statistique

La covariance constitue un outil d'investigation en ce qui concerne le lien entre deux variables X et Y. Le problème est que si la non-nullité de la covariance implique l'existence d'un lien, sa nullité ne nous donne qu'une présomption d'indépendance. L'exemple que nous fournissons ci-dessous illustre la nécessaire prudence à maintenir dans les conclusions.

Considérons le tableau des séries statistiques ci-dessous.

individus	X	Y	individus	X	Y
CE01	4	16	CE21	10	100
CE02	9	81	CE22	8	64
CE03	2	4	CE23	0	272
CE04	2	4	CE24	2	4
CE05	6	36	CE25	8	64
CE06	0	272	CE26	1	1
CE07	2	4	CE27	6	36
CE08	2	4	CE28	6	36
CE09	8	64	CE29	4	16
CE10	8	64	CE30	8	64
CE11	7	49	CE31	8	64
CE12	5	25	CE32	8	64
CE13	2	4	CE33	4	16
CE14	8	64	CE34	2	4
CE15	2	4	CE35	2	4
CE16	2	4	CE36	10	100
CE17	10	100	CE37	3	9
CE18	0	272	CE38	8	64
CE19	5	25	CE39	1	1
CE20	8	64	CE40	9	81

Tableau 18 : Tableau des séries statistiques (simulées) des variables quantitatives discrètes X et Y

Il ressort que la moyenne de X vaut 5, celle de Y vaut 55,6 et la moyenne de la variable produit XY vaut 278. Si nous procédons alors au calcul de la covariance nous trouvons que $Cov(X,Y) = 278 - 5x55{,}6 = 0$. Et pourtant si nous observons la représentation graphique, nous pouvons être surpris par la régularité de la forme.

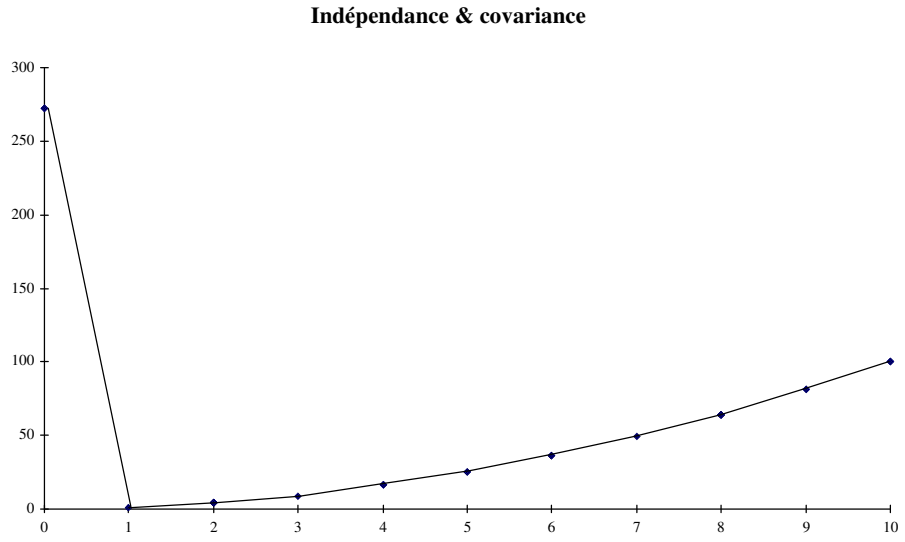


Figure 28 : nuage statistique du couple de variables statistiques (X, Y)

En fait si nous analysons, nous constatons l'existence d'une relation mathématique entre X et Y ainsi définie : 0 ---> 272 et $x \neq 0$ ---->$y = x^2$

Ainsi en face d'un tel résultat, il serait particulièrement erroné de conclure à une indépendance des deux variables bien qu'elles soient non-corrélées et que □ $(X,Y) = \frac{cov(X;Y)}{\sigma(X)\,\sigma(Y)} = 0$

Le raisonnement statistique suivant confirme le rejet de l'hypothèse d'indépendance entre les deux variables X et Y.

