Guy Tchibozo

Introduction pratique aux méthodes quantitatives en Sciences de l'éducation et de la formation

Atramenta

Introduction pratique aux méthodes quantitatives en Sciences de l'Éducation et de la Formation

Guy Tchibozo

Publié en juillet 2019 par :

Atramenta
Tampere, FINLANDE

www.atramenta.net

Imprimé en France
Imprimeur certifié Imprim'Vert

ISBN 978-952-340-501-1

Avant-propos

Le recours aux approches quantitatives est moins fréquent en recherche académique en sciences de l'éducation et de la formation qu'il ne l'est dans d'autres sciences sociales et humaines. Un corollaire en est la relative rareté des outils pédagogiques susceptibles de permettre à un large public d'étudiants, enseignants, praticiens et chercheurs en éducation et formation de s'initier rapidement et efficacement aux méthodes et outils quantitatifs. Ce qui, en retour, ne saurait favoriser ni la familiarité avec ces outils, ni le recours à ces approches. Or s'il est évident qu'il n'y a pas à tout aborder sous l'angle quantitatif en éducation et formation, il est en revanche certain que des analyses quantitatives sont parfois indispensables.

Cet ouvrage propose une introduction aux principales méthodes quantitatives contemporaines d'analyse et de recherche en éducation et formation. L'objectif en est de permettre à des éducationnistes non-familiers de ces méthodes de s'y initier aisément afin d'être capables, lorsque nécessaire, d'y avoir recours de façon pertinente dans le cadre de leurs travaux d'étude et de recherche. L'ambition est de les guider et de les soutenir dans la mise en œuvre de ces démarches, méthodes et outils. La priorité de l'ouvrage est donc de présenter de façon simple et pratique une large gamme de méthodes

existantes ; d'expliquer, pour chacune d'entre elles, ce à quoi elle sert et à quel type de question d'étude ou de recherche en éducation et formation elle peut permettre de répondre ; d'en montrer la logique d'ensemble ; d'illustrer comment, en termes pratiques, la mettre en œuvre ; et de préciser comment en interpréter les résultats.

L'ouvrage s'adresse tout particulièrement aux étudiants en Licence, Master et Doctorat en sciences de l'éducation et de la formation. Commencer à s'initier dès la première année de Licence à l'approche quantitative, puis acquérir progressivement de nouvelles méthodes au fil du cursus, donne à l'étudiant une plus grande latitude quant au choix des approches, et les meilleures chances de concevoir clairement, maitriser rapidement et mettre en œuvre efficacement des procédures quantitatives s'il y a lieu, lorsque sera venu le moment d'organiser sa recherche de mémoire ou de thèse, puis d'en entreprendre la phase d'investigation empirique.

L'objectif de cet ouvrage n'est clairement pas de former des statisticiens ou des chercheurs en statistique. Priorité est à la mise en œuvre concrète de méthodes par des utilisateurs qui n'en sont pas familiers. Il ne sera donc question ici ni de théories statistiques ni de leurs fondements mathématiques. Il s'agit essentiellement de poser les principaux repères grâce auxquels un

utilisateur pourra, de façon informée, sélectionner la méthode pertinente pour son travail ; en argumenter le choix ; en engager la mise en œuvre – le cas échéant avec appel éclairé à des personnes-ressources, et en interaction intelligente avec elles – ; et en tirer des conclusions utiles à ses analyses.

Pour que l'ensemble de ce pari puisse être tenu, il importait que l'ouvrage soit accessible. Le texte est donc présenté sans formalisation inutile ni développement mathématique superflu. L'essentiel est ainsi mis à disposition sous forme compacte, sachant que le lecteur qui souhaitera en savoir plus et aller plus loin pourra toujours se référer à la riche littérature qui existe sur ces questions dans différentes branches de l'analyse quantitative : statistique, biostatistique, économétrie, psychométrie, sociométrie, édumétrie, notamment.

Précisons à ce stade que les méthodes présentées dans cet ouvrage débordent du cadre de l'édumétrie. Telle qu'entendue traditionnellement, l'édumétrie se focalise sur la mesure dans les processus d'apprentissage. Les méthodes présentées ici peuvent – au moins pour certaines d'entre elles – s'appliquer à l'analyse des processus d'apprentissage. Mais elles ont aussi vocation à s'appliquer en dehors, dans l'analyse des *contextes* mêmes au sein desquels l'apprentissage prend place, qu'il s'agisse de contextes institutionnels,

sociaux ou économiques, ainsi que dans la perspective plus large de l'analyse des politiques d'éducation et de formation.

Cet ouvrage s'adresse à un large public d'enseignants-chercheurs, chercheurs, chargés d'études, doctorants et étudiants dans le domaine de l'éducation et de la formation au sens large. Il est conçu comme un outil à disposition de ceux qui souhaiteraient accompagner l'apprentissage et/ou étendre leur palette de compétences dans le champ du quantitatif. Il n'a pas pour objet d'imposer la perspective quantitative là où elle ne se justifierait pas. Je forme le vœu qu'il réponde aux besoins de son public.

Introduction

Cet ouvrage est un compagnon de route de l'éducationniste dans l'exploration, la maîtrise progressive et l'exploitation du potentiel des méthodes quantitatives pour l'analyse et la recherche en éducation et formation. Il est conçu de façon à accompagner l'utilisateur, des premiers stades de l'apprentissage jusqu'à une relative autonomie. À cette fin, l'étendue du contenu présenté est des plus larges, allant des notions de base de statistique descriptive jusqu'à des méthodes avancées comme la modélisation par équations structurelles. L'objectif est de permettre un apprentissage progressif et organisé de ces méthodes. Chaque chapitre fait appel à des connaissances présentées dans les chapitres précédents. Le lecteur pourra ainsi mesurer lui-même, chapitre après chapitre, le chemin parcouru. Mais bien entendu, l'ouvrage permet aussi au lecteur qui dispose déjà des bases, d'aller directement au chapitre traitant de la méthode spécifique à laquelle il envisage d'avoir recours.

Cet ouvrage s'inscrit dans une logique de construction de compétences, suivant une approche pratique et concrète, à finalité opérationnelle. Ce n'est ni une encyclopédie, ni un traité théorique, ni un recueil de savoirs abstraits dont l'intérêt pratique resterait un mystère. C'est essentiellement un guide de méthode pour, étape par étape, se doter de compétences en analyse quantitative et être capable de mettre en œuvre

ces compétences dans le cadre de travaux d'étude et de recherche en éducation et formation.

Contenu et organisation de l'ouvrage

L'ouvrage est divisé en trois parties, précédées par un chapitre liminaire introduisant les notions de base.

Le chapitre liminaire (Chapitre 1) présente et illustre des notions telles que celles d'individu, population, variable, indice, série statistique, distribution statistique, courbe normale, ainsi que les indicateurs basiques de statistique descriptive. L'accent est mis dans ce chapitre sur des outils qui reviennent fréquemment par la suite dans l'ouvrage, et plus généralement en analyse quantitative, et dont la bonne compréhension est indispensable à une progression aisée.

Suivent trois parties consacrées respectivement aux principes de collecte et de préparation de données, aux méthodes d'analyse de données, et à la modélisation des relations entre variables.

La première partie (chapitres 2 et 3) présente les principales étapes préalables à une démarche quantitative. Le chapitre 2 est consacré aux méthodes de collecte des données, mettant l'accent sur les aspects cruciaux mais rarement présentés que constituent la vérification de la validité et de la lisibilité de questionnaires, ainsi que l'échantillonnage. Le chapitre 3 rappelle les règles générales de la préparation des

données. L'attention y est attirée sur la question de la gestion des valeurs aberrantes. Les principales méthodes de standardisation des données y sont exposées.

La deuxième partie de l'ouvrage (chapitres 4 à 15) présente une gamme étendue de méthodes d'analyse de données. Les chapitres 4 et 5 introduisent respectivement l'analyse lexicométrique et l'analyse classificatoire, rarement présentées mais dont le potentiel est non négligeable pour les chercheurs en sciences de l'éducation et de la formation. Les chapitres 6 à 12 présentent la philosophie, les méthodes et les conditions d'utilisation des principaux tests statistiques, sans aucun doute les outils les plus rapidement accessibles pour une approche quantitative efficace en recherche en éducation et formation. Les outils de mesure de la corrélation et de l'association, aisément accessibles eux aussi, sont présentés au chapitre 13. Le chapitre 14 présente l'analyse factorielle exploratoire, le chapitre 15 l'analyse de variance.

La troisième partie (chapitres 16 à 22) est consacrée à la modélisation des relations entre variables. Elle présente de façon détaillée un répertoire de modèles d'analyse de variance et de régression.

N'ont été retenues que des méthodes qui présentent un intérêt évident par rapport à des problématiques de recherche en éducation et formation, et pour la mise en

œuvre desquelles, par ailleurs, l'utilisateur peut s'appuyer de façon substantielle sur des logiciels statistiques.

Tout au long de l'ouvrage, la mise en œuvre des méthodes est illustrée au travers d'exemples[1].

Le recours aux logiciels statistiques

Cet ouvrage s'appuie de façon essentielle sur l'usage de logiciels statistiques. Ceux-ci ont, en effet, l'avantage d'automatiser un certain nombre de calculs qui ne présentent pas toujours en eux-mêmes un intérêt conceptuel considérable. Les logiciels facilitent en ce sens la maîtrise progressive des méthodes quantitatives. Disposer d'emblée d'un bon logiciel permet de ne pas être bloqué par le caractère parfois un peu fastidieux du calcul manuel et des manipulations préparatoires sur un tableur non spécialisé. Disposer d'un bon logiciel statistique est indispensable pour éviter des sources inutiles de découragement et d'erreurs de calcul.

Dans cet esprit, l'équipement de base recommandé est le tableur Excel de Microsoft, à compléter avec

[1] Le fichier Excel des tableaux de données analysés dans les exemples est téléchargeable à l'adresse : https://methodes-quantitatives-en-sciences-de-leducation-71.webself.net/

XLSTAT[2], le logiciel statistique en langue française sans doute le plus pédagogique et le plus simple d'utilisation pour un utilisateur non familier à l'heure actuelle. XLSTAT, en version de base, permet de couvrir la plupart des thèmes de cet ouvrage[3]. Mais évidemment, le concept de l'ouvrage n'entraine aucune contrainte quant aux types de logiciels à utiliser, de sorte que le lecteur reste libre d'opter, en fonction de ses préférences, pour toute autre solution logicielle à sa convenance[4].

---

[2] https://www.xlstat.com/fr/. Un grand nombre de tutoriels guidant l'utilisateur d'XLSTAT sont librement accessibles sur internet.

[3] Des ressources complémentaires sont indiquées lorsque ce n'est pas le cas.

[4] Par exemple, Minitab ou SPSS représentent d'excellentes alternatives. Des solutions gratuites – mais plus partielles – existent aussi. Excel en soi, tout d'abord, qui sans constituer un logiciel statistique complet, couvre déjà, en cumulant ses fonctions statistiques (sous l'onglet *Formules*) et son utilitaire d'analyse (sous l'onglet *Données*), une large gamme de besoins. De même, XLSTAT offre depuis 2018 une version gratuite quoique limitée à seulement quelques fonctions. À titre complémentaire, divers sites internet permettent de télécharger gratuitement des modules de calcul statistique en fonction des besoins, par exemple,

- en français : AnaStats

http://www.anastats.fr/outils.php ;

- en anglais : Real Statistics (http://www.real-statistics.com/). Ce site met à disposition un ensemble d'outils, le « Real Statistics Resource Pack », téléchargeable à l'adresse http://www.real-statistics.com/free-download/real-statistics-resource-pack/. Une fois le téléchargement et l'installation effectués conformément aux instructions, il peut s'avérer nécessaire d'accéder aux *Propriétés* du fichier téléchargé (par un clic droit) et de le *Débloquer* (en face du message « *Ce fichier provient d'un autre ordinateur…* » dans l'onglet *Général*). La combinaison de touches « Ctrl m » dans Excel donne accès à la fenêtre de sélection des outils d'analyse de données (analyse de variance, régression, etc.) du pack. Le pack comprend aussi des fonctions statistiques, permettant de réaliser de nombreux tests statistiques, et qui s'utilisent comme les autres fonctions d'Excel, c'est-à-dire après avoir inscrit le signe = dans une cellule de feuille de calcul.

Enfin, des logiciels statistiques gratuits eux aussi mais beaucoup plus complets existent, par exemple SAS ou R, téléchargeables sur internet. Cependant, ils nécessitent un minimum de programmation.

Table des matières

Liste des encadrés

Chapitre 1. Notions de base

Aborder avec aisance les méthodes quantitatives nécessite de comprendre et maitriser d'entrée quelques termes et outils essentiels, d'usage récurrent. Ce premier chapitre les présente.

1.1. INDIVIDU, POPULATION, VARIABLE

L'analyse quantitative porte toujours sur un ensemble d'objets, par exemple les résultats scolaires d'un élève, ou les effectifs des classes d'un établissement, ou les salaires des enseignants, etc. Par exemple, si l'on étudie les notes d'un unique élève, l'objet n'est pas l'élève mais la note. Chacun des objets du groupe d'objets étudiés est un « individu » (ou « unité statistique »). On appelle « population » le groupe d'objets étudiés.

On appelle « variable » une relation qui, à chaque individu d'une population donnée, associe une « valeur ». Par exemple, la variable « Notes obtenues par une promotion d'étudiants à l'examen de sociologie », associe à chaque étudiant une valeur qui est la note que cet étudiant a obtenue à l'épreuve de sociologie. Ou encore, autre exemple, la variable « Diplôme universitaire initial des enseignants de Sciences de l'éducation et de la formation » associe à

chaque enseignant une valeur qui est l'intitulé du premier diplôme universitaire que cet enseignant a obtenu.

On appelle « série statistique » la liste des valeurs qu'une variable associe à tous les individus d'une population. Par exemple, la série statistique des salaires des cinq principaux enseignants de l'école primaire Lambda s'établit comme suit :

Tableau 1.1
École publique Lambda
Traitement mensuel brut des cinq principaux Professeurs des écoles

- 2 306 EUR (Mme Bernard, 10 ans d'ancienneté)
- 2 306 EUR (M. Poussin, 10 ans d'ancienneté)
- 2 765 EUR (M. Rondot, 20 ans d'ancienneté)
- 3 777 EUR (Mme Chasles, 30 ans d'ancienneté)
- 3 777 EUR (Mme Kuznets, 30 ans d'ancienneté)

On appelle « modalités » les *types de valeurs* qu'une variable associe à une population. Dans l'exemple du Tableau 1.1, la série comporte cinq valeurs représentant *trois modalités* (2 306, 2 765, et 3 777). Lorsqu'une variable n'a que deux modalités, on dit qu'elle est *binaire* ou *dichotomique*. Une variable *polytomique* a plusieurs modalités de réponse, mais le

nombre de ces modalités est limité à quelques unités. Lorsque le nombre de modalités est très élevé mais fini (par exemple le nombre d'élèves par lycée, ou le nombre d'enseignants par établissement, ou le nombre d'établissements par région, etc.), la variable est *discrète*. La variable est *continue* si le nombre de ses modalités est infini.

On appelle « caractère » l'identification générique des modalités d'une variable. Par exemple, dans la variable « Diplôme universitaire initial des enseignants en Sciences de l'éducation et de la formation », le caractère est le diplôme universitaire initial.

Suivant la nature de son caractère, une variable est dite quantitative ou qualitative.

Une variable est quantitative si ses modalités de caractère expriment une quantité. Par exemple, dans la variable « Effectifs des classes des établissements d'enseignement primaire » où un nombre d'élèves est associé à chaque classe, chaque modalité exprime une quantité. Lorsque la variable est quantitative, ses modalités peuvent être additionnées. Par exemple la modalité 15 élèves peut être additionnée avec la modalité 20 élèves (soit 35 élèves au total).

Une variable est qualitative (on dit aussi « catégorielle ») si ses modalités n'expriment pas une

quantité. Par exemple, dans un système scolaire pratiquant la notation par lettres de l'alphabet (A – B – C – D), la variable « Notes au contrôle de maths » est qualitative et non pas quantitative. Les lettres permettent sans aucun doute d'exprimer un niveau de performance, mais elles ne représentent pas des quantités, et ne sont donc pas sommables.

Les variables qualitatives peuvent être ordinales ou nominales. Une variable est ordinale si ses modalités expriment un rang ou une gradation logique (par exemple : faible – moyen – fort ; ou encore : pas d'accord – d'accord – tout-à-fait d'accord). Les variables nominales n'expriment pas de rang ou gradation (par exemple : gauche – droite ; vert – jaune ; etc.)

Il importe de ne pas confondre les termes « quantitatif » et « numérique ». Une variable peut être numérique mais qualitative. Par exemple, on peut imaginer un exercice de démocratie scolaire dans le cadre duquel les élèves doivent choisir entre un programme conservateur (codé 1) et un programme progressiste (codé 2). Les modalités de la variable sont exprimées sous forme numérique mais elles ne sont pas sommables : il n'y a aucun sens à sommer 1 et 2 dans ce cas, car 3 ne signifierait rien. De façon plus générale, les codes numériques attribués à des modalités qualitatives ne changent rien au fait que la variable

caractérisée par ces modalités est elle-même qualitative.

1.2. INDICE

Un indice est un nombre qui permet de comparer une valeur à une référence dans le temps ou dans l'espace. Soit x une valeur à un moment donné ou à un endroit donné, soit x^* la référence à laquelle on souhaite comparer x, l'indice de x par rapport à sa référence est défini par :

$$Indice = \frac{x}{x^*}$$

Par exemple, si le nombre d'élèves inscrits est de 300 dans l'établissement de référence A, de 200 dans l'établissement B et de 400 dans l'établissement C :

- l'indice du nombre d'inscrits par rapport à la référence A est de 0,66 dans l'établissement B (200/300) ; et
- l'indice du nombre d'inscrits par rapport à la référence A est de 1,33 dans l'établissement C (400/300).

De même, si le nombre d'inscrits dans un établissement D est de 150 à une date de référence (par exemple octobre 2016), de 250 en octobre 2017 et de 300 en octobre 2018, l'indice du nombre d'inscrits par rapport à la date de référence est de 1,66 en octobre 2017 et de 2 en octobre 2018.

L'usage des indices permet une comparaison directe et simple de valeurs différentes au moyen d'une référence commune. Il permet aisément aussi de mesurer des écarts dans l'espace. Par exemple, l'indice 0,66 de l'établissement B signifie aussi que l'effectif des inscrits de l'établissement B :

- *représente* 66% de celui des inscrits de l'établissement de A ;
- *est de 34% inférieur* à celui de l'établissement A (0,66 – 1 = 0,34) ;
- *est égal à 0,66 fois* celui de l'établissement A ;
- *a un coefficient multiplicateur de* 0,66 par rapport à l'effectif des inscrits de l'établissement A.

L'usage des indices permet tout aussi aisément de mesurer des variations au cours du temps. Par exemple, les indices 1,66 et 2 de l'établissement D signifient aussi que :

- le nombre d'inscrits a *augmenté de*
  - 66% en octobre 2017 par rapport à octobre 2016 (1,66 – 1 = 0,66 = 66%) ;
  - 100% en octobre 2018 par rapport à octobre 2016 (2 – 1 = 1 = 100%) ;
- le nombre d'inscrits a *été multiplié par*
  - 1,66 entre octobre 2016 et octobre 2017 ;
  - 2 entre octobre 2016 et octobre 2018.

Dans la comparaison, la valeur de référence est égale à 1, donc on dit que l'indice est en base 1. Mais une autre pratique courante consiste à exprimer l'indice plutôt en base 100 :

$$Indice\ base\ 100 = Indice\ base\ 1\ \times 100$$

Ainsi par exemple, les indices 0,66 et 1,33 ci-dessus des établissements B et C deviennent respectivement 66 et 133 si on les exprime en base 100. Cela étant, l'expression en base 100 ne change pas le principe : les indices s'interprètent toujours par comparaison avec la base.

1.3. INDICATEURS DE STATISTIQUE DESCRIPTIVE

La statistique descriptive représente le traitement de base applicable à des données quantitatives. Elle permet de mettre en évidence les premières caractéristiques des données observées. Dans le cadre d'un compte-rendu de recherche (rapport de recherche, thèse, article), ces caractéristiques doivent faire partie de la présentation générale des données analysées[5].

Supposons un ensemble de données quantitatives. La statistique descriptive vise à en fournir une présentation

[5] XLSTAT permet d'établir un résumé des statistiques descriptives d'une série à partir de la commande *Statistiques descriptives* sous l'onglet *Description des données*.

synthétique. Cette présentation s'effectue au moyen d'*indicateurs de statistique descriptive*, qui résument les principales propriétés de ces données. Deux types d'indicateurs de statistique descriptive sont d'usage particulièrement fréquent : les *indicateurs de tendance centrale* et les *indicateurs de dispersion*.

1.3.1. Indicateurs de tendance centrale

Soit une population comprenant n individus dont on étudie la variable X. Par exemple un groupe de n étudiants dont on étudie la variable « notes à l'épreuve d'anglais ». La valeur de la variable X pour un étudiant i est notée x_i (Tableau 1.2).

Tableau 1.2.
Notes à l'épreuve d'anglais

	Étudiants ($\boldsymbol{i}$)	Notes à l'épreuve d'anglais ($\boldsymbol{x_i}$)
Étudiant-e 1	Chloé	10
Étudiant-e 2	Lucas	10
Étudiant-e 3	Leo	11
Étudiant-e 4	Nathan	12
Étudiant-e 5	Paul	12
Étudiant-e 6	Noémie	13
Étudiant-e 7	Emma	14
Étudiant-e 8	Léa	14
Étudiant-e 9	Sarah	16
Étudiant-e 10	Abdel	17

Les indicateurs de tendance centrale permettent de mettre en évidence des caractéristiques représentatives de la plupart des valeurs de la série.

1.3.1.1. Fréquence

La fréquence d'une modalité traduit le nombre de fois où cette modalité est représentée dans une population. Elle est définie par le rapport :

$$fréquence = \frac{nombre\ d'occurences\ de\ la\ modalité}{effectif\ de\ la\ population}$$

Elle s'exprime en pourcentage. Par exemple, si 3 élèves sur les 25 que compte une classe ont obtenu la note de 17/20, la fréquence de la modalité 17/20 est $\frac{3}{25} = 0{,}12$. Dans le Tableau 1.2, les modalités 10, 12 et 14 ont chacune une fréquence de 0,20. La fréquence est de 0,10 pour chacune des modalités 11, 13, 16 et 17. La somme des fréquences des modalités d'une série est toujours égale à 1.

1.3.1.2. Mode

Le mode d'une série est la valeur de caractère la plus fréquente dans la série. Par exemple, dans le Tableau 1.2, il y a trois modes, 10-12-14. On dit que la série (c'est-à-dire ici la série des notes à l'épreuve d'anglais)

est *trimodale*. Une série peut avoir un seul mode (*série monomodale*), deux modes (*bimodale*), plusieurs modes (*multimodale* ou *plurimodale*).

1.3.1.3. Médiane

La médiane d'une série est la valeur de caractère qui sépare la série en deux sous-groupes de même effectif. Les valeurs de la série doivent avoir été au préalable rangées par ordre, croissant ou décroissant. C'est le cas de la série du Tableau 1.2, où les valeurs sont rangées en ordre croissant. Par exemple, dans le Tableau 1.2, la médiane est 12,5. Si l'effectif n'avait comporté que les 9 premiers étudiants, la médiane aurait été 12. De façon générale :

- lorsque l'effectif de la série est impair, la médiane est la $(\frac{n+1}{2})$-ième valeur ;
- lorsque l'effectif de la série est pair, la médiane est égale à

$$\frac{\left(\frac{n}{2}\right) ième\ valeur + \left(\frac{n}{2}+1\right) ième\ valeur}{2}$$

1.3.1.4. Moyenne arithmétique

La moyenne arithmétique (« la moyenne » tout court en langage courant) est le rapport entre la somme des valeurs de la série et l'effectif de la série. On note :

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{n} = \frac{\sum_{i=1}^{N} x_i}{n}$$

où

- $\overline{x}$ est la moyenne arithmétique ;
- $x_1, x_2, x_3, \cdots, x_N$ représentent les valeurs du caractère pour le premier, le deuxième, le troisième et ainsi de suite jusqu'au dernier individu de la série ;
- $\sum$, l'opérateur *Sigma* qui indique que les valeurs de la série sont additionnées.

Par exemple, dans la série du Tableau 1.2, la moyenne est :

$$\overline{x} = \frac{10 + 10 + 11 + 12 + 12 + 13 + 14 + 14 + 16 + 17}{10} = 12{,}9$$

1.3.1.5. Moyenne géométrique

La moyenne géométrique G d'une série d'effectif n est donnée par :

$$G = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_N} = \sqrt[n]{\prod_{i=1}^{N} x_i}$$

où

- $\sqrt[n]{\ }$ est la racine n-ième (racine carrée quand $= 2$, racine cubique quand $n = 3$, racine quatrième quand $n = 4$, et ainsi de suite) ;
- $\prod$, l'opérateur *Pi*, indique que les valeurs de la série sont multipliées.

La moyenne géométrique est utile en particulier pour calculer des rythmes moyens de croissance. Imaginons par exemple une réforme mise en place en 2010 en vue de réduire le décrochage scolaire. Pendant les cinq années qui suivent la réforme, les effectifs inscrits augmentent aux rythmes annuels présentés dans le Tableau 1.3.

Tableau 1.3. Taux de croissance annuels des inscrits suite à la réforme

Année	Taux de croissance
2011	0,1 %
2012	0,5 %
2013	1,2 %
2014	4,1 %
2015	5,5 %

Étant donné que le taux d'une année s'entend par rapport à l'année précédente (par exemple, le nombre d'inscrits en 2015 est de 5,5% supérieur au nombre d'inscrits en 2014), on ne peut pas additionner ces taux et en faire la moyenne arithmétique car chacun de ces taux est exprimé en une unité de compte différente

(c'est-à-dire le nombre d'inscrits de l'année précédente). Il faut donc plutôt utiliser la moyenne géométrique, qui permet de tenir compte du caractère composé des taux de croissance (la croissance d'une période sert de base au calcul de la croissance de la période suivante).

On commence par traduire les taux de croissance en indices de croissance, puis on calcule ensuite la moyenne des indices (Tableau 1.4) :

Tableau 1.4. Indices de croissance annuels des inscrits suite à la réforme

Année	Indices de croissance
2011	1,001
2012	1,005
2013	1,012
2014	1,041
2015	1,055

L'indice de croissance annuel moyen est la moyenne géométrique des indices de croissance annuels :

$$\begin{aligned} &\textit{indice de croissance annuel moyen} \\ &= \sqrt[5]{1{,}001 \times 1{,}005 \times 1{,}012 \times 1{,}041 \times 1{,}055} \\ &= 1{,}0225 \end{aligned}$$

Le taux de croissance annuel moyen des inscrits suite à la réforme est donc de 2,25%.

1.3.2. Indicateurs de dispersion

Les indicateurs de dispersion cherchent plutôt à souligner les éléments de variabilité et d'hétérogénéité au sein de la série.

1.3.2.1. Minimum, maximum, extremum, étendue

La valeur la plus basse (minimum) et la plus haute (maximum) d'une série en sont les extrema (par exemple 10 et 17 dans le Tableau 1.2).

L'étendue d'une série est l'écart entre son maximum et son minimum (7 dans l'exemple).

1.3.2.2. Quartiles, déciles, centiles

Lorsqu'une série a un effectif suffisamment élevé, on peut en étudier l'hétérogénéité en la décomposant en sous-groupes et en comparant les caractéristiques et comportements de ces sous-groupes.

Une méthode de décomposition usuelle consiste à découper la série en quarts, ou en dixièmes, ou en centièmes. On appelle quartiles, déciles et centiles les valeurs de caractère qui délimitent les sous-groupes.

Dans une décomposition par quarts, la série, ordonnée par valeurs croissantes (comme par exemple dans le Tableau 1.2), est découpée en quatre quarts. On définit comme suit 3 quartiles, Q_1, Q_2 et Q_3:

- Q_1, le premier quartile, est la valeur telle qu'un quart de l'effectif a une valeur de caractère inférieure à Q_1 ;
- Q_2, le deuxième quartile, est la valeur telle que la moitié de l'effectif a une valeur inférieure à Q_2. Q_2 est égal à la médiane ;
- Q_3, le troisième quartile, est la valeur telle que trois quarts de l'effectif ont une valeur inférieure à Q_3.

Avec les extrema, les quartiles constituent les bornes de chaque quart. Le premier quart est donc délimité par le minimum (borne inférieure) et le premier quartile (borne supérieure) ; le deuxième quart, par le premier quartile et le deuxième quartile ; le troisième quart, par le deuxième quartile et le troisième quartile ; et le quatrième quart, par le troisième quartile et le maximum.

EXEMPLE 1.1

Considérons le pourcentage d'étudiants salariés dans différents pays (Tableau 1.5).

Tableau 1.5.
Part des étudiants salariés dans la tranche des 15-29 ans dans les pays de l'OCDE (%) – Année 2011

Source : OCDE (2013), *Regards sur l'éducation* (Annexe Tableau C5.2a).

NB : la moyenne de l'OCDE s'établit à 11%.

N°	Pays	%
1	Slovaquie	2,1
2	Grèce	2,2
3	Hongrie	2,2
4	Italie	2,5
5	Belgique	3,5
6	République Tchèque	3,6
7	Espagne	4,7
8	Corée du sud	5,2
9	Portugal	5,3
10	Turquie	5,5
11	France	5,9
12	Luxembourg	5,9
13	Chili	6,8
14	Mexique	6,9
15	Irlande	7,1
16	Pologne	7,8
17	Allemagne	8,5
18	Autriche	9,6
19	Israël	10,5
20	Estonie	10,9

N°	Pays	%
21	Suède	11,1
22	Royaume Uni	11,3
23	Suisse	11,9
24	États-Unis	15,1
25	Norvège	15,3
26	Finlande	16,0
27	Slovénie	16,9
28	Nouvelle Zélande	17,4
29	Canada	17,6
30	Australie	21,1
31	Islande	26,6
32	Danemark	32,1
33	Pays-Bas	32,4

Une méthode simple pour calculer les quartiles est celle des *médianes partielles*[6].

Calcul de Q_1

On considère la première moitié de la série (ou la première moitié à laquelle on ajoute la médiane, si l'effectif de la série est impair), et on en calcule la médiane. Ici par exemple, l'effectif de la série est impair ($n = 33$), la médiane de l'ensemble de la série

[6] Il en existe d'autres, qu'utilisent par exemple certains logiciels comme Excel.

est la 17ème valeur, soit 8,5% (pourcentage pour l'Allemagne). La première moitié de la série comprend donc les valeurs de 2,1 à 8,5. La médiane de cette première moitié (première « médiane partielle ») est la 9ème valeur, c'est-à-dire 5,3% (pourcentage du Portugal). Cette première médiane partielle est le premier quartile, Q_1.

Calcul de Q_3

On considère à présent la seconde moitié de la série (augmentée de la médiane si l'effectif de la série est impair), et on en calcule la médiane partielle. Dans l'exemple, la seconde moitié de la série comprend les valeurs de 8,5 à 32,4. La médiane partielle de cette seconde moitié (seconde « médiane partielle ») est la 25ème valeur de la série, c'est-à-dire 15,3 (pourcentage de la Norvège). Cette seconde médiane partielle est le troisième quartile, Q_3.

Résultats

$$\begin{cases} Q_1 = 5{,}3 \\ Q_2 \text{ est la médiane : } Q_2 = 8{,}5 \\ Q_3 = 15{,}3 \end{cases}$$

L'identification des quarts permet ainsi de souligner les éléments d'hétérogénéité. Par exemple, alors que la moyenne de l'OCDE s'établit à 11%, on voit que les pourcentages sont de plus de moitié inférieurs à la

moyenne dans le premier quart, et peuvent lui être jusqu'à deux fois supérieurs dans le quatrième.

On peut observer toutefois que les valeurs extrêmes sont parfois aberrantes (c'est-à-dire considérablement plus faibles ou plus élevées que le reste de la série) et dans ce cas influencent très sensiblement le comportement des premiers et derniers quarts. Dans l'exemple, les deux dernières valeurs de la série (Danemark et Pays-Bas) pèsent fortement sur la caractérisation du dernier quart. L'influence excessive des valeurs extrêmes conduit souvent à les écarter de l'analyse. Pour ce faire, soit on élimine de la série les valeurs extrêmes elles-mêmes, soit on élimine de la segmentation les sous-groupes qui contiennent les valeurs extrêmes. Lorsque le découpage est par quarts, éliminer les premier et dernier quarts revient cependant à amputer la série de la moitié de son effectif, ce qui est beaucoup. On recourt alors plutôt – si l'effectif de la série est suffisamment élevé – à un découpage en dixièmes ou centièmes. On définit 9 déciles (D_1 à D_9), qui partagent la série en 10 dixièmes ; ou 99 centiles (C_1 à C_{99}), qui la partagent en 100 centièmes. On peut alors exclure les premier et dixième sous-groupes et ne conserver pour l'analyse que les 8 dixièmes restants (soit 80% de la série) ; ou exclure les premier et centième sous-groupes et ne conserver que les 98 centièmes restants (98% de la série). L'intérêt de l'opération est de disposer de fourchettes qui donnent

une idée de la dispersion de la majorité des valeurs de la série :

- 50% des valeurs de la série sont comprises entre Q_1 et Q_3 ;
- 80% des valeurs de la série sont comprises entre D_1 et D_9 ;
- 98% des valeurs de la série sont comprises entre C_1 et C_{99}.

1.3.2.3. Écart à la moyenne : écart moyen, écart-type et variance

L'idée ici est de représenter la dispersion des valeurs de la série par une mesure de l'écart des valeurs à la moyenne de la série.

Une première méthode consiste à calculer *l'écart moyen*, qui est la moyenne des écarts à la moyenne :

$$écart\ moyen = \frac{\sum_{i=1}^{N}(x_i - \bar{x})}{n}$$

EXEMPLE 1.2

Soit la série décrite dans la première colonne (1) du Tableau 1.6. La deuxième colonne décrit les étapes du calcul de l'écart moyen.

Tableau 1.6.
Calcul d'écart moyen

x	$x - \overline{x}$
(1)	(2)
12	5,4
8	1,4
7	0,4
4	-2,6
2	-4,6
Moyenne = 6,6	Total = 0
	Écart moyen = 0/5 = 0

L'inconvénient de l'écart moyen est que l'écart d'une valeur à la moyenne peut être positif ou négatif. Donc les écarts des différentes valeurs se compensent au moins en partie[7], de sorte que, *in fine*, l'écart moyen reflète mal les distances entre les valeurs et la moyenne. Dans l'exemple, l'écart moyen est nul, suggérant que

[7] Dans cet exemple, les écarts s'annulent, pour mieux montrer comment l'écart moyen peut ne pas refléter les écarts réels. Mais dans l'absolu il n'y a aucune raison pour que l'écart moyen soit toujours nul. Il peut aussi – tout dépend de la série – être positif ou négatif.

toutes les valeurs sont très proches de la moyenne, alors que ce n'est pas du tout le cas.

Une méthode plus satisfaisante consiste alors à utiliser *l'écart-type*, défini comme la moyenne des carrés des écarts à la moyenne. L'écart-type est noté σ et calculé par :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n}}$$

EXEMPLE 1.3

Reprenons dans le Tableau 1.7 ci-dessous la série décrite dans la colonne (1) du Tableau 1.6. La troisième colonne du Tableau 1.7 décrit les étapes du calcul de l'écart-type.

Tableau 1.7.
Calcul d'écart moyen et d'écart-type

x	$x - \bar{x}$	$(x - \bar{x})^2$
(1)	(2)	(3)
12	5,4	29,16
8	1,4	1,96
7	0,4	0,16
4	-2,6	6,76
2	-4,6	21,16
Moyenne : 6,66	Total = 0	Total = 59,2
	Écart moyen = 0/5 = 0	Variance = 59,2/5 = 11,84
		Écart-type $= \sqrt{11,84} \cong \pm 3,44$

Par rapport à l'écart moyen, l'avantage de l'écart-type est que, pour chaque valeur, on prend le carré de son écart à la moyenne ; il n'y a donc plus de terme négatif ni de compensation. L'écart-type reflète donc mieux les écarts à la moyenne que ne le fait l'écart-moyen, ce qu'illustre bien l'exemple.

On appelle *variance* le radicande (l'expression sous le radical $\sqrt{\ }$) de l'écart-type :

$$variance = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n}$$

La variance est donc la somme des carrés des écarts divisée par l'effectif. C'est le carré de l'écart-type : $variance = \sigma^2$.

Écart-type et variance sont des indicateurs très utilisés en statistique. Du point de vue de la dispersion, chacun des deux permet de représenter de façon synthétique par un chiffre unique la tendance des valeurs de la série à s'écarter de la moyenne. Soit une série : plus sa dispersion est forte, plus son écart-type et sa variance sont élevés.