X / Y	0	1	2	3	4	5	6	7	8	9	10	effectifs
1	0	**2**	0	0	0	0	0	0	0	0	0	**2**
4	0	0	**10**	0	0	0	0	0	0	0	0	**10**
9	0	0	0	**1**	0	0	0	0	0	0	0	**1**
16	0	0	0	0	**3**	0	0	0	0	0	0	**3**
25	0	0	0	0	0	**2**	0	0	0	0	0	**2**
36	0	0	0	0	0	0	**3**	0	0	0	0	**3**
49	0	0	0	0	0	0	0	**1**	0	0	0	**1**
64	0	0	0	0	0	0	0	0	**10**	0	0	**10**
81	0	0	0	0	0	0	0	0	0	**2**	0	**2**
100	0	0	0	0	0	0	0	0	0	0	**3**	**3**
272	**3**	0	0	0	0	0	0	0	0	0	0	**3**
effectifs	**3**	**2**	**10**	**1**	**3**	**2**	**3**	**1**	**10**	**2**	**3**	**40**

Tableau 19 : Tableau de la distribution statistique conjointe de (X, Y)

Ainsi Prop{ X= 2 et Y = 4 }= $\frac{10}{40}$ = 0,025 alors que Prop{ X= 2 }= Prop{ Y = 4 }= $\frac{10}{40}$. Donc Prop{ X= 2 et Y = 4 }≠ Prop{ X= 2 }Prop{ Y = 4 }.

Nous pouvons même remarquer que la relation d'indépendance n'est vérifiée par aucun des résultats du couple (X, Y). En effet si nous recourons à la fréquence conditionnelle, nous remarquons :

Prop { Y = y sachant X = x} = 1 si x = 0 et y = 274 ou si x ≠ 0 et y = x^2 .

Prop { Y = y sachant X = x} = 0 si x = 0 et y ≠ 274 ou si x ≠ 0 et y ≠ x^2 .

de là Prop{ Y = y sachant X = x} ≠ Prop{ Y = y } pour toutes les valeurs de y de l'ensemble des résultats.

Pour analyser de plus près ce phénomène, nous pourrions recourir à l'autre outil que nous avons abordé dans la section 5.4.3. Il s'agit du **rapport de corrélation de Y en X**,

$\eta^2_{Y|X} = \frac{V(m(Y|X))}{V(Y)}$. Ce rapport est maximal et égal à 1 si la variable Y est **fonctionnellement**[6] liée à la variable X.

[6] au sens mathématique du terme qui indique l'existence d'un lien univoque de X vers Y: à un résultat x de X ne correspond qu'un et un seul résultat y de Y. Si la réciproque est vraie, nous avons affaire à un lien biunivoque qui caractérise une fonction bijective.

valeurs de X : k / moyennes conditionnelles de Y	0	1	2	3	4	5	6	7	8	9	10
m(Y | X=k)	272	1	4	9	16	25	36	49	64	81	100
fréquences	0,075	0,05	0,25	0,025	0,075	0,05	0,075	0,025	0,25	0,05	0,075

Tableau 20 : Tableau d'aide au calcul du rapport de corrélation de Y en X

De ce tableau, il ressort que la moyenne de Y vaut 55,6 et que la variance de Y qui vaut V(Y) = 4773,44, est égale à la variance V(m(Y|X)) des moyennes conditionnelles de Y. Si nous calculons maintenant le rapport de corrélation de Y en X, $\eta = \sqrt{\frac{V(m(Y|X))}{V(Y)}}$, nous obtenons la valeur 1 qui révèle l'existence de la liaison fonctionnelle de Y avec X lisible sur la représentation graphique.

Nous pourrions obtenir un résultat plus général en considérant **X comme une variable centrée et symétrique** car alors m(X) = 0 et le moment centré d'ordre 3, $m[(X-0)^3] = m[X^3] = 0$. De là si nous considérons la variable $Y = X^2$ nous obtenons alors

$$cov(X,Y) = m[XY] - m[X]\, m[Y] = m[X^3] - m[X]\, m[X^2] = 0.$$

Cependant la condition de symétrie n'est pas nécessaire comme l'illustre l'exemple que nous donnons ci-après.

5.5.3 Autre exemple de couple de variables dépendantes de covariance nulle

X / Y	-11	- 7	- 5	- 4	- 2	-1	0	4	5	9	effectifs
0	0	0	0	0	0	0	**4**	0	0	0	**4**
1	0	0	0	0	0	**4**	0	0	0	0	**4**
4	0	0	0	0	**8**	0	0	0	0	0	**8**
16	0	0	0	**4**	0	0	0	**4**	0	0	**8**
25	0	0	**2**	0	0	0	0	0	**6**	0	**8**
49	0	**2**	0	0	0	0	0	0	0	0	**2**
81	0	0	0	0	0	0	0	0	0	**4**	**4**
121	**2**	0	0	0	0	0	0	0	0	0	**2**
effectifs	**2**	**2**	**2**	**4**	**8**	**4**	**4**	**4**	**6**	**4**	**40**

Tableau 21 : Tableau de la distribution statistique conjointe de (X, Y)