1.3.2.4. Coefficient de variation

Le coefficient de variation est parfois appelé aussi *coefficient de dispersion*. La question à laquelle le coefficient de variation permet de répondre est la suivante : comment comparer la dispersion de deux séries (ou plus) dont les moyennes sont différentes ? Par exemple, le niveau des élèves est-il homogène dans les classes de différents enseignants d'une même matière ? Ou bien : y a-t-il homogénéité dans l'origine socio-économique des élèves dans différents établissements ou régions ? Etc.

Le coefficient de variation est défini par le rapport écart-type / moyenne. Il permet de comparer la dispersion de séries de moyennes et écarts-types différents : la série est d'autant plus dispersée que le coefficient de variation est élevé.

EXEMPLE 1.4

On considère la dépense publique par élève dans le primaire, le secondaire et le supérieur dans différents pays de l'OCDE (Tableau 1.8). On cherche à savoir si les écarts entre pays sont plus particulièrement marqués dans l'un ou l'autre des trois niveaux d'enseignement.

Tableau 1.8.
Dépense publique par élève dans l'enseignement primaire secondaire et supérieur dans les différents pays de l'OCDE – Année 2010 – En USD

Source : OCDE (2013), *Regards sur l'éducation* (Annexe Tableau B1.1a).

	Primaire	Secondaire	Supérieur
Australie	9 463	10 350	15 142
Autriche	10 244	12 551	15 007
Belgique	8 852	11 004	15 179
Canada	8 933	–	22 475
Chili	3 301	3 110	7 101
République Tchèque	4 120	6 546	7 635
Danemark	10 935	11 747	18 977
Estonie	5 140	6 444	6 501

	Primaire	Secondaire	Supérieur
Finlande	7 624	9 162	16 714
France	6 622	10 877	15 067
Hongrie	4 684	4 553	8 745
Islande	9 482	7 841	8 728
Irlande	8 384	11 380	16 008
Israël	5 758	5 616	10 730
Italie	8 296	8 607	9 580
Japon	8 353	9 957	16 015
Corée du sud	6 601	8 060	9 972
Luxembourg	21 240	17 633	–
Mexique	2 331	2 632	7 872
Pays-Bas	7 954	11 838	17 161
Nouvelle Zélande	6 842	8 170	10 418
Norvège	12 255	13 852	18 512
Pologne	5 937	5 483	8 866
Portugal	5 922	8 882	10 578
Slovaquie	5 732	4 806	6 904
Slovénie	8 935	8 187	9 693
Espagne	7 291	9 608	13 373
Suède	9 987	10 185	19 562
Suisse	11 513	14 972	21 893
Turquie	1 860	2 470	–
Royaume uni	9 369	10 452	15 862
États-Unis	11 193	12 464	25 576
Moyenne OCDE	*7 974*	*9 014*	*13 528*
Écart-type	*3561,76*	*3601,65*	*5170,37*
Coeff. de variation	*0,44*	*0,39*	*0,38*

On constate que l'écart-type est plus élevé dans l'enseignement supérieur, ce qui semble suggérer que les pratiques d'un pays à l'autre varient beaucoup plus dans le supérieur que ce n'est le cas dans le primaire et le secondaire. Mais les séries ayant des moyennes différentes, cette conclusion est hasardeuse. Le coefficient de variation la remet en effet en cause : il est plus faible dans le supérieur, indiquant que les pratiques des pays sont en réalité, pour ce niveau d'enseignement, plus homogènes qu'elles ne le sont dans le primaire et le secondaire.

1.4. COVARIANCE

Soient deux variables, X et Y. La covariance indique si la variation de l'une est liée à la variation de l'autre, et dans l'affirmative, si cette relation est linéaire. Par exemple, y a-t-il une relation linéaire entre le nombre d'heures d'enseignement et la réussite scolaire ? Entre l'origine sociale et le retard scolaire ? Entre le soutien de l'institution et la motivation des enseignants ? Etc.

S'il n'y a pas de relation linéaire entre les deux variables (mais il peut y avoir entre elles une relation non-linéaire : voir Encadré 1.1), la covariance est nulle. S'il y a une relation linéaire, la covariance est non nulle et est alors positive si les deux variables varient dans le

même sens, ou négative si elles varient en sens contraire.

Pour deux variables, $X = \{x_1, x_2, \dots, x_N\}$ d'effectif n et de moyenne $\bar{x}$, et $Y = \{y_1, y_2, \dots, y_N\}$ d'effectif n et de moyenne $\bar{y}$, la covariance est mesurée par

$$Cov = \frac{\sum[(x - \bar{x}) \times (y - \bar{y})]}{n}$$

EXEMPLE 1.5

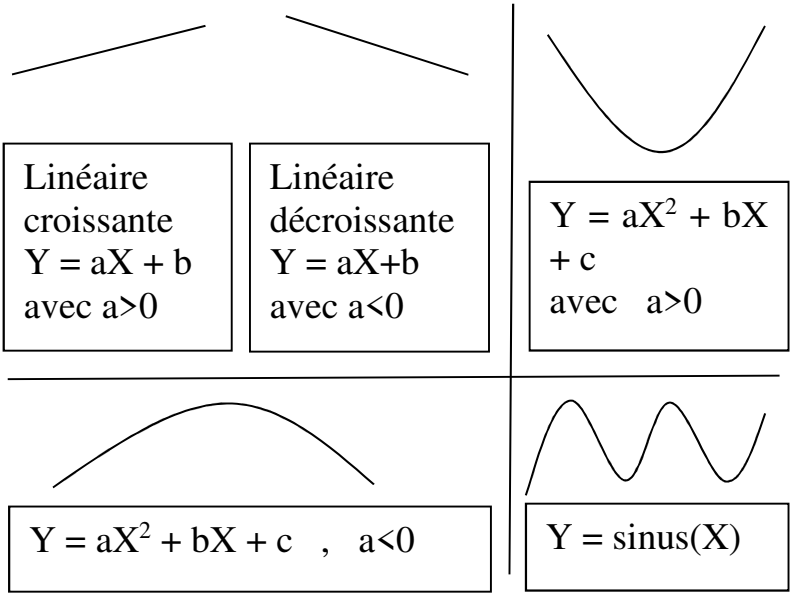
On calcule la covariance des variables *X* et *Y* dont les valeurs sont répertoriées dans les deux premières colonnes du Tableau 1.9 ci-après.

Tableau 1.9
Calcul de covariance

X	Y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) \times (y - \bar{y})$
12	25	5,4	9,6	51,84
8	35	1,4	19,6	27,44
7	7	0,4	-8,4	-3,36
4	9	-2,6	-6,4	16,64
2	1	-4,6	-14,4	66,24
$\bar{x} = 6,6$	$\bar{y} = 15,4$	$\sum[(x - \bar{x}) \times (y - \bar{y})] = 158,8$		
		$Cov_{XY} = 158,8/5 = 31,76$		

ENCADRÉ 1.1 – RELATIONS LINÉAIRES ET NON-LINÉAIRES

Une relation est linéaire si elle peut être représentée graphiquement par une droite. Elle est non-linéaire dans le cas contraire. L'expression mathématique d'une relation linéaire entre X et Y est : $Y = aX + b$ où a et b sont des paramètres. Exemples de représentations graphiques de relations linéaires et non linéaires :



Dans cet exemple, la valeur de la covariance indique qu'il y a entre les variables *X* et *Y* une relation linéaire positive.

1.5. DISTRIBUTION STATISTIQUE ET COURBE NORMALE

On appelle « distribution statistique » (ou « distribution des fréquences ») la liste des fréquences des modalités d'une variable. La distribution statistique se représente sous forme de tableau ou de courbe indiquant la fréquence de chacune des modalités. Par exemple le Tableau 1.10 et le Graphique 1.1 ci-dessous décrivent la distribution statistique d'une variable « Catégorie socio-professionnelle du chef de famille » pour les élèves d'un établissement scolaire.

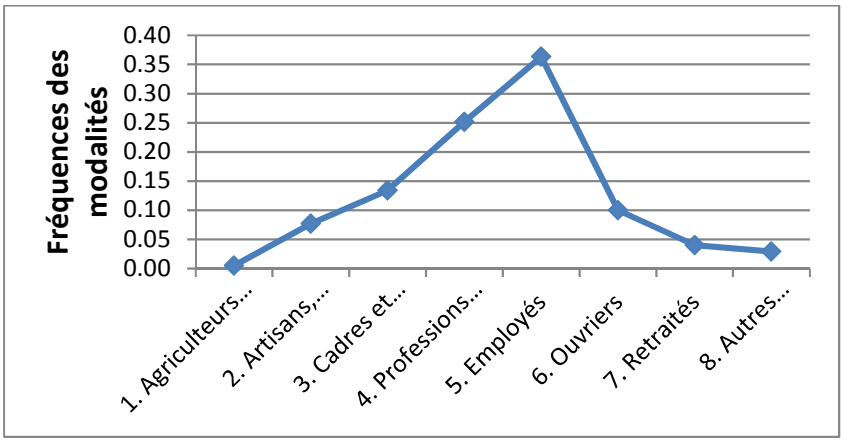
Tableau 1.10
Tableau de la distribution statistique de la variable « Catégorie socio-professionnelle du chef de famille »

Catégorie socio-professionnelle	Effectif de la modalité	Fréquence de la modalité
1. Agriculteurs exploitants	15	0,01
2. Artisans, commerçants et chefs d'entreprise	230	0,08
3. Cadres et professions intellectuelles supérieures	402	0,13

Catégorie socio-professionnelle	Effectif de la modalité	Fréquence de la modalité
4. Professions Intermédiaires	755	0,25
5. Employés	1090	0,36
6. Ouvriers	300	0,10
7. Retraités	121	0,04
8. Autres personnes sans activité professionnelle	87	0,03
Total	*3000*	*1*

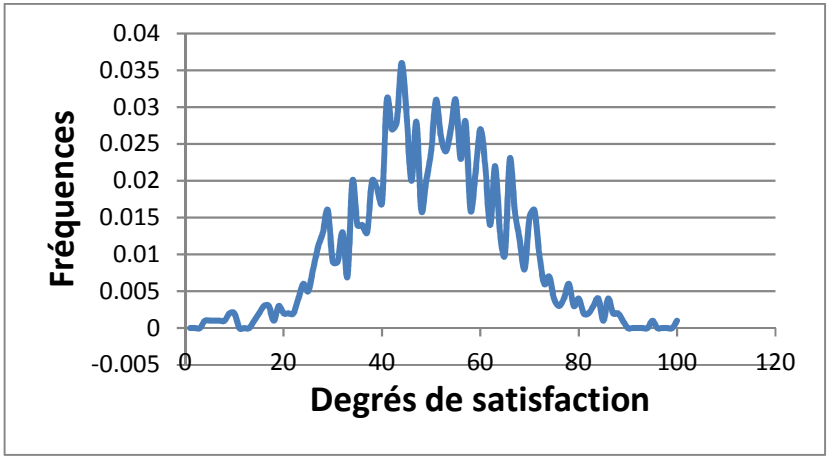
Graphique 1.1.

Courbe de la distribution statistique de la variable « Catégorie socio-professionnelle du chef de famille »



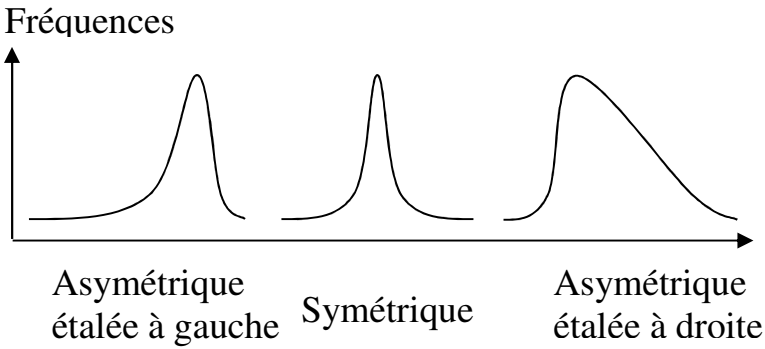
Le principe est le même lorsque la variable est quantitative. Par exemple, le Graphique 1.2 montre la courbe (fictive) des fréquences des degrés de satisfaction (sur une échelle de 1 à 100) exprimés par 1000 répondants à l'égard de la politique de la carte scolaire :

Graphique 1.2.
Fréquences des degrés de satisfaction exprimés



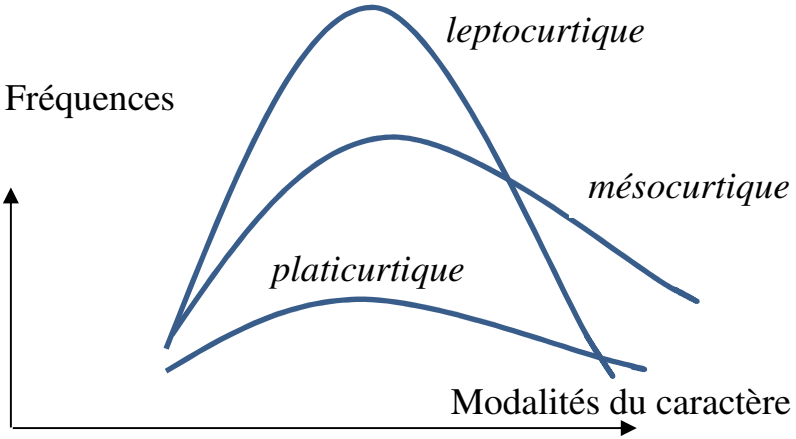
Une distribution de fréquences peut être plus ou moins symétrique :

Graphique 1.3.
Symétrie (*skewness*) de la courbe des fréquences



Une distribution de fréquences peut être plus ou moins aplatie :

Graphique 1.4.
Aplatissement (*kurtosis*) de la courbe des fréquences



La courbe des fréquences peut être plutôt aplatie (« *platicurtique* »), ou plutôt moyennement aplatie (« *mésocurtique* »), ou plutôt pointue (« l*eptocurtique* »).

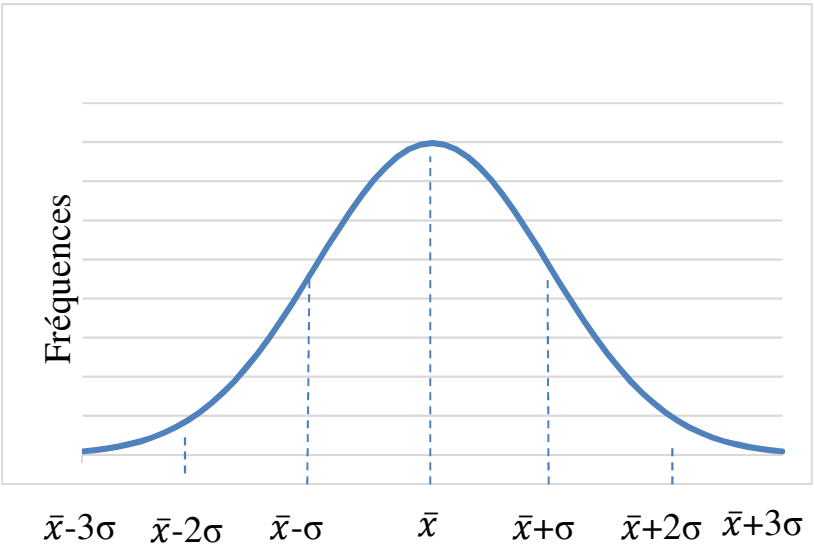
Une distribution est « normale » lorsqu'elle suit la « loi normale » (ou loi de Gauss ou loi de Laplace-Gauss). La loi normale décrit la tendance des résultats d'une expérience aléatoire (c'est-à-dire une expérience dont le résultat dépend du hasard) lorsque cette expérience est répétée un très grand nombre de fois. Une distribution normale est à la fois symétrique et mésocurtique. La symétrie signifie que la moyenne de la série est égale à son mode et à sa médiane. Une importante propriété de la distribution normale est que les valeurs de la série se répartissent suivant un schéma spécifique, à savoir que :

- 68% des valeurs sont comprises dans l'intervalle $[\bar{x} - \sigma\ ;\ \bar{x} + \sigma]$ où $\bar{x}$ est la moyenne et σ l'écart-type de la série ;
- 95% des valeurs sont comprises dans l'intervalle $[\bar{x} - 2\sigma\ ;\ \bar{x} + 2\sigma]$;
- 99,8% des valeurs sont comprises dans l'intervalle $[\bar{x} - 3\sigma\ ;\ \bar{x} + 3\sigma]$.

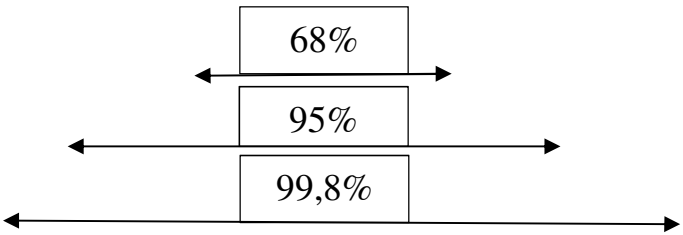
Cette répartition est représentée graphiquement sur la courbe de la loi normale par la surface entre la courbe

elle-même et l'axe des abscisses pour chacun de ces intervalles.

Graphique 1.5.
Courbe de la loi normale



$\bar{x}$ = moyenne =mode=médiane



Première partie : collecter et préparer les données

La démarche quantitative commence par la collecte des données qui retracent le phénomène à analyser. Les données susceptibles d'être analysées par approche quantitative sont de formes et natures variées.

Du point de vue de la forme, on peut distinguer entre données textuelles et données numériques.

Dans le premier cas, il s'agit par exemple de textes littéraires, de documents politiques, d'articles de presse ou de transcriptions d'entretiens. Bien qu'habituellement analysés par approche qualitative suivant des méthodes d'analyse de contenu, les textes sont aussi susceptibles, le cas échéant, d'être approchés sous l'angle quantitatif. C'est l'objet de la lexicométrie[8].

Les données numériques peuvent être soit quantitatives soit qualitatives (ordinales ou nominales). Elles représentent la majorité des données habituellement soumises à analyse quantitative.

Du point de vue de leur nature, les données susceptibles d'être traitées par analyse quantitative peuvent être réparties entre données simulées et données observées,

[8] Voir chapitre 4.

d'une part ; et entre données perceptuelles et données factuelles d'autre part. Les données simulées sont des données fictives, créées par générateurs de données suivant des procédures plus ou moins sophistiquées. Elles servent par exemple dans les phases de mise au point ou d'essai de modèles, ou encore en appui à des illustrations pédagogiques. Au contraire, les données observées sont tirées de situations réelles et décrivent le monde réel. Les données perceptuelles expriment des opinions, par opposition aux données factuelles (par exemple démographiques ou physiques), qui décrivent des faits indépendamment des opinions.