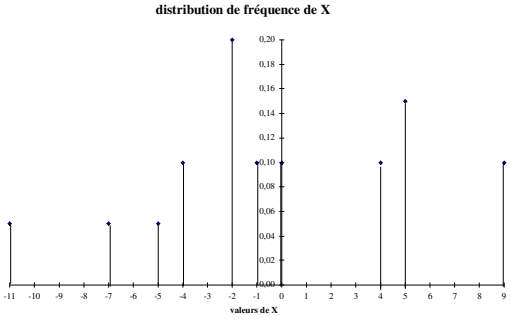


Figure 29 : Diagramme en bâtons de la variable statistique quantitative X

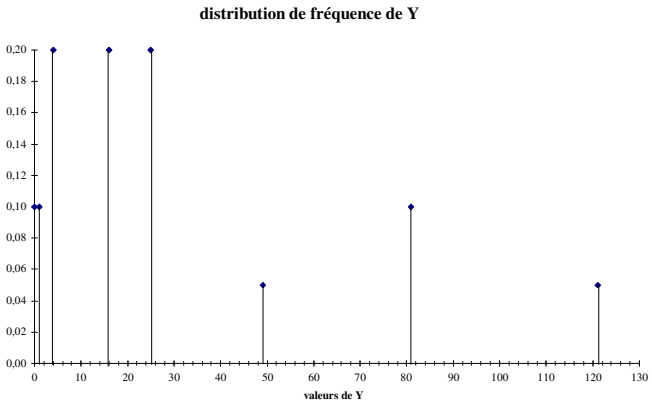


Figure 30 : Diagramme en bâtons de la variable statistique quantitative Y

Ci-après nous rappelons le tableau des séries statistiques de X et de Y.

individus	X	Y	individus	X	Y
I001	-11	121	I021	-1	1
I002	-11	121	I022	-1	1
I003	-7	49	I023	0	0
I004	-7	49	I024	0	0
I005	-5	25	I025	0	0
I006	-5	25	I026	0	0
I007	-4	16	I027	4	16
I008	-4	16	I028	4	16
I009	-4	16	I029	4	16
I010	-4	16	I030	4	16
I011	-2	4	I031	5	25
I012	-2	4	I032	5	25
I013	-2	4	I033	5	25
I014	-2	4	I034	5	25
I015	-2	4	I035	5	25
I016	-2	4	I036	5	25
I017	-2	4	I037	9	81
I018	-2	4	I038	9	81
I019	-1	1	I039	9	81
I020	-1	1	I040	9	81

Tableau 22 : Tableau des séries statistiques de X et de Y

Nous traduisons alors ce tableau sous la forme du nuage statistique des points de coordonnées (x, y). Il faut tenir dans l'interprétation du fait que certains points sont superposés, c'est-à-dire qu'à chaque point est attachée une masse qui n'est autre que l'effectif ou la fréquence.

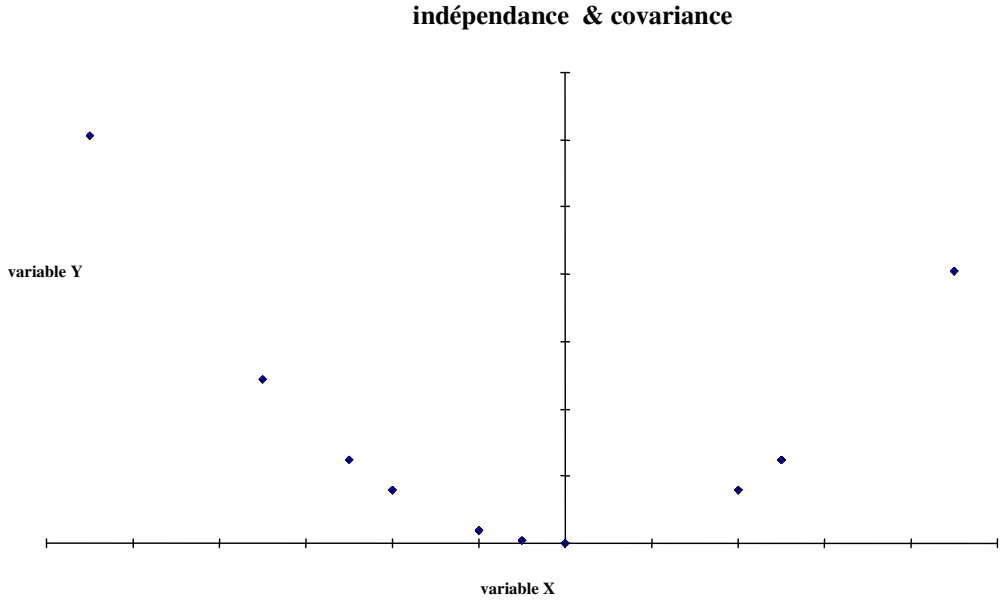


Figure 31 : nuage statistique du couple de variables statistiques (X, Y)

Ici m(X) = 0 mais aussi m[X^3] = 0. De là si nous considérons la variable Y = X^2 nous obtenons encore cov(X,Y) = m[XY] - m[X] m[Y] = m[X^3] - m[X] m[X^2] = 0.

Ainsi Prop{ X= 5 et Y = 25 }= $\frac{6}{40}$ alors que Prop{ X= 5 }= $\frac{6}{40}$ et Prop{ Y = 25 }= $\frac{8}{40}$. Donc Prop{ X= 5 et Y = 25 }≠ Prop{ X= 5 }Prop{ Y = 25}. Pour que ces deux événements soient indépendants, il aurait fallu obtenir Prop{ X= 5 et Y = 25 }= 0,03. Quoi qu'il en soit ceci conduit à rejeter l'hypothèse d'indépendance de X et de Y pourtant suggérée par la nullité de la covariance.

Tout cela confirme la nécessité de grande prudence quant à l'information issue de la nullité de la covariance ou du coefficient de corrélation linéaire qui en découle pour porter un jugement à l'égard de l'indépendance de deux variables quantitatives. Qui plus est, quand la nullité de la covariance est elle-même le résultat d'un test statistique conduisant à ne pas rejeter l'hypothèse Ho au niveau α , c'est à dire à conserver Ho avec un risque de second espèce de niveau β, nous pouvons imaginer combien la conservation de l'hypothèse d'indépendance est risquée. Toutefois en utilisant une représentation graphique pour repérer les points de coordonnées (x;y) et le rapport de corrélation $\eta^2_{Y|X}$, nous pouvons confronter la vraisemblance de la présomption d'indépendance de X et de Y à une forme remarquable du nuage de points. À cela s'ajoute d'autres éclairages que le contexte dans lequel le problème de recherche de liaison entre X et Y a été posé de manière pertinente, doit pouvoir apporter.

valeurs de X : k / moyennes conditionnelles de Y	-11	- 7	- 5	- 4	- 2	-1	0	4	5	9
m(Y | X=k)	121	49	25	16	4	1	0	16	25	81
fréquences	0,05	0,05	0,05	0,1	0,2	0,1	0,1	0,1	0,15	0,1

valeurs de Y : k / moyennes conditionnelles de X	0	1	4	16	25	49	81	121
m(X | Y=k)	0	-1	-2	0	2,5	-7	9	-11
fréquences	0,1	0,1	0,2	0,2	0,2	0,05	0,1	0,05

Tableau 23 : Tableau d'aide au calcul du rapport de corrélation de Y en X

Les calculs des divers paramètres utiles nous fournissent les résultats suivants :

$$m(X) = m[m(X|Y)] = 0 \qquad m(Y) = m[m(Y|X)] = 25,7$$

$$V(X) = 25,7 \qquad V(Y) = 1027,21$$

$$V[m(X|Y)] = 18,75 \qquad V[m(Y|X)] = 1027,21$$

D'où d'une part la covariance cov(X,Y) = 0 et par conséquent le coefficient de corrélation (linéaire) $\rho = 0$, d'autre part les deux rapports de corrélation prennent respectivement les valeurs suivantes :

rapport de corrélation de Y en X, $\eta^2_{Y|X} = \frac{V(m(Y|X))}{V(Y)} .= \frac{1027,21}{1027,21} = 1$

rapport de corrélation de X en Y, $\eta^2_{X|Y} = \frac{V(m(X|Y))}{V(X)} .= \frac{18,75}{25,7} \approx 0,729$

Ainsi la nullité du coefficient de corrélation (linéaire) traduit l'absence de liaison linéaire et non l'indépendance des deux variables X et Y. Par ailleurs la valeur du rapport de corrélation de Y en X, $\eta^2_{Y|X} = 1$, confirme l'existence d'une liaison fonctionnelle, ici $Y = X^2$, alors que celle du rapport de corrélation de X en Y, $\eta^2_{X|Y} \approx 0,729$, rend compte d'une liaison non-fonctionnelle.

Cet exemple confirme toute l'attention que nous devons porter à la formulation des conclusions relatives aux recherches de liaisons, quand le raisonnement statistique prend appui sur ces outils. L'étude théorique de ces outils mathématiques paraît être un bon moyen pour mieux maîtriser ces formulations.