Chapitre 2. Collecter les données

Toute donnée n'est pas bonne à prendre. L'identification des bonnes données à collecter s'effectue sur la base de la question de recherche, et en fonction du cadre conceptuel et théorique adopté pour ladite recherche. Les données à collecter doivent évidemment porter de façon précise sur l'objet de la recherche, et en respecter les conceptualisations. Elles doivent constituer la meilleure représentation empirique possible des concepts à l'œuvre dans cette recherche.

Un grand nombre de données statistiques relatives à l'éducation et à la formation existent aujourd'hui en France et à travers le monde. Elles sont produites par une multitude de sources, institutionnelles ou individuelles (recherches académiques notamment), publiques ou privées, internationales, nationales ou locales. Au niveau international, l'Organisation des Nations Unies pour l'éducation, la science et la culture (UNESCO)[9], la Banque mondiale[10], l'Organisation pour la Coopération et le Développement Économiques (OCDE)[11], ou encore l'Office statistique de l'Union

[9] http://data.uis.unesco.org/?lang=fr

[10] https://donnees.banquemondiale.org/

[11] http://www.oecd.org/fr/edu/regards-sur-l-education-19991495.htm

européenne (Eurostat)[12], par exemple, mettent à disposition du grand public une abondance de données statistiques sur l'éducation et la formation. Ces données sont généralement agrégées (statistiques pour l'ensemble d'un pays dans tel ou tel domaine), mais comprennent souvent aussi des *micro-données*, c'est-à-dire des données obtenues de chacun des répondants individuels lors d'enquêtes. En France, deux importants producteurs de données sur l'éducation sont le ministère de l'Éducation nationale[13] et le ministère de l'Enseignement supérieur[14]. Au niveau local, de très nombreuses sources de données sur l'éducation et la formation existent également. Par exemple, chaque établissement d'enseignement dispose de données, et en particulier de données administratives sur ses étudiants et de données sur leurs performances académiques.

Pour autant, il n'est pas rare de manquer de données appropriées lorsqu'il s'agit de faire de la recherche. Bien souvent en effet, les données disponibles sont trop générales et non adaptées à l'objet spécifique de telle ou telle recherche. En particulier, les définitions adoptées par les producteurs de données pour

[12] http://ec.europa.eu/eurostat/data/database

[13] http://www.education.gouv.fr/bcp/mainFrame.jsp?p=1

[14] http://www.enseignementsup-recherche.gouv.fr/pid24748/statistiques-et-analyses.html

construire leurs indicateurs et collecter les données correspondantes ne sont pas nécessairement compatibles avec le cadre théorique et conceptuel de tel ou tel travail de recherche.

Il n'est donc pas rare que le chercheur en éducation et formation soit contraint de collecter lui-même les données pertinentes pour sa recherche. C'est l'objet de l'enquête de terrain. Dans l'absolu, l'enquête peut prendre la forme de l'observation en immersion, du questionnaire ou de l'entretien. Seules ces deux dernières formes (et davantage le questionnaire que l'entretien) sont susceptibles de générer des données traitables par approche quantitative.

2.1. L'ENQUÊTE PAR QUESTIONNAIRE

D'un point de vue quantitatif, l'enquête par questionnaire présente l'avantage de permettre de collecter, auprès d'un grand nombre de répondants, des données textuelles et/ou numériques susceptibles d'être traitées par lexicométrie et/ou méthodes statistiques. Collecter des réponses par questionnaire auto-administré accessible sur internet à des milliers de répondants est aujourd'hui pratique courante. Or il parait évident que plus le nombre de répondants est élevé, plus il y a de chances que les conclusions de l'analyse soient généralisables.

De nombreux ouvrages présentent les méthodes d'élaboration de questionnaire et de conduite d'enquête, et il existe également des logiciels[15] qui en facilitent la mise en œuvre. Sur certains thèmes, il existe des questionnaires psychométriques déjà validés qui peuvent servir de point de départ[16]. On n'y reviendra donc pas en détail dans le présent ouvrage, consacré plutôt à l'analyse des données déjà collectées. Quelques rappels essentiels s'imposent cependant.

2.1.1. Préparation du questionnaire

La préparation d'un questionnaire s'articule en deux principales phases : la phase de conception et la phase pilote.

2.1.1.1. Conception du questionnaire

Tout d'abord, construire un questionnaire nécessite de respecter un minimum de règles. L'anonymat doit être garanti, et les règles relatives au traitement des données personnelles (y compris l'âge, la profession et toute

[15] Par exemple Modalisa, Sphynx, SurveyMonkey.

[16] Par exemple, le *Burros Institute of Mental Measurement* de l'Université du Nebraska (http://buros.org/) dispose d'une base de données répertoriant un grand nombre de ces instruments, qui nécessitent cependant traduction et adaptation pour des répondants en France.

information susceptible de permettre l'identification du répondant par croisement des données) respectées[17].

[17] Le Règlement général de l'Union européenne pour la protection des données à caractère personnel (RGPD) est entré en vigueur le 25 mai 2018. Le texte complet en est téléchargeable à l'adresse :
https://www.cjoint.com/doc/17_12/GLnmzFxp4tM_rgpd.pdf.
Ses principales implications pour l'enquête par questionnaire dans le cadre de la recherche quantitative s'établissent comme suit :

- Sauf sous certaines conditions, sont interdits : le traitement de données *révélant* l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique (article 9).
- Lorsqu'elles ne sont pas anonymes, les données collectées doivent être au moins « pseudonymisées » (article 89).
- Le consentement du répondant au traitement de ses données aux fins qui lui sont indiquées doit être obtenu, et trace de ce consentement doit être conservée. Le répondant doit être informé de la durée de conservation des données. Le répondant a le droit de retirer son consentement à tout moment (ce qui ne remet pas en cause cependant la licéité de l'usage fait des données avant le retrait). Le répondant doit être informé de son droit de retrait avant de donner son consentement. La procédure pour retirer le consentement doit être « aussi simple » que la procédure pour le donner. Le répondant doit être informé de l'identité et des coordonnées du responsable auquel adresser sa

En principe, un questionnaire commence par une introduction qui, d'une part, présente le cadre général dans lequel le questionnaire s'inscrit, de façon que le répondant comprenne la démarche à laquelle on lui demande de participer ; et d'autre part valorise et encourage la participation du potentiel répondant à la démarche proposée.

La première partie de l'introduction (présentation) indique (brièvement) la nature, l'objet, le cadre institutionnel[18] et scientifique, les objectifs et les enjeux de la recherche, en soulignant plus particulièrement les aspects de nature à intéresser et motiver le type de répondants visé.

---

demande. Sa requête doit être satisfaite gratuitement dans un délai d'un mois, et notification du retrait doit lui être faite (articles 6, 7, 12, 13).

- Le répondant doit pouvoir conserver une copie de ses réponses (article 20).
- Le responsable de l'enquête doit, à l'aide de mesures techniques ou organisationnelles appropriées, assurer la sécurité et la confidentialité des données collectées, y compris la protection contre le traitement non autorisé ou illicite (articles 5, 25).

[18] Le chercheur responsable doit également être mentionné (voire présenté) s'il ne l'a déjà été lors de la procédure de prise de contact avec le répondant potentiel, par exemple dans le courriel d'invitation à répondre au questionnaire.

La deuxième partie de l'introduction (valorisation et encouragement) consiste à expliquer au répondant potentiel que sa participation est indispensable, et que rien ne peut se faire sans lui. On peut aussi expliquer que par ses réponses, le répondant contribuera à apporter des améliorations dans le futur (par exemple améliorer les pratiques pédagogiques, ou bien tel ou tel aspect des politiques d'éducation). Une forme d'incitation consiste, par exemple, à donner au répondant la possibilité d'être, par la suite, tenu informé des résultats de la recherche.

L'ensemble de l'introduction doit être assez bref, clair et lisible. L'introduction s'achève par une indication sur la durée de temps nécessaire pour répondre au questionnaire. C'est aussi là (ou alors en fin de questionnaire) que l'on remercie le répondant pour sa participation.

L'introduction est suivie par des consignes qui indiquent au répondant comment procéder pour répondre, et qui lui fournissent toutes indications nécessaires à cette fin. Les consignes doivent être concises, de façon à éviter que le répondant ne les lise pas. Elles doivent en même temps être complètes et claires, afin d'éviter que le répondant puisse se sentir perdu ou désorienté et finisse par abandonner. Le cas échéant, les modalités suivant lesquelles le répondant peut obtenir de l'assistance peuvent être indiquées.

Les questions, ensuite, doivent être de nature à apporter une information utile par rapport à la question de recherche. Elles doivent être formulées de façon claire, être précises, et être compréhensibles pour tout répondant (éviter le jargon et les acronymes, fournir des exemples si nécessaire). Elles doivent être suffisamment peu nombreuses pour ne pas lasser le répondant et pour éviter de l'amener à répondre distraitement et au hasard (ou même pas du tout) en fin de questionnaire. Il faut également veiller à la neutralité du questionnaire : la façon de formuler les questions ou l'ordre dans lequel les questions sont posées, par exemple, peuvent engendrer des biais dans les réponses. À chaque question, les modalités de réponses proposées au répondant doivent être peu nombreuses et clairement différenciées, afin d'éviter confusion et réponse au hasard. Les modalités additionnelles « Autre » et « Sans opinion » permettent de ne pas forcer le répondant à fournir des réponses trompeuses.

En outre, il importe de garder à l'esprit que les réponses relèvent toujours de la déclaration : un répondant est rarement supposé fournir les preuves de ses assertions. De même, le répondant peut vouloir donner de soi une image qu'il juge plus flatteuse, ou dont il pense qu'elle correspond mieux à ce qu'il croit être attendu de lui

(effet de désirabilité sociale ; effet Bradley[19]). Par conséquent, la formulation des questions doit éviter de suggérer ou d'inspirer au répondant des pistes de réponses qu'il pourrait trouver plus tentantes que la stricte exactitude. Il peut parfois être nécessaire aussi de prévoir, au fil du questionnaire, des questions de recoupement, dont l'objectif est d'aborder sous un angle différent un aspect déjà évoqué dans une autre question, de façon à vérifier si les réponses concernant un même aspect sont cohérentes entre elles.

2.1.1.2. Phase pilote

Une fois le questionnaire conçu, la phase pilote vise à en vérifier la validité et l'intelligibilité, de façon à en améliorer la qualité et l'efficacité avant administration aux répondants. La vérification de la validité porte sur trois principaux volets : couverture du champ, fiabilité, et cohérence interne. La vérification de l'intelligibilité vise à s'assurer que le questionnaire sera lisible et compréhensible pour le public auquel il s'adresse.

2.1.1.2.1. Couverture du champ

Le questionnaire doit, tout d'abord, couvrir effectivement le champ qui lui est assigné, et n'en

[19] Du nom de Tom Bradley, candidat afro-américain au poste de gouverneur de Californie qui perdit l'élection de 1982 alors qu'il était en tête dans les sondages.

manquer aucun aspect essentiel. Il importe donc de vérifier que les domaines (sections thématiques) et items (questions) du questionnaire couvrent bien tous les aspects et dimensions de nature à permettre au chercheur de collecter l'information qu'il vise au travers du questionnaire. Pour traiter ce premier volet de la validité, le principe est de soumettre le questionnaire à des experts (chercheurs et praticiens) familiers du champ, afin de recueillir leur avis sur le point de savoir si toutes les principales thématiques et dimensions sont couvertes. Il est de bonne méthode de disposer d'un critère objectif sur la base duquel on peut décider que les experts ont ou non entériné le questionnaire qui leur a été soumis, et un indicateur reconnu à cette fin est l'*alpha de Krippendorff*[20]. On

[20] Supposons un questionnaire soumis à des experts afin d'obtenir leur avis sur le caractère satisfaisant (ou non) de ses différentes questions. Chaque expert attribue à chaque question du questionnaire une note, par exemple de 1 à 5. Les notes sont répertoriées dans un *tableau de notation* du type suivant :

	Premier Expert	Deuxième expert	…	Dernier expert
Première question				
Deuxième question				
⋮	⋮	⋮	⋮	⋮
Dernière question				

Alternativement ou complémentairement, on peut plutôt/aussi demander aux experts un avis sur tel ou tel aspect du questionnaire, par exemple :

- Donnez une note de 1 à 5 suivant que le questionnaire reflète ou non la réalité du champ ; ou bien
- Donnez une note de 1 à 5 suivant que le questionnaire est pertinent ; etc.

Ces questions (une par ligne) figureront, avec les items du questionnaire proprement dits, ou à leur place, dans le tableau de notation. On peut faire noter plutôt/aussi les domaines du questionnaire. L'échelle de notation de 1 à 5 elle-même ne s'impose nullement : on peut adopter des codes de catégories (par exemple « le questionnaire vous parait original – 1 ; ou plutôt classique – 2), ou bien une échelle ordinale (par exemple de type Likert), ou encore une échelle quantitative (par exemple de 1 à 100).

L'alpha de Krippendorff mesure le degré de consensus entre experts. Il est calculé par la formule :

$$\alpha = \frac{p_a - p_e}{1 - p_e}$$

ajuste ensuite le questionnaire à la lumière des avis des experts. Il ne s'agit pas nécessairement de suivre à la lettre chacun de ces avis, mais au moins de pouvoir clarifier et argumenter les choix méthodologiques que l'on effectue.

---

où p_a est la proportion des experts qui convergent dans leurs appréciations, et p_e la convergence simplement due au hasard. La valeur de l'alpha s'établit entre 0 et 1. En règle générale, la convergence entre experts est considérée comme suffisante si l'alpha est au moins égal à 0,8.

Un outil commode pour calculer l'alpha de Krippendorff est la fonction KALPHA du « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9). Par exemple, pour un tableau de notation figurant dans la plage A1:E25 d'une feuille de calcul Excel, et sachant que les experts devaient distribuer des notes allant de 1 à 7, on inscrit dans la plage F1:F7 la liste des notes (1, 2, …, 7), et dans la cellule G1 la formule

= KALPHA(KTRANS(A1:E25), « type des notes », F1:F7)

où « type des notes » doit être remplacé par

- 0 si les notes sont de type catégoriel ;
- 1 si les notes sont de type ordinal ;
- 2 si les notes sont de type quantitatif ;
- 3 si les notes sont des pourcentages.

Puis on appuie sur la touche Entrée. La valeur de l'alpha s'inscrit en G1.

2.1.1.2.2. Fiabilité

Le questionnaire est fiable s'il est stable, c'est-à-dire si les réponses ne dépendent ni de l'enquêteur, ni du moment auquel le questionnaire est administré. Il y aurait évidemment problème si les réponses devaient dépendre de qui est l'enquêteur, par exemple si un questionnaire d'évaluation des enseignements par les étudiants devait engendrer des réponses différentes suivant que l'enquêteur est ou non l'enseignant lui-même. De même, un questionnaire n'est pas fiable si les réponses diffèrent suivant le moment auquel il est administré, par exemple, dans le cas précédent, juste avant ou juste après l'examen terminal.

Pour déceler un problème de fiabilité, on peut administrer le questionnaire à deux groupes distincts de testeurs, suivant la méthode du *split-half*. Chaque groupe doit être d'effectif suffisant (au moins trente participants[21]). Les groupes doivent être définis soit comme différents, soit comme identiques. Des critères de différence ou similarité doivent donc être préalablement établis, par exemple l'âge, la profession, le genre, le niveau d'études, les caractéristiques psychométriques, etc. On compare ensuite les réponses

[21] Pour que des tests statistiques standards puissent être appliqués.

des groupes à l'aune de critères statistiques (moyennes, médianes, variances, proportions) au moyen de tests d'hypothèses[22]. Le questionnaire est alors considéré comme fiable si les réponses ne sont pas significativement différentes d'un groupe à l'autre. Si les groupes ont été composés de façon à être différents, l'écart initial entre eux est connu, et il en est tenu compte lorsqu'on évalue l'éventuelle différence dans les réponses aux questionnaires.

Une autre approche consiste à administrer le questionnaire à un même groupe de testeurs à au moins deux reprises (« test-retest »). On peut faire varier l'enquêteur pour vérifier la fiabilité-enquêteur. Il y a fiabilité si les réponses lors du test et du retest sont suffisamment identiques (par exemple à 85%). Plus la durée entre les deux passations du questionnaire est élevée, plus le risque d'attrition augmente, donc l'effectif initial du groupe doit être suffisamment élevé pour que l'effectif lors du retest soit encore suffisant (au moins trente participants).

Cependant on ne peut exclure que les testeurs eux-mêmes aient évolué d'une passation à l'autre, par exemple par effet d'apprentissage, de sorte que l'écart de réponses entre test et retest peut être important même

[22] Voir chapitres 6 à 12.

si en réalité le questionnaire est fiable. Une parade possible est l'usage de questionnaires « parallèles » (*parallel forms*[23]), c'est-à-dire de versions alternatives d'un questionnaire identique sur le fond mais dont la présentation et la formulation comportent des différences à la fois non-susceptibles d'influencer le sens des réponses et suffisantes pour neutraliser l'effet d'apprentissage.

Les différences entre questionnaires initial et parallèle doivent être de nature strictement formelle, touchant éventuellement à la l'introduction/présentation, mais surtout aux questions : vocabulaire utilisé (par exemple langue courante dans le questionnaire initial et langage technique dans le questionnaire parallèle), style (plutôt parlé/familier ou plutôt littéraire/formel), ton (plutôt direct ou plutôt distancié), canal d'administration (par exemple d'abord électronique puis ensuite en répondant oralement à un enquêteur en face à face), ordre des questions, nombre de questions (par exemple une question dans le questionnaire initial est décomposée en deux questions dans le questionnaire parallèle), échelle d'évaluation (par exemple de 0 à 20 puis de 0 à 100, ou encore en faisant passer une échelle de Likert de 5 à 7 modalités), etc. Plus les sources de différenciation sont nombreuses, plus le risque d'effet d'apprentissage est réduit.

[23] Voir par exemple Henchy (2013).

Mais les différences entre questionnaire initial et parallèle ne doivent pas être de nature à modifier le contenu / sens des réponses. Par exemple, il convient de s'assurer que les synonymes employés sont compris comme tels par les répondants. De même, passer d'une échelle de Likert à choix impair à une échelle à choix forcé peut affecter la signification. L'effet possible des changements entre questionnaire initial et questionnaire parallèle doit donc être mûrement réfléchi.