5.6 Réflexion sur certains obstacles lors de la mise en œuvre de démarches statistiques

La lecture-compréhension de représentations symboliques sous la forme de formules algébriques et l'expertise calculatoire que nécessite leur mise en œuvre par les opérations arithmétiques pour résoudre des problèmes au niveau du cadre théorique des mathématiques constituent des facteurs générateurs d'obstacles de diverse nature à l'engagement dans des approches statistiques.

Prenons, dans un premier temps, le cas du calcul de l'écart-type d'une variable quantitative. Nous partons de la définition rapportée à son algorithme de calcul formulé sous une forme discursive dans le registre de la connaissance prédicative.

Définition algorithmique *: L'écart-type est la racine carrée de la moyenne arithmétique des carrés des écarts (des valeurs) à la moyenne arithmétique de ces valeurs.*

Sous la forme d'une représentation sémiotique du registre symbolique algébrique, nous avons la formule :

$$\sigma(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{i=k} n_i(x_i - \bar{X})^2}$$

La lecture tant de la forme discursive que de la forme symbolique s'effectue de gauche à droite.

Maintenant examinons ce concept dans le registre de la connaissance opératoire. Analysons finement le déroulement de la réalisation des calculs :

Étape	**Description de l'opération**	**Opération représentée symboliquement**	**Opérations arithmétiques**
1	Calcul de la moyenne arithmétique $\bar{X}$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{i=k} n_i\, x_i$	Addition Multiplication Division
2	Calcul des k écarts e_i à la moyenne	Pour i= 1 à k, $e_i = x_i - \bar{X}$	Soustraction
3	Calcul des carrés de écarts e_i	Pour i= 1 à k, $e_i^2 = (x_i - \bar{X})^2$	Multiplication
4	Calcul de la moyenne arithmétique des carrés de écarts e_i^2 qui est la variance	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{i=k} n_i e_i^2$	Addition Multiplication Division
5	Calcul de la racine carrée de la variance	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{i=k} n_i e_i^2}$	Racine carrée

Tableau 24 : Déroulement pas à pas de l'algorithme de calcul de l'écart-type

Nous pouvons alors constater que les opérations élémentaires mobilisées font partie des connaissances acquises et compétences développées au cours de la formation scolaire de base. Mais quand nous analysons les données construites au vu des productions écrites d'étudiants de licence ou master en sciences de l'éducation et de la formation, nous constatons de nombreuses difficultés à réaliser cette suite d'opérations à partir des erreurs commises. Il ressort une fois de plus que le tout que traduit la formule, est bien plus que la somme des opérations arithmétiques élémentaires qui le constituent.

Une autre remarque que nous faisons est l'inversion de l'ordre des opérations par rapport à l'ordre de lecture. Il est possible que cela conduise les utilisateurs à mobiliser des opérations mentales d'un niveau de complexité supérieur.

Dans un second temps, nous allons porter notre attention sur le cas de l'intervalle de confiance à un niveau .95 d'une proportion dans lequel sont mobilisées les mêmes opérations algébriques. L'extrait d'une production écrite met en évidence l'obstacle calculatoire auquel est confronté un étudiant lors de la résolution du problème suivant construit à partir d'une situation-problème issue des travaux d'une thèse en sciences de l'éducation de Jean-Bruno Bernard (2007) sur la thématique de *l'enfant "aidé" à l'école : de la position de difficulté d'apprentissage passagère à la situation d'échec scolaire avéré. L'influence des représentations sociales des maîtres généralistes et des maîtres E* soutenue en 2007.

Considérez-vous qu'il existe un profil type d'enfants prédisposés à être en difficulté d'apprentissage ?

Tableaux statistiques de la variable V 07

Tableau n° T 7	V07 = profil -type	oui	non	Je ne sais pas
« Maîtres généralistes »	Effectifs	108	160	30
« Maîtres E »	Effectifs	52	57	14

Q 401 Donner la proportion des individus déclarant « oui, il existe un profil type… » dans **l'échantillon** des « maîtres généralistes ».

Q 402 Donner une **estimation ponctuelle** de la proportion des individus déclarant « oui, il existe un profil type… » dans la **population** des « maîtres généralistes »

Q 403 Donner une **estimation par intervalle de confiance à 95%** de la proportion des individus « oui, il existe un profil type… » dans la population des « maîtres généralistes »

Figure 32 : Énoncé d'une situation-problème mobilisant l'estimation statistique

Pour plus de précision, nous rappelons que, dans ce cas, la proportion des individus sur l'échantillon est donnée par le rapport $f_n = p = \frac{108}{298} \approx 36.24$. Cette proportion est considérée comme une estimation ponctuelle de la proportion inconnue π sur la population d'étude. Pour obtenir une estimation par intervalle de confiance, nous avons recours à la formule suivante :

$$f_n - k\sqrt{\frac{f_n\,(1-f_n)}{n-1}} < \pi < f_n + k\sqrt{\frac{f_n\,(1-f_n)}{n-1}}$$

Nous pouvons constater que cette formule est bien identifiée et rappelée dans le texte ci-contre. Les valeurs de réalisation de la proportion empirique sont correctes :

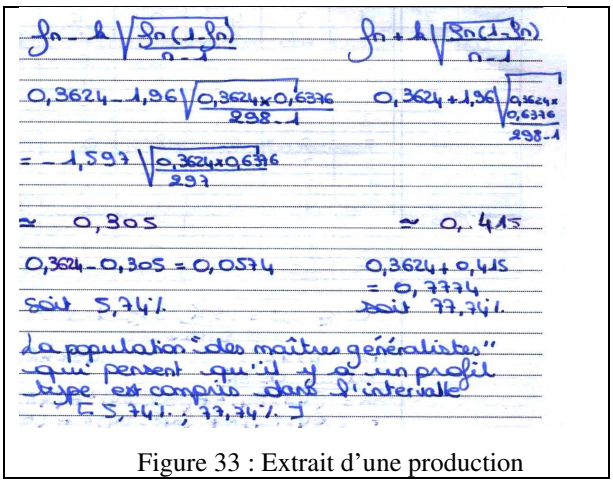


Figure 33 : Extrait d'une production

$$0{,}3624 \pm 1{,}96\sqrt{\frac{0{,}3624(1-0{,}3624)}{298-1}}$$

La valeur k=1.96 est déterminée par le niveau de confiance de 0.95 à partir de la loi de Laplace-Gauss de paramètres (0 ; 1) en résolvant l'équation :

$$Prob\{-k < LG(0;1) < k\} = 0.95$$

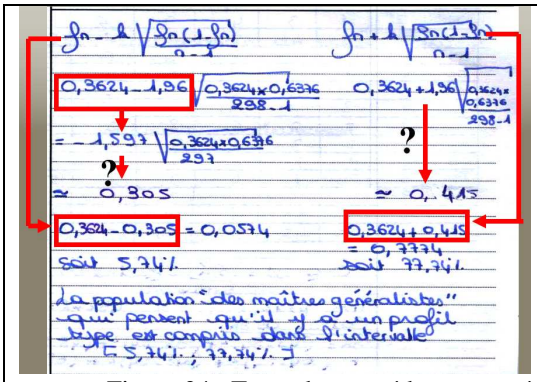


Figure 34 : Trace des procédures erronées

C'est alors au niveau de la réalisation même des calculs que les erreurs se produisent comme nous le faisons apparaître dans la figure ci-contre. Nous pouvons repérer la première erreur commise à partir de la non-prise en compte de la règle dite de la priorité de la multiplication sur l'addition dans les notations simplifiées.

6 Conclusion

Nous avons tenté ici de proposer une réflexion sur l'usage du cadre théorique de la statistique dans les travaux de recherche dans le domaine de l'éducation, en particulier dans celui des sciences de l'éducation et de la formation en tant que champ disciplinaire universitaire. Ce texte reprend en partie les propos exposés lors de la conférence et les complète. Il s'appuie aussi sur une expérience déjà longue d'enseignement de la statistique pour des non-spécialistes et d'encadrement de travaux de recherche pour lesquels l'approche statistique était tout à fait pertinente. Une réflexion devrait aussi être menée sur la question de l'erreur et de sa signification dans le raisonnement statistique. Des auteurs ont consacré leurs réflexions sur la question générale de l'erreur chez les êtres humains comme, par exemple au XIXème, la thèse de Victor Brochard soutenue en 1879 à la faculté des lettres de Paris sous le titre : de l'erreur (Brochard, 1927). Un article de Jean-Pierre Algoud (1982) porte sur une question qui reste, 30 ans plus tard, tout à fait actuelle, celle de la place de l'erreur dans le cadre de l'intelligence artificielle. Cela concerne aussi le domaine du raisonnement statistique dans la mesure où ce qui fut dénommé maladroitement *intelligence artificielle* n'a rien à voir avec l'intelligence humaine malgré ses apparences. Les algorithmes sont des créations humaines et ne sont en aucune façon en mesure de comprendre les contenus et instructions à la manière des êtres humains. Ils ne font qu'exécuter des ordres paramétrés par des langages construits par les êtres humains. Ils sont donc d'abord au service de leurs créateurs qui les paramètrent. L'opérationnalité de ces algorithmes emprunte largement aux approches statistiques et en retour ceux-ci fournissent des outils d'aide aux traitements statistiques. Enfin Vittorio Girotto et Michel Gonzalez (2000) abordent une question qui concerne tout à fait le raisonnement statistique à savoir celle des erreurs dans le raisonnement probabiliste quotidien. Ils partent de la question suivante : « *Des personnes qui ne maîtrisent pas le calcul de probabilités sont-elles*