Le questionnaire parallèle peut être considéré comme équivalent au questionnaire initial s'il engendre, pour chaque question ou modalité de réponse, les mêmes résultats auprès du même groupe de testeurs. Par exemple, le pourcentage de répondants choisissant telle modalité de réponse à telle question est le même avec les deux questionnaires. Ou encore, si la réponse attendue est de nature quantitative, la moyenne des réponses est similaire dans les deux questionnaires. On ne peut évidemment s'attendre à une similarité parfaite, de sorte qu'un intervalle d'écart admissible doit être défini. Il n'existe pas de pratique consensuelle sur l'écart qui peut être considéré comme admissible, donc la décision sur ce point est de la responsabilité du chercheur. Plusieurs méthodes de définition de l'écart admissible sont envisageables. Par exemple, on peut considérer que le pourcentage de répondants

choisissant telle modalité de réponse à telle question ne doit pas varier de plus de 5 points (ou 10 points, ou 15 points par exemple) d'un questionnaire à l'autre. Si les réponses sont quantitatives, on peut considérer que le coefficient de corrélation[24] entre les réponses à telle question du questionnaire initial et les réponses à la question équivalente du questionnaire parallèle doit être au moins de 85% (ou 90%, ou 95%, par exemple) ; ou encore que la différence de moyennes entre les deux questionnaires pour cette question ne doit pas être statistiquement significative, ou qu'elle doit être inférieure à un seuil[25]. Mais quelle que soit la méthode choisie, le choix doit être argumenté, par exemple par référence à la littérature, ou à une théorie, ou aux spécificités de l'objet de recherche, etc. La règle est d'éviter l'arbitraire.

2.1.1.2.3. Cohérence interne

Le questionnaire est cohérent si les questions regroupées au sein d'un même domaine portent effectivement sur ce domaine. Si le questionnaire est cohérent, il n'y a dans chaque domaine que des questions relatives à ce domaine. Une procédure

[24] Voir chapitre 13.

[25] Voir en particulier les méthodes de test présentées aux chapitres 8 et 12.

courante pour vérifier la cohérence interne de questionnaires consiste à calculer le coefficient *alpha de Cronbach*[26] pour les différents domaines du questionnaire. Dans la pratique, il est d'usage de considérer le domaine comme cohérent si l'alpha de Cronbach est supérieur à 0,7. À défaut, il importe d'identifier les raisons du niveau insuffisant de cohérence des domaines concernés et de remédier à cette insuffisance, par exemple en reformulant, ou en remplaçant, ou en retirant les items à l'origine du problème.

Par exemple, imaginons une recherche sur le climat scolaire qui, dans son questionnaire, considère le domaine « bien-être de l'élève à l'école », entre autres.

---

[26] Le coefficient alpha (α) de Cronbach est défini par :

$$\alpha = \frac{I \times \bar{r}}{1 + [(I-1) \times \bar{r}]}$$

où

- I est le nombre d'items composant le domaine ; et
- $\bar{r}$ est la moyenne des intercorrélations entre réponses à ces items (intercorrélations calculées en utilisant le coefficient de corrélation linéaire de Pearson : voir chapitre 13).

En principe, la valeur du coefficient alpha de Cronbach est comprise entre 0 et 1 (en fait on peut parfois observer des valeurs négatives ou supérieures à 1, par exemple lorsque le nombre d'observations est trop faible). L'homogénéité du domaine est d'autant plus forte que le coefficient alpha est proche de 1.

Le domaine comprend trois items. Le répondant note chaque item sur une échelle de 1-« très mauvais » à 5-« très bien ». Les réponses obtenues auprès de quarante répondants s'établissent comme suit (Tableau 2.1) :

Tableau 2.1.
Scores de 40 répondants à trois items du domaine « bien-être de l'élève à l'école »

Identifiant du répondant	Conditions matérielles de travail	Relations avec les autres élèves	Sentiment d'être bien encadré
321	2	4	2
322	3	4	3
323	1	4	2
324	3	5	4
325	5	4	4
326	3	2	3
327	4	3	4
328	2	3	5
329	1	2	4
330	5	5	1
331	5	5	1
332	4	1	1
333	3	5	2
334	3	4	4
335	2	3	1
336	5	4	1
337	4	4	3

Identifiant du répondant	Conditions matérielles de travail	Relations avec les autres élèves	Sentiment d'être bien encadré
338	3	2	3
339	4	5	2
340	4	2	5
341	2	2	5
342	1	4	5
343	2	1	2
344	5	5	1
345	3	3	1
346	3	4	5
347	1	2	1
348	3	1	2
349	2	5	3
350	1	5	2
351	5	4	2
352	3	1	1
353	2	2	4
354	4	5	4
355	3	2	1
356	5	4	2
357	2	5	5
358	3	5	2
359	4	1	2
360	4	1	5

La plupart des logiciels fournissent directement la valeur de l'alpha de Cronbach[27]. On la calcule ici en trois étapes[28] : $\alpha = -0{,}07$. On en conclut que la

[27] Dans XLSTAT, sous l'onglet *Description des données*, soit par la commande *Analyse de fiabilité*, soit parmi les sorties de la commande *Matrices de similarité / dissimilarité* avec, sous l'onglet *Général*, les spécifications *Similarités* et *Coefficient de corrélation de Pearson*.

[28] - Étape 1 : Calcul des intercorrélations entre items (voir le calcul du coefficient de corrélation de Pearson pages 319 et suivantes)

Variables	Conditions matérielles de travail	Relations avec les autres élèves	Sentiment d'être bien encadré
Conditions matérielles de travail	1		
Relations avec les autres élèves	0,161	1	
Sentiment d'être bien encadré	-0,220	-0,009	1

- Étape 2 : Moyenne des intercorrélations entre items

$$\bar{r} = \frac{0{,}161 - 0{,}220 - 0{,}009}{3} = -0{,}022$$

- Étape 3 : Calcul du coefficient alpha

$$\alpha = \frac{3 \times (-0{,}022)}{1 + [(3-1) \times (-0{,}022)]} = -0{,}07$$

cohérence du domaine n'est pas établie. Les items qui le composent n'appartiennent pas à une même dimension « bien-être de l'élève à l'école ; ou ne sont pas perçus comme tels par ces 40 répondants. Il est aussi possible que les répondants aient interprété et compris les questions de façons très différentes. Donc la cohérence – et par conséquent la validité – de cette partie du questionnaire sont douteuses. Mais on ne peut exclure non plus que le problème provienne du fait que les perceptions des répondants vis-à-vis de la situation de terrain au regard de ces trois items soient tout simplement extrêmement hétérogènes.

2.1.1.2.4. Intelligibilité

Le questionnaire doit être compris par le public auquel il s'adresse. Il doit donc, d'abord, être lisible pour ce public. Il existe de nombreux indicateurs de lisibilité adaptés aux textes en langue française[29], utilisés par exemple dans le monde de l'édition, mais les outils permettant aux chercheurs de les mettre en œuvre de façon rapide et simple sont peu nombreux. On peut citer néanmoins la version de l'*indice de Flesch* pour textes

[29] Voir par exemple en liste des références la revue effectuée par Noëlle Sorin (1996).

en langue française[30] développée en 2010 par Ménoni *et al.*[31]. Basé sur l'hypothèse que l'usage de mots courts et phrases courtes favorise la lisibilité, l'indice de Flesch mesure la lisibilité en fonction du nombre moyen de mots par phrase et du nombre moyen de syllabes par mot suivant la formule :

$$\begin{aligned} &Indice\ de\ Flesch \\ &= 206{,}835 \\ &- (1{,}015 \times Nombre\ moyen\ de\ mots\ par\ phrase) \\ &- (84{,}6 \times Nombre\ moyen\ de\ syllabes\ par\ mot) \end{aligned}$$

L'indice peut prendre une valeur entre 0 (texte très complexe) et 100 (texte très simple). L'indice permet de vérifier si la lisibilité du texte est adaptée au public auquel il est destiné[32], et d'effectuer les adaptations nécessaires sinon.

---

[30] Accessible à : https://www.rechercheclinique pariscentre.fr/?page_id=3169

[31] Voir liste des références.

[32] La version du test de Flesch pour textes en langue française proposée par Ménoni *et al* (2010) établit la correspondance suivante entre valeur de l'Indice de Flesch, caractéristiques de lisibilité, et niveau scolaire minimal du public cible :

L'intelligibilité peut être vérifiée plus avant en administrant le questionnaire à un petit groupe (une vingtaine) d'individus appartenant à la population-cible, mais qui – afin d'éviter tout biais – ne feront pas

Indice de Flesch	Nb de mots / phrases	Nb de syllabes / mots	Niveau stylistique	Niveau scolaire
0 à 30	29 et plus	1,92 et plus	Très complexe	Universitaire
30 à 50	25	1,67	Complexe	1er cycle universitaire
50 à 60	21	1,55	Assez complexe	Lycée
60 à 70	17	1,47	Standard	Quatrième / troisième
70 à 80	14	1,39	Assez simple	Cinquième
80 à 90	11	1,31	Simple	Sixième
90 à 100	8 ou moins	1,23 ou moins	Très simple	Cours moyen

partie de l'échantillon qui sera finalement interrogé. L'objectif de la démarche est de vérifier que les questions et consignes sont comprises et traitées de la façon souhaitée, et de déceler d'éventuels problèmes, par exemple que telle ou telle formulation n'est pas claire et que les répondants ne la comprennent pas ; ou qu'une formulation est trompeuse et oriente le répondant vers une fausse piste ; etc. Les correctifs nécessaires peuvent alors être apportés.

2.1.2. Échantillonnage

Une fois au point, le questionnaire peut être administré, mais encore faut-il déterminer à qui. Il ne s'agit plus à ce stade de définir la population d'intérêt, mais de déterminer lesquels de ses membres interroger. Parfois, la population d'intérêt est de taille limitée. C'est le cas par exemple dans une recherche sur les facteurs locaux de l'échec au bac dans une petite ville de quelques milliers d'habitants. Lorsque l'effectif de la population d'intérêt se limite à quelques dizaines ou centaines d'individus, on peut envisager d'administrer le questionnaire à l'ensemble de ces individus. Souvent cependant, l'effectif de la population d'intérêt peut être très élevé, atteignant par exemple plusieurs centaines de milliers d'individus. Dans ce cas, administrer le questionnaire à l'ensemble de ces individus peut soulever de redoutables problèmes de moyens (notamment trouver et financer des enquêteurs) et des

problèmes d'ordre technique (en particulier la gestion et le traitement des données) de nature à rendre l'ensemble des processus de collecte et de traitement lourds, complexes, couteux et longs. La solution alternative consiste alors à travailler sur échantillon. Mais dans ce cas, l'objectif restant bien d'aboutir à des conclusions généralisables, il faut éviter l'échantillonnage sauvage, et au contraire procéder avec méthode. Deux principaux points méritent alors attention : la composition de l'échantillon, et sa taille.

2.1.2.1. Composition de l'échantillon

On distingue deux approches en matière de composition de l'échantillon : l'échantillonnage non-probabiliste et l'échantillonnage probabiliste.

2.1.2.1.1. Échantillonnage non-probabiliste

L'échantillonnage est non-probabiliste s'il est conduit sans prendre en compte la probabilité qu'a chaque membre de la population-mère d'être sélectionné pour faire partie de l'échantillon (« probabilité de sélection », on dit encore « probabilité d'inclusion »). Une méthode connue d'échantillonnage non-probabiliste est le micro-trottoir, qui consiste à se poster à un endroit et à interroger tout ou partie des passants. Il est clair que toute la population n'avait pas la même probabilité de passer à cet endroit à ce moment, donc

que les probabilités de sélection des individus ne sont pas égales, (par exemple la probabilité de sélection des riverains est plus élevée), sans que l'on en sache plus sur le niveau exact de ces probabilités.

La méthode la plus courante d'échantillonnage non-probabiliste est la *méthode des quotas*. Le principe en est de constituer un échantillon tel que chaque composante de la population-mère soit représentée au sein de l'échantillon dans les mêmes proportions que dans la population-mère. Cette méthode repose sur le postulat que le phénomène que l'on cherche à étudier (par exemple l'absentéisme scolaire, ou encore tel ou tel style d'apprentissage, ou quoi que ce soit d'autre), même si on en ignore l'ampleur au sein de la population-mère, existera dans l'échantillon de la même façon qu'il existe au sein de la population-mère, dès lors que la structure connue de la population-mère est respectée. En d'autres termes, on suppose qu'il existe une certaine corrélation entre toutes les caractéristiques possibles d'une même population.

D'un point de vue théorique, la limite de la méthode des quotas est que son postulat fondateur reste à démontrer. D'un point de vue pratique, une de ses limites tient au fait qu'il n'existe pas de règle quant aux caractéristiques à prendre en compte pour définir la structure d'une population. Les caractéristiques possibles sont innombrables : dans l'absolu, elles

peuvent être démographiques, sociales, éducationnelles, professionnelles, culturelles, religieuses, politiques, économiques, géographiques, par exemple. Sélectionner certaines caractéristiques et pas d'autres est donc toujours discutable, et doit par conséquent être argumenté. Cependant, en admettant même que toutes les caractéristiques possibles puissent être prises en compte, cela impliquerait de prendre en compte de micro-composantes de la population, dont l'effectif est restreint à quelques unités ou dizaines d'unités. Il faudrait alors que ces micro-composantes soient représentées dans l'échantillon dans la même proportion que dans la population-mère. Par suite, il faudrait inclure dans l'échantillon suffisamment d'individus appartenant aux composantes numériquement les plus importantes de la population-mère pour que ces composantes importantes soient représentées à leur exacte proportion dans l'échantillon. Il en résulterait que l'effectif de l'échantillon risquerait de n'être plus très éloigné de celui de la population-mère, ce qui à l'évidence annulerait l'intérêt même de recourir à un échantillon.

Pour pallier cette limite, une pratique courante consiste à s'en tenir à seulement quelques caractéristiques (appelées « variables de contrôle ») comme par exemple le genre, l'âge, et le statut socioprofessionnel. Cela étant, le fait même de ne plus prendre en compte qu'un nombre restreint de ces variables de contrôle

souligne davantage encore la fragilité du postulat fondateur de la méthode des quotas.

En tout état de cause, en pratique, la représentativité de l'échantillon dans la méthode des quotas doit s'entendre en termes d'un nombre limité de caractéristiques de la population-mère, et non comme une représentativité absolue au regard de toutes ses caractéristiques possibles.

Il appartient au chercheur de recruter comme membres de l'échantillon, des individus qui correspondent aux caractéristiques retenues comme variables de contrôle. Le recrutement de répondants est une difficulté récurrente des enquêtes car, en dehors du cadre du recensement général de la population, il n'y a pas d'obligation réglementaire de participation s'imposant aux répondants potentiels. Une pratique courante est donc le *recrutement de commodité* (ou *recrutement de convenance*), qui consiste à recruter comme répondants des personnes plus aisément accessibles (par exemple parmi les proches), ce qui exclut les personnes moins accessibles, et introduit par conséquent un risque supplémentaire de biais. Une autre pratique tout aussi courante est le *recrutement volontaire*, qui consiste à accepter comme répondants des volontaires autorecrutés, mus par leur propre intérêt pour l'objet spécifique de l'enquête en question. Il est évident que les personnes intéressées par l'enquête sont différentes

de celles qui ne le sont pas, et que la relative absence de ces dernières dans l'échantillon est aussi potentiellement source de biais. En outre, les risques de manipulation (et donc à nouveau de biais), par des groupes de répondants qui voudraient influencer les résultats d'une enquête à laquelle ils s'intéresseraient tout particulièrement, ne sont pas nuls (réponses convergentes concertées, pluri-participation des mêmes individus à l'enquête). Le chercheur devra prêter une attention particulière à scruter les réponses obtenues afin de déceler et d'écarter des réponses ou tendances suspectes.

Quelle que soit la procédure adoptée pour le recrutement de répondants, le chercheur devra veiller à ce que les proportions requises pour chaque variable de contrôle soient respectées. Lorsque ces proportions ne sont pas respectées, un *redressement* doit être effectué. Le redressement est effectué soit par suppression soit par pondération. La suppression consiste à retirer de l'échantillon des individus appartenant aux composantes surreprésentées. La pondération consiste à attribuer à chaque répondant un coefficient tel que la structure de l'échantillon redressé (les individus affectés de leurs coefficients) soit identique à celle de la population-mère. Les logiciels d'enquête et les

logiciels statistiques permettent d'automatiser les opérations de redressement[33].

2.1.2.1.2. Échantillonnage probabiliste

La forme canonique de l'échantillonnage probabiliste est l'*échantillonnage aléatoire simple*. Le principe de l'échantillonnage aléatoire simple est que les personnes auxquelles le questionnaire sera soumis sont sélectionnées au hasard, de sorte que toutes ces personnes ont chacune la même probabilité n/P d'être sélectionnées pour faire partie de l'échantillon (n étant la taille de l'échantillon et P l'effectif de la population-mère). La méthode nécessite de disposer de la liste complète des individus appartenant à la population-mère (« base de sondage »), et d'attribuer à chacun de ces individus un numéro d'ordre. On procède ensuite au tirage au sort en utilisant un générateur de nombres aléatoires[34].

---

[33] Sous XLSTAT, les redressements s'effectuent à partir de la commande *Redressement d'enquêtes* sous l'onglet *Préparation des données*.

[34] Par exemple la fonction « Alea entre bornes » d'Excel. Par exemple, si la population-mère comporte 3 millions d'individus (à chacun desquels est assigné un numéro d'ordre) et si la taille de l'échantillon est fixée à 2000, la formule « =ALEA.ENTRE.BORNES(1,3000000) mise en œuvre pour 2000 cellules de la feuille de calcul fournira les 2000 numéros

Plusieurs variantes de l'échantillonnage probabiliste existent. L'échantillonnage aléatoire « systématique » (ou « échantillonnage par intervalles »), d'abord, dans lequel le premier membre de l'échantillon est tiré au sort, puis chaque n-ième[35] membre de la base de sondage à partir du premier sélectionné est inclus dans l'échantillon jusqu'à ce que celui-ci soit au complet.

Dans l'échantillonnage probabiliste « stratifié », on définit une « partition » de la population-mère, c'est-à-dire qu'on décompose celle-ci en sous-groupes disjoints (« strates »), mutuellement exclusifs : les membres d'une strate n'appartiendront qu'à cette strate, et l'ensemble des strates constitue la population-mère. En général, les strates sont choisies de façon à constituer des groupes homogènes au regard de la

---

aléatoires désignant les 2000 individus de la base de sondage à inclure dans l'échantillon. En réalité cependant, il faut générer plus de numéros aléatoires qu'il n'y a d'individus dans l'échantillon, par exemple ici 2100, car les numéros générés peuvent comporter des doublons.

[35] Par exemple chaque dixième membre, ou chaque quinzième membre, ou chaque vingtième membre, etc. Par exemple, si la population-mère compte 100 individus et si la taille de l'échantillon est fixée à 10, on sélectionne aléatoirement le premier membre de l'échantillon, puis le 10ème après lui, puis le 20ème après lui, puis le 30ème après lui, etc., jusqu'à ce que l'effectif de l'échantillon soit au complet.

variable d'intérêt. Par exemple, les élèves répartis par catégorie socioprofessionnelle du chef de famille peuvent être des strates lorsque l'objet auquel on s'intéresse est l'origine sociale des élèves ; ou encore, autre exemple, les sortants du système éducatif répartis par niveau d'études peuvent être des strates lorsqu'on analyse les effets du niveau d'études sur l'insertion professionnelle. Une fois les strates définies, on applique le tirage au sort au sein de chaque strate pour désigner les individus de la strate qui feront partie de l'échantillon[36].

[36] Il importe de ne pas confondre les strates de l'échantillonnage stratifié et les quotas de la méthode des quotas. Dans la méthode des quotas, un même individu peut faire partie de plusieurs quotas. Par exemple, si le genre, l'âge et la catégorie socioprofessionnelle sont les variables de contrôle, un même individu, par exemple Monsieur Bernard, médecin de 45 ans, fera partie du quota Hommes, mais aussi du quota des 34-55 ans, et également du quota des professions libérales. Dans l'échantillonnage stratifié, les strates sont des segments mutuellement exclusifs d'une seule variable, par exemple l'âge. Si Monsieur Bernard est sélectionné au titre de la strate des 34-55 ans, il ne peut pas être sélectionné au titre d'une autre strate. Les critères de sélection sont multidimensionnels dans la méthode des quotas, et unidimensionnels dans l'échantillonnage stratifié. Au-delà, les différences entre méthode des quotas et échantillonnage stratifié sont que la méthode des quotas ne nécessite ni base de sondage ni sélection aléatoire des répondants.

Dans l'absolu, bien d'autres variantes des méthodes probabilistes existent[37], comme par exemple le sondage en grappes[38], le sondage à plusieurs degrés[39], ou le sondage à probabilités inégales[40]. Ces variantes

---

[37] Pour des illustrations, voir par exemple :

- Statistique Canada
https://www.statcan.gc.ca/edu/power-pouvoir/ch13/prob/5214899-fra.htm; ou

- le CEGO – Centre Canadien d'expertise des grands organismes
http://grandsorganismes.gouv.qc.ca/outils/guides-pratiques-mesure-de-la-satisfaction-de-la-clientele/lechantillionnage/.

[38] La population-mère est d'abord décomposée en une liste de grappes, une grappe étant un ensemble d'unités statistiques / individus. Un même individu ne peut appartenir qu'à une seule grappe. Les grappes sont tirées au sort, puis *tous* les individus membres des grappes désignées par le tirage au sort sont incorporés à l'échantillon. Dans l'échantillonnage en grappes, il n'est pas nécessaire que la base de sondage comprenne la liste des individus membres de la population-mère car il n'y a pas de tirage au sort de ces individus. Le tirage au sort ne porte que sur les grappes, donc ce qui est indispensable, c'est la liste des grappes.

[39] La population-mère est décomposée en grappes, puis certaines grappes sont tirées au sort, puis une liste des individus appartenant à ces grappes est établie et *un certain nombre* d'individus de ces grappes sont tirés au sort pour composer l'échantillon.

[40] Certains individus ou types d'individus de la population-mère, dont le chercheur considère qu'il est particulièrement important de les inclure dans l'échantillon, se voient attribuer une

répondent généralement aux besoins de cas de figure spécifiques et/ou à des contraintes de coût.

Les méthodes probabilistes ne visent pas à ce que les composantes de la population-mère soient représentées dans l'échantillon. Elles cherchent à rendre l'échantillon non pas représentatif mais significatif, c'est-à-dire de nature telle que les résultats d'analyse obtenus sur l'échantillon puissent être extrapolés à la population-mère, avec un degré de précision connu. On peut ainsi, par exemple, estimer la moyenne d'une population-mère, ou l'ampleur relative (%) d'un phénomène dans une population-mère, à partir de la moyenne ou de l'ampleur relative du phénomène dans l'échantillon (voir Encadré 2.1).

ENCADRÉ 2.1 – ESTIMER LA MOYENNE OU UNE PROPORTION DANS LA POPULATION-MÈRE À PARTIR DE LA MOYENNE OU DE LA PROPORTION DANS L'ÉCHANTILLON

Estimation de la moyenne

Dans l'échantillonnage probabiliste, les membres de l'échantillon sont sélectionnés de façon

---

probabilité plus élevée que les autres de faire partie de l'échantillon.

aléatoire. Or le théorème central limite implique que lorsqu'un grand nombre d'échantillons aléatoires sont tirés d'une même population, les moyennes de ces échantillons sont distribuées normalement autour de la moyenne de la population. On sait par ailleurs (voir chapitre 1) que dans une distribution normale :

- 68% des valeurs de la série sont comprises dans l'intervalle $[\overline{x} - \sigma\ ;\ \overline{x} + \sigma]$ où $\overline{x}$ est la moyenne de la série et σ son écart-type ;
- 95% des valeurs sont comprises dans l'intervalle $[\overline{x} - 2\sigma\ ;\ \overline{x} + 2\sigma]$; et
- 99,8% des valeurs sont comprises dans l'intervalle $[\overline{x} - 3\sigma\ ;\ \overline{x} + 3\sigma]$.

Par conséquent, si l'échantillon est aléatoire, il y a :

- 68% de chances que sa moyenne soit comprise dans l'intervalle $[\overline{x} - \sigma\ ;\ \overline{x} + \sigma]$ où $\overline{x}$ et σ sont respectivement la moyenne et l'écart-type de la population-mère ;
- 95% de chances que sa moyenne soit comprise dans l'intervalle $[\overline{x} - 2\sigma\ ;\ \overline{x} + 2\sigma]$; et
- 99,8% de chances que sa moyenne soit comprise dans l'intervalle $[\overline{x} - 3\sigma\ ;\ \overline{x} + 3\sigma]$.

On peut donc estimer la moyenne de la population-mère à partir de la moyenne de l'échantillon en se donnant une certaine marge d'erreur. La marge d'erreur est définie par :

$$marge\ d'erreur = coefficient\ de\ confiance \times \frac{écart\ type\ de\ l'échantillon}{\sqrt{taille\ de\ l'échantillon}}$$

Le *coefficient de confiance* lui-même se calcule à partir du « *degré de confiance* » que le chercheur fixe pour son estimation :

$$coefficient\ de\ confiance = Z_{\alpha/2}$$

où α est le degré de confiance.

Pour un degré de confiance α donné, le coefficient de confiance $Z_{\alpha/2}$ est fourni par la table de Z (voir Annexe 1, page 721) : on divise α par 2 et on cherche le résultat dans la table. Le coefficient de confiance est la valeur de Z en tête de ligne, suivie par les décimales en intitulé de la colonne à laquelle appartient $\frac{\alpha}{2}$. Par exemple, si le chercheur choisit un degré de confiance de 80%,

alors $\frac{\alpha}{2} = 0{,}40$. Dans la table de Z, 0,40 correspond à la ligne 1,2 et à la colonne 0,09, donc le coefficient de confiance est 1,29.

En pratique, les degrés de confiance les plus souvent utilisés sont :

Degré de confiance (α)	Coefficient de confiance ($Z_{\alpha/2}$)
99%	2,576
98%	2,326
95%	1,960
90%	1,645

À titre d'illustration, imaginons par exemple que l'on cherche à mesurer le sentiment d'exposition au harcèlement ressenti par les étudiantes dans l'enseignement supérieur. On constitue un échantillon aléatoire de 150 étudiantes auquel on administre un questionnaire permettant de mesurer le niveau de sentiment perçu. Les résultats indiquent que le niveau de sentiment perçu s'établit à 27 en moyenne sur une échelle de 1 à 100, avec un écart-type de 6,43. On peut donc estimer, avec un degré de confiance de 99%, que le sentiment d'exposition au harcèlement dans la

population-mère s'établit à 27 en moyenne sur une échelle de 1 à 100, avec une marge d'erreur de $\pm 1{,}35$:

$$\begin{aligned} &marge\ d'erreur \\ &\quad = coefficient\ de\ confiance \\ &\quad \times \frac{écart\ type\ de\ l'échantillon}{\sqrt{taille\ de\ l'échantillon}} \\ &\quad = 2{,}576 \times \frac{6{,}43}{\sqrt{150}} = 1{,}35 \end{aligned}$$

En d'autres termes, selon cet exemple fictif, il y a 99% de chances pour que dans la population des étudiantes de l'enseignement supérieur, le niveau du sentiment d'exposition au harcèlement soit en moyenne compris entre 25,65 et 28,35 sur une échelle de 1 à 100.

L'intervalle

$$[moyenne - écart\ type\ ;\ moyenne + écart\ type]$$

($[25{,}65\ ;\ 28{,}35]$ dans cet exemple) est appelé « intervalle de confiance ».

Estimation d'une proportion

On peut de même estimer une proportion dans la population-mère à partir de cette proportion dans

l'échantillon. La marge d'erreur est donnée par :

$$marge\ d'erreur\ =\ Z_{\alpha/2}\ \times \sqrt{\frac{\rho \times (1-\rho)}{n}}$$

où

- ρ (rho) est la proportion dans l'échantillon ; et
- n la taille de l'échantillon.

Les méthodes probabilistes sont peu utilisées dans la recherche académique en éducation et formation en France, l'une des raisons étant qu'il n'existe généralement pas de base de sondage à partir desquelles effectuer le tirage au sort. Là encore, les moyens disponibles ne favorisent pas la constitution de telles bases, *a fortiori* pour les besoins des recherches doctorales.

2.1.2.2. Taille de l'échantillon

La question est de savoir quelle taille minimum l'échantillon doit avoir pour que l'on puisse considérer qu'il constitue un substitut opératoire de la population-

mère[41]. Par exemple, peut-on se contenter d'un échantillon de 10 individus pour représenter une population-mère en comprenant un million ? Ou bien la taille de l'échantillon doit-elle être proportionnelle à l'effectif de la population-mère ?

Il n'existe pas de méthode standard qui serait applicable à toute enquête pour calculer la bonne taille d'échantillon. Tout dépend de la méthode adoptée pour la composition de l'échantillon, et des objectifs de l'analyse.

Si la composition de l'échantillon s'effectue par méthode non-probabiliste, la décision relative à la taille de l'échantillon dépend essentiellement du chercheur. Dans la méthode des quotas, l'échantillon est composé sous forme d'une maquette de la population-mère que le chercheur juge appropriée à ses objectifs. Il lui revient alors de décider de l'échelle de sa maquette, par exemple au cent-millième (l'effectif de chaque composante de l'échantillon sera égal à l'effectif de cette composante dans la population-mère divisé par 100 000), ou au dix-millième, etc. La décision relative à l'échelle tient compte du volume de données que le chercheur souhaite traiter dans le cadre des moyens (équipe d'enquêteurs ; moyens de gestion et de

[41] La problématique de la taille d'échantillon en approche quantitative n'a donc rien à voir avec celle de la saturation dans laquelle s'inscrit souvent la recherche qualitative.

traitement des données) et du temps dont il dispose. Le chercheur peut aussi décider de se caler sur la taille d'échantillon habituellement utilisée dans la littérature de son domaine de recherche, quitte à l'adapter en tirant parti de l'expérience accumulée par lui-même et ses pairs[42].

Si l'échantillonnage s'effectue de façon aléatoire simple, la problématique est totalement différente. La taille de l'échantillon peut être calculée à partir de la formule de la marge d'erreur pour l'estimation d'une proportion, présentée en Encadré 2.1, page 97 :

$$e = Z_{\alpha/2} \times \sqrt{\frac{\rho \times (1-\rho)}{n}} \quad => \quad n = \frac{Z_{\alpha/2}^2 \times \rho \times (1-\rho)}{e^2}$$

où

- n est la taille de l'échantillon ;
- e est la marge d'erreur ;

[42] Mais il importe de conserver à l'esprit que les méthodes non-probabilistes n'ont pas de fondements statistiques qui permettraient d'extrapoler avec une précision connue les résultats de l'échantillon à la population-mère. En ce sens, ce sont davantage des outils qui fournissent une photographie de la situation dans l'échantillon à un instant donné, plutôt que des outils qui renseignent avec certitude et précision sur la population-mère, et encore moins des outils prédictifs.

- ρ (rho) est la proportion qui indique dans quelle mesure le phénomène d'intérêt est présent dans la population-mère. Par exemple, si la recherche porte sur l'illettrisme, ρ pourra être le taux d'illettrisme ; si la recherche porte sur le décrochage scolaire, ρ pourra être le taux de décrochage scolaire ; etc. Lorsque ρ est inconnue, on utilise $\rho = 0{,}5$ car l'expression $\rho \times (1 - \rho)$ représente la variabilité du phénomène, et plus un phénomène est variable, plus l'échantillon nécessaire pour l'étudier doit être grand. C'est lorsque $\rho = 0{,}5$ que l'expression $\rho \times (1 - \rho)$ prend sa valeur la plus élevée[43]. Donc on s'assure (par précaution) de la

[43] On vérifie aisément que :

Si ρ =	$(1 - \rho)$ =	Et $\rho \times (1 - \rho)$ =
0,1	0,9	0,09
0,2	0,8	0,16
0,3	0,7	0,21
0,4	0,6	0,24
0,5	0,5	0,25
0,6	0,4	0,24
0,7	0,3	0,21
0,8	0,2	0,16
0,9	0,1	0,09
1	0	0

taille d'échantillon la plus élevée en prenant $\rho = 0{,}5$.

Cependant, on ne peut exprimer la variabilité au moyen de proportions que si la variable d'intérêt est dichotomique, les individus pouvant alors être répartis entre ceux qui vérifient la propriété d'intérêt (ρ) et ceux qui ne la vérifient pas ($1 - \rho$). Parfois, la variable d'intérêt est polytomique (par exemple le niveau de diplôme), ou discrète (par exemple le nombre de candidats au bac par établissement, ou le salaire des enseignants, etc.). Dans ce cas, on ne peut pas représenter la variabilité par une proportion. On utilise alors, à la place, la variance du phénomène. La taille n de l'échantillon s'écrit alors :

$$n = \frac{Z_{\alpha/2}^2 \times \sigma^2}{e^2}$$

où σ^2 est la variance.

Par exemple, on veut connaitre l'ampleur ainsi que les formes et mécanismes particuliers du retard scolaire à l'école primaire dans une région frontalière. On opte pour un échantillonnage aléatoire simple permettant d'approcher la situation dans la population-mère avec une marge d'erreur maximale de 2% et un degré de

confiance de 99%. On s'interroge sur la bonne taille d'échantillon.

$$n = \frac{Z_{\alpha/2}^2 \times \rho \times (1-\rho)}{e^2} = \frac{2{,}576^2 \times 0{,}5 \times 0{,}5}{0{,}02^2} = 4147{,}36$$

La valeur de ρ a été fixée à 0,5 car l'ampleur du phénomène est inconnue et fait précisément partie des questions de recherche. L'échantillon devra donc comprendre 4148 élèves d'écoles primaires de la région.

Il importe d'observer que la taille de l'échantillon ne dépend pas de la taille de la population-mère. S'il apparait que la taille requise pour l'échantillon est du même ordre que la taille de la population-mère, plusieurs options sont possibles, notamment :

- reconsidérer la possibilité d'analyser l'ensemble de la population-mère ; ou bien
- relaxer les exigences de précision en adoptant un degré de confiance et une marge d'erreur plus faibles ; ou bien
- appliquer une minoration de taille d'échantillon pour cas de population-mère de faible taille. Par exemple, Yamane (1967)[44] propose de s'appuyer uniquement

[44] Voir liste des références.

sur la taille P de la population-mère et la marge d'erreur :

$$n^{-} = \frac{P}{1 + (P \times e^2)}$$

où n^{-} représente la taille minorée d'échantillon.

Dans l'exemple ci-dessus, à supposer que le nombre total d'élèves en école primaire dans la région frontalière considérée s'établisse à 10 000, la taille minorée d'échantillon serait de 2000 élèves selon la formule de Yamane.

Une autre correction possible (Israel 2012[45]) est de la forme :

$$n^{-} = \frac{n}{1 + \frac{n-1}{P}}$$

Ici, toujours en se référant à l'exemple précédent, la taille minorée d'échantillon s'établirait à 2932 élèves.

Si l'échantillon est stratifié, une méthode logique consiste à partir de la taille nécessaire pour un échantillon aléatoire simple, et à affecter à chaque strate la même part relative dans l'échantillon que dans la population-mère.

[45] Voir liste des références.

Dans l'échantillonnage en grappes, on peut estimer le nombre moyen d'individus par grappe, et en déduire le nombre de grappes à tirer au sort. Dans le sondage à deux degrés, on peut, sachant le nombre moyen de personnes par grappe, fixer le nombre de personnes à tirer au sort dans chaque grappe, et en déduire le nombre de grappes à tirer au sort. En principe, la taille de l'échantillon devrait en outre être ajustée en tenant compte de « l'effet de plan » (effet du plan de sondage) défini par le rapport entre variance de l'estimation en sondage par grappe ou par degré et variance de l'estimation en sondage aléatoire simple. En pratique cependant, l'estimation de l'effet de plan est complexe et nécessite l'appui de statisticiens[46].

Enfin, une fois connue la taille de l'échantillon, il reste à déterminer le nombre de répondants potentiels à contacter. En effet, ce n'est pas parce qu'un répondant potentiel est sollicité qu'il répondra effectivement : il peut y avoir des non-réponses. Du reste, même si toutes les personnes sollicitées répondaient, toutes les réponses ne seraient pas forcément exploitables : il peut y avoir des réponses incomplètes ou fantaisistes, par exemple. Donc le chercheur doit anticiper le taux de non-réponses et de réponses non exploitables, se donner une marge, et en déduire le nombre de

[46] Voir en liste des références la revue des méthodes d'estimation de l'effet de plan effectuée par Bem *et al*. (2008).

répondants potentiels à contacter. Si par exemple le chercheur est pessimiste (ou réaliste !) et pense que seuls 10% des personnes sollicitées répondront effectivement et de façon utile / exploitable, et si la taille n d'échantillon nécessaire a été estimée à 400, alors le nombre de répondants potentiels à contacter s'établit à 4000 :

$$\begin{aligned} &nombre\ de\ répondants\ potentiels\ \text{à}\ contacter \\ &= \frac{taille\ d'échantillon\ nécessaire\ (n)}{pourcentage\ de\ réponses\ exploitables} \\ &= \frac{400}{0{,}10} = 4000. \end{aligned}$$

2.2. L'ENQUÊTE PAR ENTRETIEN

La littérature relative aux méthodes d'entretien est riche. De façon générale, l'entretien se définit comme un échange verbal entre deux personnes, à l'initiative d'un chercheur et contrôlé par lui, en vue d'obtenir du répondant des informations pertinentes par rapport à un objet d'étude.