capables de faire des inférences probabilistes correctes ? (…) Par exemple, une personne qui doit décider d'une façon raisonnée s'il est préférable d'investir son épargne en actions ou en obligations doit évaluer la probabilité de diverses éventualités telles qu'une hausse des taux directeurs, surchauffe de l'économie, et une reprise de l'inflation. » (*ibid.*, 2000, p.133) Pour eux, « *les compétences probabilistes sont limitées par des contraintes cognitives qui conduisent à certaines erreurs de jugement. Mais aussi (…) il existe une réelle compétence probabiliste qui permet des évaluations de probabilité correctes dans des situations où l'information disponible n'est pas celle de notre environnement naturel.* » (*ibid.*, 2000, p.134) Ce questionnement concerne à l'évidence le rôle de l'approche statistique et la mise en œuvre du raisonnement statistique dans la mesure où l'inférence statistique mobilise des outils du domaine des théories des probabilités. L'usage même des termes plausibilité, vraisemblance renvoie au langage des probabilités et au raisonnement probabiliste.

7 Références

Parmi ces références figurent des articles qui portent sur une réflexion relative des questions relevant du domaine de la statistique dont les textes sont accessibles à partir de la base de données HAL

ACADÉMIE DES SCIENCES (2000) *La statistique* Rapport sur la science et la technologie n°8, Paris : Éditions TEC&DOC

Algoud, J.P., (1982) Erreur et intelligence artificielle. J. Oudot, A. Morgon, J.-P Revillard (Eds) *L'erreur*. pp.139-156 Lyon : PUL

Bernard, J.B., (2007) *L'enfant "aidé" à l'école : de la position de difficulté d'apprentissage passagère à la situation d'échec scolaire avéré. L'influence des représentations sociales des maîtres généralistes et des maîtres E* Thèse de doctorat de l'Université Lumière Lyon2 [http://theses.univ-lyon2.fr/documents/lyon2/2007/bernard_jb/]

Brochard, V., (1879, 1897, 1927) *De l'erreur*. Paris : Felix Alcan

Brousseau, G., (2004) *Situations fondamentales et processus génétique de la statistique.* XII École d'été de didactique des mathématiques.

Chevallard, Y. (1978) *Notes pour la didactique de la statistique* I.R.E.M. d'Aix-Marseille.

Cournot, A. A., (1843, 1984) *A.A. Cournot, œuvres complètes.* Tome 1 : *exposition de la théorie des chances et des probabilités*, B. Bru (Ed.), Paris : Librairie J. Vrin, 385 p.,

Cournot, A. A., (1851) *Essai sur les fondements de nos connaissances les caractères de critique philosophique*. Paris : Hachette

Eduscol (octobre 2016) *Esprit critique* [https://eduscol.education.fr/1538/former-l-esprit-critique-des-eleves]

Escoffier, B., Pagès, J. (1990) *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation*, Paris : Dunod.

Girard, JC, Régnier, JC, (1998d) Pourquoi "faire des statistiques» ? J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment ?*, Irem de Lyon - Université Lyon1, pp.1-4. ⟨halshs-00405959⟩

Girotto, V., Gonzalez, M., (2000) Les erreurs dans le raisonnement probabiliste quotidien. L'erreur. *Revue Le temps des savoirs*. (2) Paris : Odile Jacob. pp.133-146

Guénot, M., Régnier, JC, (2010) L'histoire de l'art à l'épreuve de l'analyse statistique implicative : l'exemple de la structure iconique de l'image médiévale. ASI 5 *Proceedings, Quaderni di Ricerca in Didattica*, Université de Palerme, Nov 2010, Palerme, Italy. ⟨ halshs-01501347 ⟩

Huberman, A. M., Miles, M. B., (1991) *Analyse des données qualitatives. Recueil de nouvelles méthodes.* (Traduit de l'anglais par C. De Backer, V. Lamongie) Bruxelles De Boeck Université pp. 21-24

Jenny Jacques (2004) *«Quanti / Quali» = distinction artificielle, fallacieuse et stérile !* [http://jacquesjenny.com/legs-sociologique/?page_id=1159]

John W. Creswell, Vicki L. Plano Clark (2017) *Designing and Conducting Mixed Methods Research* SAGE Publications

Lalande, A., (1926, 1991) *Vocabulaire technique et critique de la philosophie* Paris : PUF