On en distingue deux cas polaires. Le premier est l'entretien libre / non-directif, dans lequel le chercheur adapte librement l'ordre et la formulation des questions (à partir toutefois d'un « guide d'entretien » préétabli), tandis que le répondant est maitre du format de ses

réponses. Dans le deuxième cas, l'entretien est structuré, directif et standardisé. Le chercheur (ou l'enquêteur qu'il a mandaté) ne s'écarte pas de l'ordre et de la formulation des questions fixés dans le guide d'entretien. Dans sa forme extrême, l'entretien standardisé ne laisse au répondant que la possibilité de choisir ses réponses sur une liste qui lui est proposée. L'entretien standardisé permet de comparer terme à terme les réponses de répondants différents. Entre ces deux cas polaires, l'entretien semi-structuré (« semi-directif ») impose l'ordre et la formulation des questions, mais prévoit des questions ouvertes, qui laissent au répondant une certaine marge dans le format de ses réponses.

Quel que soit le type d'entretien qu'il envisage, le chercheur doit commencer par établir un guide d'entretien comprenant une présentation de la recherche (et du chercheur lui-même) destinée à expliquer au répondant ce dont il s'agit ; et une liste des points sur lesquels le répondant devra être amené à s'exprimer. En principe, l'ordre des questions donne la priorité à celles auxquelles il est agréable et facile au répondant de répondre, de façon à amorcer un échange fluide et à créer un contexte propice à développer plus avant l'entretien. Les questions plus difficiles, qui pourraient entrainer tensions et blocages, viennent seulement en fin d'entretien, lorsqu'un maximum de réponses ont déjà été obtenues. La formulation des

questions doit être sans ambiguïté, précise, et éviter les connotations non désirées. L'entretien ne doit pas être trop long, afin d'éviter l'effet de lassitude.

Dans la perspective quantitative toutefois, l'entretien occupe une place marginale. La raison en est que, sauf à disposer de moyens d'enquête considérables, l'enquête par entretien ne peut pas viser un nombre très élevé de répondants. Dans une recherche académique (et *a fortiori* dans une recherche doctorale) en éducation et formation, effectuer une vingtaine d'entretiens est déjà une très belle performance. Mais un tel volume est bien inférieur au seuil nécessaire pour pouvoir appliquer des traitements statistiques standards, par exemple des tests d'hypothèses. En conséquence, dans la perspective d'une approche quantitative, l'entretien comme mode de collecte de données n'est guère pertinent que si l'analyse des données est envisagée par voie lexicométrique[47].

[47] Voir chapitre 4.

Chapitre 3. Préparer les données

Les données destinées à une analyse quantitative sont en principe regroupées dans une feuille de calcul de tableur (par exemple Excel). Lorsque le questionnaire a été administré en ligne, les plateformes d'enquête permettent au chercheur de récupérer directement le fichier des données. Dans les autres cas, le répondant communique ses réponses oralement ou sur document papier à l'enquêteur / chercheur qui les saisit dans la feuille de calcul. Le principe général est d'organiser la feuille de calcul de façon à avoir une colonne par question ou par modalité de réponse, et une ligne par répondant. Les réponses sont saisies telles que fournies par le répondant lorsque la variable est quantitative ou textuelle courte (commentaire libre)[48]. Lorsque la variable est qualitative et les modalités de réponse possibles préétablies, il est généralement plus commode de représenter chaque modalité de réponse par un numéro de code[49].

---

[48] S'il s'agit de données textuelles longues (par exemple interviews, articles, discours, textes littéraires, etc.) à traiter par édumétrie, elles doivent plutôt être présentées dans un fichier de traitement de texte (par exemple Word).

[49] Par exemple 0 pour « Non », 1 pour « Oui », 90 pour « Non-concerné / Non-applicable », 99 pour « Non-réponse ». Autant que possible la signification d'un code devrait être intuitive et rester stable d'une question à l'autre, de façon à réduire les risques d'erreur lors de la saisie dans le fichier et les risques de

Une fois le fichier rempli, il doit être finalisé. La finalisation exige d'abord de scruter les données à la recherche d'erreurs à éliminer. Les sources de possible erreur sont multiples. Des erreurs peuvent intervenir aussi bien lorsque le répondant fournit ses réponses que lors de la saisie de ces dernières dans le fichier. Dans les questionnaires auto-administrés, certains répondants peuvent aussi parfois fournir des réponses fantaisistes.

Mais la finalisation renvoie aussi à deux autres points importants : le traitement des valeurs aberrantes et la standardisation des données.

3.1. VALEURS ABERRANTES

Même en l'absence d'erreurs, les données peuvent contenir des valeurs « aberrantes » (*outliers*), inhabituellement faibles ou élevées, et plus généralement non-représentatives en ce que leur profil tranche radicalement avec le ou les profils habituellement observés dans les données sur lesquelles

---

confusion lors de l'interprétation des résultats. Autant que possible, les codages devraient avoir été prévus lors de l'élaboration du questionnaire, de façon que l'ordre des questions coïncide avec le codage.

porte l'analyse. Elles peuvent être détectées à la lecture des données, ou en visualisant ces dernières au moyen de graphiques (par exemple nuages de points, histogrammes ou courbes[50]). Les logiciels statistiques proposent aussi souvent des tests de détection des valeurs aberrantes[51]. Les valeurs aberrantes peuvent sensiblement influencer la moyenne de la série ou le résultat de tel ou tel calcul, et engendrer une impression d'ensemble incorrecte. On peut vérifier l'incidence des valeurs aberrantes en effectuant les calculs avec puis sans elles, et comparer les résultats. Il importe que le chercheur s'interroge, en fonction de la nature spécifique de sa recherche, sur la nécessité de retirer ou non les valeurs aberrantes, et plus généralement sur la façon de les gérer.

3.2. STANDARDISATION

Un traitement préparatoire additionnel qui peut s'avérer nécessaire dans la phase de finalisation du fichier des données est la standardisation des données. La

[50] XLSTAT permet d'établir des nuages de points ou d'autres types de graphiques à partir de l'onglet *Visualisation des données*.

[51] Dans XLSTAT, la détection de valeurs aberrantes univariées peut s'effectuer notamment au moyen du test de Grubbs, accessible sous l'onglet *Tests pour les valeurs extrêmes*. Voir page 324 un exemple de détection graphique et par test de valeurs aberrantes univariées.

problématique ici est celle de données d'échelles différentes qui, par conséquent, ne peuvent être visualisées sur un même plan, et par ailleurs ne sont pas directement comparables. C'est par exemple le cas lorsque les séries considérées sont exprimées en unités de compte différentes (nombres de personnes, nombre d'incidents, unités de temps, salaire, etc.), ou sont d'ordres de grandeur disproportionnés, allant par exemple de 1 à 10 dans la première série et de 100 à 1000 dans la seconde.

Pour rendre ces données comparables, on procède habituellement à leur standardisation[52].

Il existe un grand nombre de méthodes de standardisation, adaptées à différents types de problèmes. On présente ici quelques formes de standardisation parmi les plus courantes : la standardisation par centration-réduction, les standardisations max-min, et la standardisation par moyenne ou écart-type.

---

[52] On utilise parfois le terme de "normalisation" comme synonyme de standardisation. Cet usage peut être source de confusion car la normalisation au sens strict vise à doter une distribution des propriétés de la loi normale. Ce n'est pas l'objectif des méthodes de standardisation présentées dans ce chapitre. Voir en liste des références la revue des méthodes de normalisation (au sens strict) effectuée par Daumas (1982).

3.2.1. Standardisation par centration - réduction

La standardisation par centration-réduction consiste à transformer chaque série :

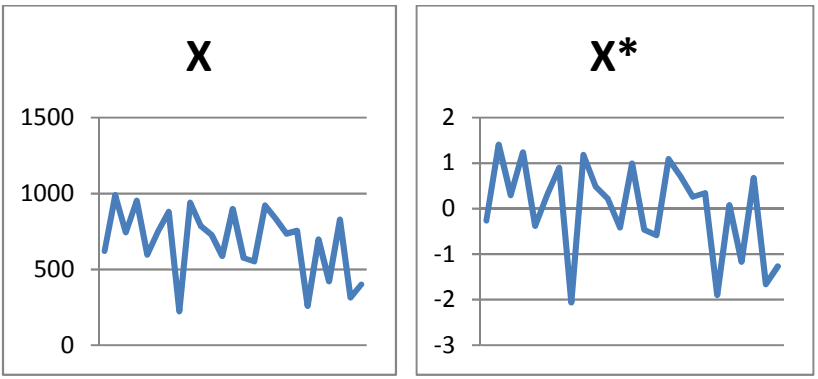
- en la centrant, i.e. en soustrayant à chaque valeur la moyenne de la série ; puis
- en la réduisant, i.e. en divisant le résultat par l'écart-type de la série.

Il s'agit donc de calculer :

$$\frac{x_i - \overline{x}}{\sigma} \quad pour\ chaque\ valeur\ i = 1, 2, \cdots, N\ de\ la\ série$$

Les variables centrées réduites n'ont pas d'unité. Elles sont de moyenne nulle et d'écart-type 1. L'ensemble des valeurs sont centrées autour de 0, certaines sont négatives, d'autres positives. Comme l'illustre par exemple le Graphique 3.1 ci-après, la centration-réduction d'une série ne modifie pas la structure de la courbe de cette série.

Graphique 3.1
Représentation de la courbe d'une même série avant (X) et après (X*) centration-réduction



EXEMPLE 3.1

On dispose des montants consacrés à l'éducation par le pays (colonne A du Tableau 3.1). On veut observer graphiquement si la courbe de l'évolution de la dépense d'éducation depuis 35 ans suit celle du nombre d'élèves (colonne B) et/ou celle du nombre d'enseignants (colonne C) dans les premier et second degrés. Étant donné les différences d'ordres de grandeur, les trois courbes ne peuvent être représentées sur le même plan. On commence donc par standardiser les trois séries (Tableau 3.2), ce qui permet ensuite de les visualiser (Graphique 3.2).

Tableau 3.1.
Évolution de la dépense d'éducation, du nombre d'élèves et du nombre d'enseignants en premier et second degrés entre 1981 et 2014

	Dépense d'éducation Milliards d'EUR (A)	Nombre d'élèves en 1er et 2nd degrés (B)	Nombre d'enseignants en 1er et 2nd degrés (C)
1981	78,8	12294900	707149
1982	82,9	12250700	723402
1983	83,7	12269800	723899
1984	85,3	12160536	730790
1985	87,4	12182183	758041
1986	87,5	12147553	759105
1987	88,6	12239432	757358
1988	91,3	12268382	766333
1989	94,8	12853500	799976
1990	99,4	12902613	810396
1991	105,0	12859300	820013
1992	112,7	12854700	829828
1993	116,2	12858300	839828
1994	119,3	12832401	846804
1995	122,8	12766049	848703
1996	124,4	12701828	853970
1997	126,6	12615566	864927
1998	129,5	12542854	867085
1999	132,6	12481516	874365

	Dépense d'éducation Milliards d'EUR (A)	Nombre d'élèves en 1er et 2nd degrés (B)	Nombre d'enseignants en 1er et 2nd degrés (C)
2000	134,4	12414667	880097
2001	134,8	12381952	888283
2002	136,4	12373004	894174
2003	136,5	12374782	894038
2004	137,3	12370320	887561
2005	137,0	12355628	884021
2006	138,6	12307266	878947
2007	139,1	12262904	870025
2008	140,3	12233449	857260
2009	144,3	12236185	852907
2010	145,1	12274743	859294
2011	144,5	12366533	849647
2012	143,5	12387813	841667
2013	145,7	12483010	855844
2014	147,0	12538258	855028

Tableau 3.2.
Données centrées-réduites de l'évolution de la dépense d'éducation, du nombre d'élèves et du nombre d'enseignants en premier et second degrés entre 1981 et 2014

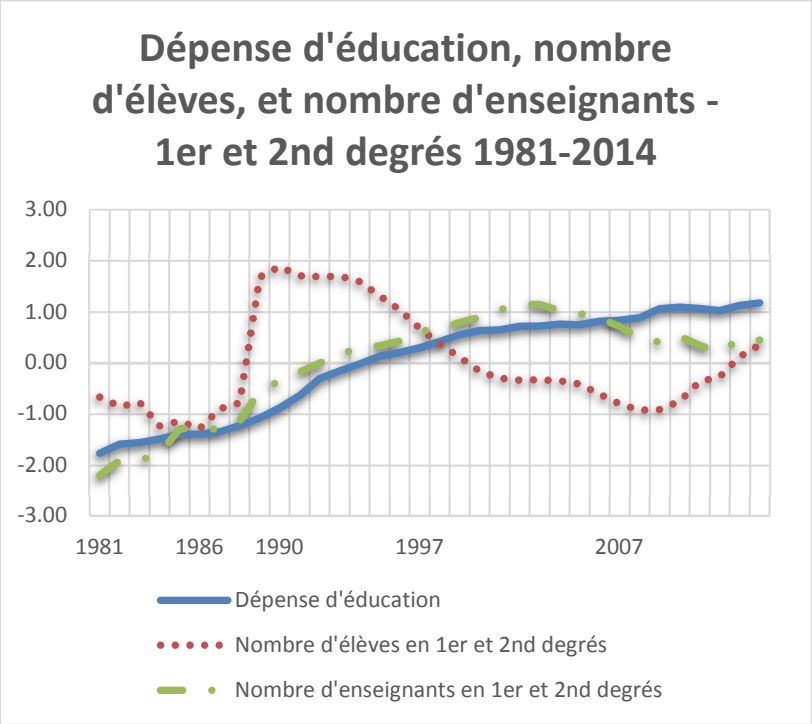
	Dépense d'éducation (A)	Nombre d'élèves en 1er et 2nd degrés (B)	Nombre d'enseignants en 1er et 2nd degrés (C)
1981	-1,77	-0,67	-2,21
1982	-1,60	-0,86	-1,92
1983	-1,56	-0,78	-1,91
1984	-1,49	-1,24	-1,78
1985	-1,40	-1,14	-1,30
1986	-1,40	-1,29	-1,28
1987	-1,35	-0,90	-1,31
1988	-1,23	-0,78	-1,15
1989	-1,08	1,68	-0,54
1990	-0,88	1,89	-0,36
1991	-0,64	1,70	-0,18
1992	-0,31	1,69	-0,01
1993	-0,16	1,70	0,17
1994	-0,02	1,59	0,30
1995	0,13	1,31	0,33
1996	0,20	1,04	0,42
1997	0,29	0,68	0,62
1998	0,42	0,37	0,66

	Dépense d'éducation (A)	Nombre d'élèves en 1er et 2nd degrés (B)	Nombre d'enseignants en 1er et 2nd degrés (C)
1999	0,55	0,11	0,79
2000	0,63	-0,17	0,89
2001	0,65	-0,30	1,04
2002	0,72	-0,34	1,15
2003	0,72	-0,33	1,14
2004	0,76	-0,35	1,03
2005	0,74	-0,41	0,96
2006	0,81	-0,62	0,87
2007	0,83	-0,81	0,71
2008	0,89	-0,93	0,48
2009	1,06	-0,92	0,41
2010	1,09	-0,76	0,52
2011	1,07	-0,37	0,35
2012	1,03	-0,28	0,20
2013	1,12	0,12	0,46
2014	1,18	0,35	0,44

Les données standardisées n'ont pas d'unité de compte et sont du même ordre de grandeur. On peut ainsi observer sur le Graphique 3.2 ci-après qu'entre 1981 et 2007, la dépense d'éducation a augmenté parallèlement au nombre d'enseignants en 1er et 2nd degrés. On peut observer aussi que sur l'ensemble de la période,

l'évolution de la dépense n'est pas liée à celle du nombre d'élèves en 1er et 2nd degrés.

Graphique 3.2



3.2.2. Standardisations max-min

On distingue trois variantes de standardisation max-min : la standardisation max-min en intervalle [0 ; 1] ; la standardisation max-min en intervalle [-1 ; 1] ; et la standardisation max-min en intervalle quelconque.

3.2.2.1. Standardisation max-min en intervalle [0 ; 1]

La méthode consiste à transformer la série de sorte que toutes ses valeurs soient comprises dans un intervalle [0 ; 1]. Pour ce faire, on remplace chaque valeur x_i de la série par

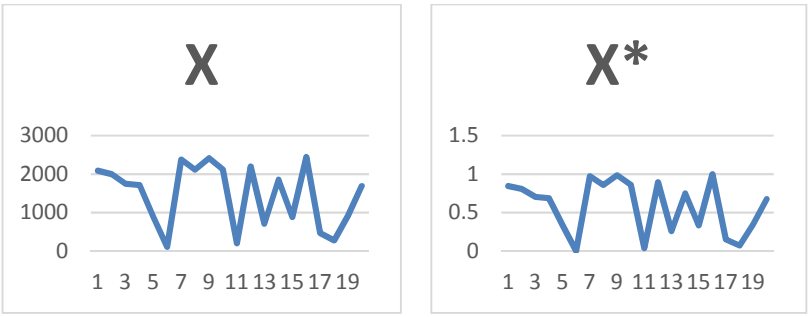
$$x_i^* = \frac{x_i - Min}{Max - Min}$$

où
Min est le minimum de la série considérée ; et
Max son maximum.

La série sera par conséquent transformée en une série bornée par 0 (transformée du minimum de cette série) et 1 (transformée du maximum de la série). On voit qu'il n'y a pas, avec cette méthode, distribution des valeurs standardisées autour de zéro, comme c'était le cas avec la standardisation par centration-réduction.

Comme l'illustre par exemple le Graphique 3.3, la standardisation des données d'une série par cette méthode ne modifie pas la structure de la courbe de cette série.

Graphique 3.3
Représentation de la courbe d'une même série avant (X) et après (X*) standardisation max-min en intervalle [0 ; 1]



EXEMPLE 3.2

L'indice de développement humain (IDH) des Nations-Unies a été créé afin de mesurer les progrès des pays en matière de développement socio-économique (éducation, santé, revenus), par opposition à une mesure exclusivement économique du développement. L'IDH combine quatre indicateurs :

- I_1, indicateur d'espérance de vie à la naissance, qui reflète les progrès en matière de santé ;
- I_2, indicateur de durée attendue de scolarisation (18 ans au maximum, ce qui correspond dans la plupart des pays a la durée d'études nécessaire pour atteindre le niveau master) ;

- I_3, indicateur de durée moyenne de scolarisation (15 ans au maximum, ce qui est l'objectif des Nations-Unies pour 2025) ; et
- I_4, indicateur de revenu par habitant.