Lauren H. Bryant The Structure of Mixed Method Studies in Educational Research: A Content Analysis *Journal of Research in Education* – 2011, Volume 22, Number 1

Lecoutre, JP, Tassi, Ph. (1987) *Statistique non paramétrique et robustesse* Paris : Economica

Mazet, S. (2016) *Esprit critique* https://www.dailymotion.com/eduscol

Miles, M.B. (1979) Qualitative data as an attractive nuisance: The problem of analysis. *Administrative Science Quarterly*, 24, 590-601

Régnier, JC, (1998a) Lire un article de journal de la presse ordinaire. J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment ?* IREM de Lyon Université Lyon1, pp.127-133. [https://halshs.archives-ouvertes.fr/halshs-00406105/]

Régnier, JC, (1998b) La prise de décision risquée en situation incertaine : élément pour une séquence didactique visant l'acquisition du raisonnement statistique. J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment ?* IREM de Lyon Université Lyon1, pp.189-201. [https://halshs.archives-ouvertes.fr/halshs-00406126/]

Régnier, JC, (1998c) Finalités et enjeux de l'enseignement de la statistique. J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment ?*, IREM de Lyon Université Lyon1, pp.5-20. ⟨halshs-00405986⟩

Régnier, JC, (1998e) Lire un article de journal de la presse ordinaire. J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment ?*, IREM de Lyon Université Lyon1, pp.127-133. ⟨halshs-00406105⟩

Régnier, JC, (1998f) De la vérité autoproclamée à la vraisemblance reconnue. J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment ?*, IREM de Lyon Université Lyon1, pp.107-118. ⟨halshs-00406007⟩

Régnier, JC, (1998g) Histogramme. J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment ?* Irem de Lyon Université Lyon1, pp.21-42. ⟨halshs-00405993⟩

Régnier, JC, (1998h) Danger! Approximations.... J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment?* IREM de Lyon Université Lyon1, pp.99-105, 1998. ⟨halshs-00405999⟩

Régnier, JC, (1998i) La prise de décision risquée en situation incertaine : élément pour une séquence didactique visant l'acquisition du raisonnement statistique. J.C. Girard et al. (Eds) *Enseigner la Statistique du CM à la Seconde Pourquoi? Comment ?*, IREM de Lyon Université Lyon1, pp.189-201, ⟨halshs-00406126⟩

Régnier, JC, (2000) *Auto-évaluation et autocorrection dans l'enseignement des mathématiques et de la statistique Entre praxéologie et épistémologie scolaire*. Note de synthèse HDR. Université Marc Bloch - Strasbourg II. [https://tel.archives-ouvertes.fr/tel-00361408]

Régnier, JC, (2002) A propos de la formation en statistique. Approches praxéologiques et épistémologiques de questions du champ de la didactique de la statistique. *Revue du Centre de Recherche en Éducation*, 22/23, pp.157-201. ⟨halshs-00363427⟩

Régnier, JC, (2005a) Formation de l'esprit statistique et raisonnement statistique. Que peut-on attendre de la didactique de la statistique ? *Séminaire National de Didactique des Mathématiques*. Paris : France. pp.13-38. ⟨halshs-00391741⟩

Régnier, JC, (2005b) Étude des difficultés d'apprentissage de la statistique dans le cadre d'un enseignement à distance. Jean-Pierre Gaté; Noëlle Zendrera; Alain Bihan-Poudec; Christelle Chevallier-Gaté. *Mesurer. Actes du Symposium "Pédagogie de la statistique à l'Université"*, L'Harmattan, pp.15-47, Éduquer. ⟨halshs-00361957⟩

Régnier, JC, (2014) Instrumentalisation technocratique des statistiques et alternatives citoyennes. Martine Boudet, Florence Saint-Luc. *Le système éducatif à l'heure de la société de la connaissance*, Presses Universitaires du Mirail, pp.83-106. Questions d'éducation. ⟨hal-01094725⟩

Régnier, JC, Koroleva, B., D., (2019) Statistique, Langage, Culture - Réflexions sur des questions langagières et culturelles dans l'apprentissage, l'enseignement et les usages de la statistique. *XXX Ежегодная международная научная конференция «ЯЗЫК и КУЛЬТУРА»*, Faculté des Langues étrangères - National Research Tomsk State University, Sep 2019, Tomsk, Russie. ⟨hal-02292032⟩

Saporta, G., (1990) *Probabilités, Analyse des données et Statistique*. Paris : Éditions Technip.

Vergnaud, G., (1991) La théorie des champs conceptuels, *Recherches en Didactique des mathématiques*, 10/2.3, pp 133-169