L'IDH d'un pays est calculé comme suit :

$$IDH = \sqrt[3]{I_1 \times \left(\frac{I_2 + I_3}{2}\right) \times I_4}$$

Comme on ne peut pas additionner ou multiplier directement des durées en années et des montants de revenus, chacune de ces composantes (espérance de vie, scolarisation, revenu) a dû d'abord être transformée en un indicateur permettant de la combiner avec les autres. La méthode utilisée est la standardisation max-min en intervalle [0 ; 1]. Ainsi par exemple, pour un pays i quel qu'il soit, l'indicateur de durée attendue de scolarisation est défini par :

$$I_{2_i} = \frac{Durée\ attendue\ de\ scolarisation\ en\ Pays\ i - 0}{18 - 0}$$

De même, l'indicateur de durée moyenne de scolarisation est défini par :

$$I_{3_i} = \frac{Durée\ moyenne\ de\ scolarisation\ en\ Pays\ i - 0}{15 - 0}$$

Par exemple, pour un pays dans lequel la durée moyenne de scolarisation s'élève à 12 ans, l'indicateur de durée moyenne de scolarisation s'établira à

$$Indicateur\ de\ durée\ moyenne\ de\ scolarisation = \frac{12}{15} = 0{,}8$$

3.2.2.2. Standardisation max-min autour de zéro

La méthode ici consiste à transformer la série en une suite de valeurs comprises entre -1 et 1, en remplaçant chaque valeur x_i par :

$$x_i^* = \frac{x_i - \left(\frac{Max + Min}{2}\right)}{\frac{Max - Min}{2}}$$

où
Min est le minimum de la série considérée ; et
Max son maximum.

EXEMPLE 3.3

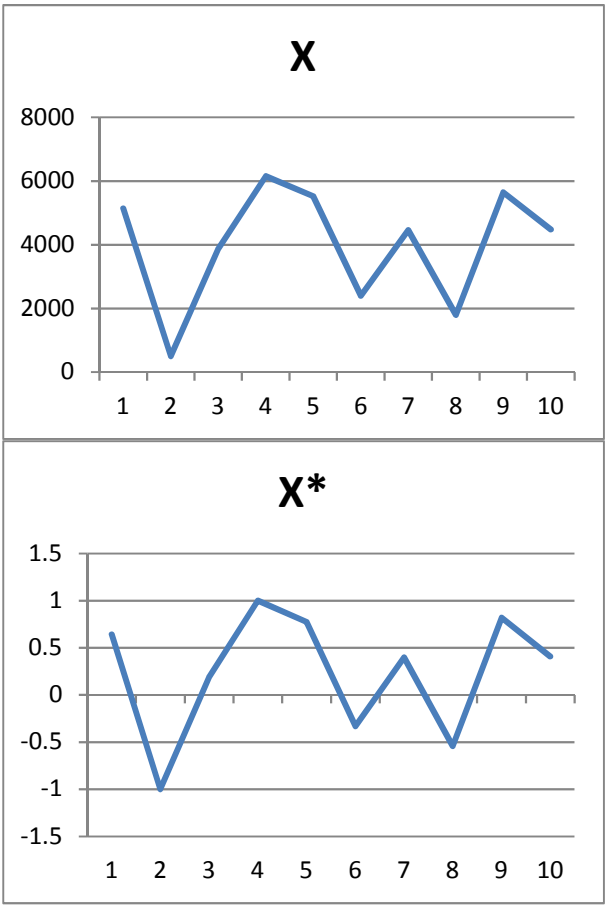
Le Tableau 3.3 ci-après décrit les étapes de la transformation max-min autour de zéro pour la série en colonne A. Le Graphique 3.4 montre la courbe de la série avant et après standardisation.

Tableau 3.3

Étapes de standardisation max-min autour de zéro

A	B	C
x_i	$x_i - \left(\frac{Max + Min}{2}\right)$	$\frac{x_i - \left(\frac{Max + Min}{2}\right)}{\frac{Max - Min}{2}}$
5147	1818,5	0,643602902
503	-2825,5	-1
3867	538,5	0,190585737
6154	2825,5	1
5520	2191,5	0,775614935
2395	-933,5	-0,330384003
4456	1127,5	0,399044417
1800	-1528,5	-0,540966201
5646	2317,5	0,820208813
4480	1151,5	0,407538489
$\frac{Max + Min}{2} = 3328{,}5$		
$\frac{Max - Min}{2} = 2825{,}5$		

Graphique 3.4
Courbe avant (X) et après (X*) standardisation max-min autour de zéro



3.2.2.3. Standardisation max-min dans un intervalle quelconque

On cherche ici à ce que toutes les valeurs transformées s'inscrivent dans un intervalle $[a, b]$. On remplace chaque valeur x_i de la série par

$$x_i^* = a + \frac{(x_i - Min)(b-a)}{Max - Min}$$

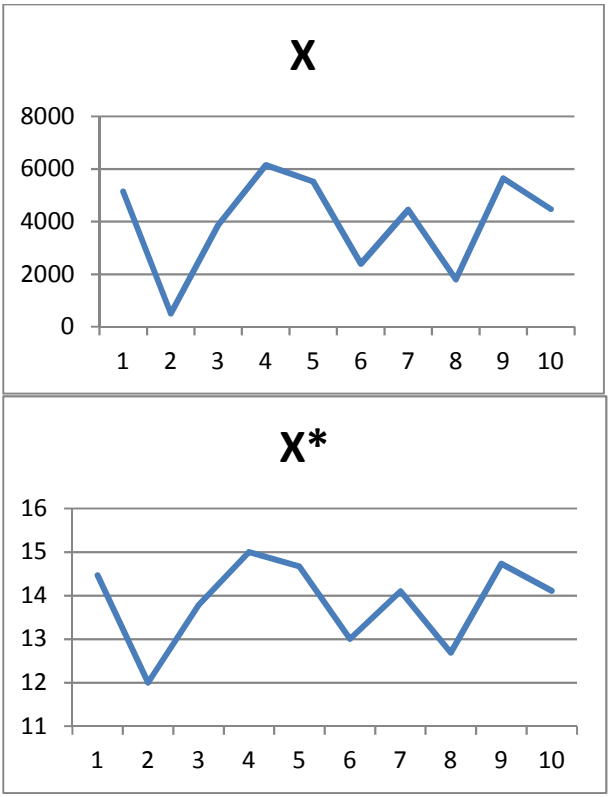
EXEMPLE 3.4

On cherche à ramener toutes les valeurs de la série du Tableau 3.3 *supra* dans un intervalle [12 ; 15]. Le Tableau 3.4 ci-après illustre la démarche.

Tableau 3.4
Standardisation max-min en intervalle [12 ; 15]

A	B	C
x_i	$x_i - Min$	$a + \frac{(x_i - Min)(b-a)}{Max - Min}$
5147	4644	14,47
503	0	12,00
3867	3364	13,79
6154	5651	15,00
5520	5017	14,66
2395	1892	13,00
4456	3953	14,10
1800	1297	12,69
5646	5143	14,73
4480	3977	14,11
$Max - Min = 5651$		

Graphique 3.5.
Courbe avant (X) et après (X*) standardisation max-min dans l'intervalle [12 ; 15]



3.2.3. Standardisation par moyenne ou écart-type

La méthode consiste à diviser chaque terme de la série soit par la moyenne de la série (méthode 1), soit par son écart-type (méthode 2) :

- Méthode 1

$$x_i^* = \frac{x_i}{\overline{x}}$$

- Méthode 2

$$x_i^* = \frac{x_i}{\sigma}$$

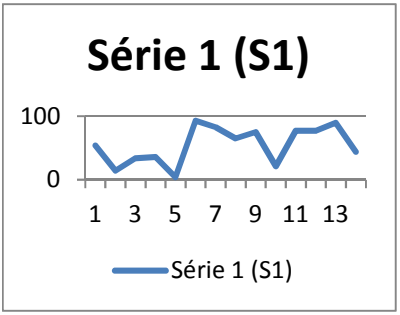
EXEMPLE 3.5

Soit le Tableau 3.5 suivant des séries S1 et S2 :

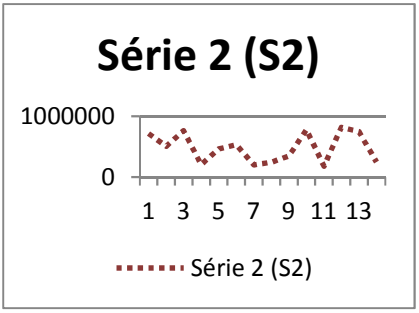
Série 1 (S1)	Série 2 (S2)
91	589654
79	259527
4	444002
54	630323
40	239168
54	182812
30	145015
61	519817
82	201862
72	361784
23	828524
36	599089
76	284791
33	478318

Graphique 3.6
Courbes avant et après standardisation par moyenne ou écart-type

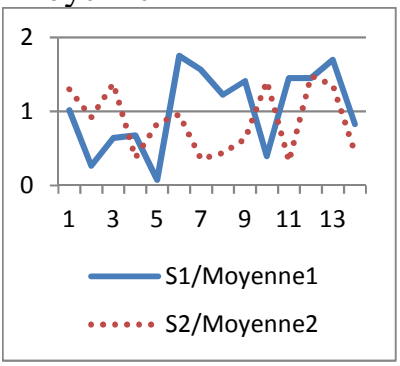
Série 1 avant standardisation



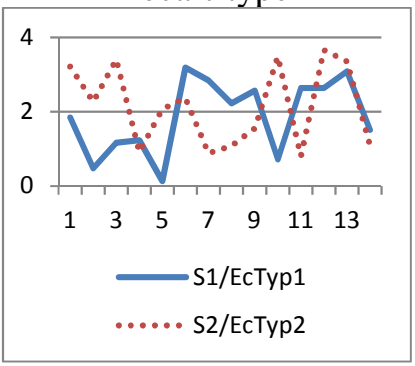
Série 2 avant standardisation



Séries 1 et 2 après standardisation par la moyenne



Séries 1 et 2 après standardisation par l'écart-type



Les deux séries ne peuvent évidemment être représentées sur un même plan. En revanche, elles peuvent l'être après avoir été réduites en les divisant par leurs moyennes ou par leurs écarts-types respectifs. Le Graphique 3.6 montre que la transformation ne modifie pas la structure des courbes.

Deuxième Partie : Analyser les données

Une fois les données collectées et préparées, il s'agit de les analyser. La gamme des méthodes d'analyse quantitative de données est riche et en constante évolution. La première étape pour le chercheur consiste à sélectionner la méthode pertinente à mettre en œuvre pour sa recherche. Le critère le plus important à prendre en compte pour effectuer cette sélection est l'objectif de l'analyse. Par exemple, les méthodes descriptives peuvent être suffisantes si l'objectif est de décrire un échantillon sans intention d'en inférer des conclusions généralisables à l'ensemble de sa population-mère. Il est donc important d'être, dès le départ, au clair avec l'objectif de l'analyse, afin de sélectionner la méthode la plus appropriée, et d'être ensuite en mesure de justifier son choix.

Il n'est pas rare cependant que pour un objectif donné, plusieurs méthodes concurrentes soient disponibles. Par exemple, on peut comparer des moyennes en appliquant des tests d'hypothèses ou en effectuant une analyse de variance. Lorsque plusieurs méthodes sont disponibles au service d'un même objectif, d'autres critères de choix peuvent intervenir, par exemple le fait que certaines méthodes peuvent être plus lourdes à mettre en œuvre que d'autres. De façon générale, il est essentiel d'éviter l'arbitraire, d'examiner rigoureusement toutes les facettes du problème, d'effectuer un choix de méthode, et de soigneusement

étayer son choix. En revanche, la pratique de la « triangulation », fréquente en recherche qualitative et qui consiste à approcher le même problème par plusieurs méthodes concurremment, n'est pas vraiment d'usage courant en recherche quantitative.

Enfin, toutes les méthodes d'analyse quantitative nécessitent que des conditions de validité soient remplies. La vérification de ces conditions fait partie du travail d'analyse et contribue à déterminer la validité et la fiabilité des résultats obtenus. Le cas échéant, la phase de vérification des conditions de validité permet de repérer et de traiter des sources de fragilité. Rendre compte de la vérification des conditions de validité fait partie intégrante du rapport d'analyse. Les conditions qui s'avèrent ne pas être remplies doivent être explicitement signalées dans le rapport d'analyse comme autant de réserves sur la validité des résultats obtenus.

Chapitre 4. Lexicométrie : l'étude quantitative de textes

L'application des méthodes quantitatives n'est pas réservée aux données numériques. Les données textuelles peuvent, elles-aussi, être traitées de façon quantitative. Il en va ainsi des réponses aux entretiens et des réponses aux questions ouvertes de questionnaires, et plus généralement de tout texte littéraire, journalistique ou politique. Habituellement, les données textuelles sont traitées par analyse de contenu, c'est-à-dire un ensemble de méthodes qualitatives qui permettent d'identifier la signification d'un discours à partir de ses caractéristiques lexicales et syntaxiques, et de cerner l'univers de référence et les attitudes du locuteur (Bardin, 2013). La lexicométrie vise, elle aussi, à appréhender le sens d'un propos, mais privilégie les observations quantitatives. Les deux types d'approches ont des points communs : l'analyse de contenu fait elle aussi appel au dénombrement d'occurrences, tandis que la lexicométrie de son côté nécessite l'identification de catégories sémantiques.

L'analyse lexicométrique s'organise en trois principales étapes[53].

[53] On ne revient pas ici sur la préparation technique du fichier, qui dépend du logiciel utilisé. Un logiciel très adapté à cette démarche est Lexico, développé par l'Université Paris 3 : http://www.lexi-co.com/Produits.html

4.1. ÉTABLIR LE CADRE DE RÉFÉRENCE

La première étape consiste à établir l'univers conceptuel et théorique dans le cadre duquel le matériau textuel (*corpus*) sera analysé. Il s'agit ici de répondre à la question : quelle est la théorie et quels sont les concepts à la lumière desquels l'objet d'étude peut être analysé. Le cadre de référence indique ce que devrait être, d'après la théorie, le contenu sémantique du corpus étudié : quels thèmes, quelles idées, quelles catégories, quels termes. Les catégories elles-mêmes sont de différents types (catégories d'acteurs, de contextes, d'actions, de moyens, etc.).

En découle une *grille d'analyse* du contenu du corpus.

Imaginons par exemple une recherche sur la gouvernance de l'école. Il est prévu d'effectuer une enquête par entretien auprès d'acteurs de terrain. Les répondants seront interrogés sur les défis prioritaires à relever au cours de la prochaine décennie. Une grille pour guider l'analyse des réponses qui seront obtenues pourrait comporter les éléments suivants :

- catégories d'acteurs possibles : gouvernement, autorités publiques régionales et locales, directions d'établissements, représentants des personnels, représentants des élèves, représentants des familles, représentants des milieux économiques ;

- processus de décision : cogestion, concertation, votes, consultation facultative, consultation obligatoire, avis conforme, autocratie ;
- processus d'interaction : échange d'informations, négociations, conflits sociaux, arbitrages, procédures judiciaires ;
- processus de régulation : inspections, audits externes, contrôles administratifs, contrôles hiérarchiques, évaluation par les pairs, auto-évaluation, évaluation par les usagers ;
- outils d'élaboration du projet collectif : prospective, expérimentations, recherche académique, planification, innovation, réforme ;
- valeurs partagées : transparence, participation, équité, égalité, fiabilité, hiérarchie, respect, liberté, compétence, éthique, humanisme, solidarité, efficacité, légalité, développement professionnel, internationalisation, autonomie, responsabilité, intégration sociale, excellence.

La grille balise l'espace des propos possibles étant donné la question de recherche. Elle décrit les thèmes, idées, termes et catégories que l'on peut s'attendre à rencontrer dans un discours sur l'objet d'étude. Elle précède l'analyse du corpus et n'en dépend pas. Elle constitue en ce sens une base de référence qui permet de renforcer le caractère objectif du travail du chercheur. En effet, sans grille de référence, le

chercheur ne trouverait dans son corpus que ce que sa propre subjectivité y mettrait. Si c'était le cas, le résultat de l'analyse serait dépendant de la subjectivité du chercheur. Au contraire, si une grille a été établie avant l'analyse du corpus, le chercheur pourra interpréter le corpus par référence à la grille, et mettre en évidence quelles composantes de la grille se retrouvent dans le corpus, quelles composantes ne s'y retrouvent pas, et s'interroger sur les contenus du corpus étrangers à la grille, ces derniers conduisant à actualiser la grille ou à mettre en évidence les particularités du discours analysé. Tout chercheur auquel la grille serait communiquée et qui analyserait le même corpus tendrait à faire les mêmes observations et à aboutir aux mêmes conclusions.

4.2. RÉPERTORIER LES FORMES PAR FRÉQUENCE D'OCCURRENCE

Il s'agit ici d'établir la liste des termes / mots (« formes ») contenus dans le corpus, classés par fréquence d'apparition. L'analyse se concentre sur les formes ayant un contenu sémantique, et ignore articles, pronoms, prépositions, conjonctions, et tous autres éléments de structure grammaticale, sauf bien entendu si le style littéraire fait partie de l'objet d'étude.

En règle générale, cette liste primaire des formes nécessite un travail d'élaboration supplémentaire. Il

faut en effet tenir compte des *synonymes*, des *associations de formes*, et des *groupes de formes*.

Les synonymes, d'abord, doivent être identifiés. Ils expriment en effet un même contenu sémantique, or ils apparaissent dans la liste primaire comme des formes distinctes, donc comptabilisées séparément. Il importe donc d'établir des listes de synonymes, de façon à disposer d'une vision plus exacte de la récurrence d'un contenu sémantique dans le corpus.

Les associations de formes sont des expressions dans lesquelles deux formes sémantiques ou plus sont utilisées conjointement. Or une forme n'a pas forcement la même signification lorsqu'elle est utilisée seule ou dans une expression. Il importe donc de pouvoir repérer les associations de formes, de façon à distinguer leur signification spécifique de celle des mêmes formes employées isolément. La signification des associations de formes (ou des formes elles-mêmes du reste) peut être notamment précisée grâce aux « concordances », c'est-à-dire aux contextes / phrases dans lesquelles ces associations apparaissent. Chaque association de formes ayant une signification spécifique doit être répertoriée sur la liste des formes, et ses occurrences faire l'objet d'une comptabilisation distincte.

Les groupes de formes doivent être identifiés aussi. Un groupe de formes est l'ensemble des formes présentant une caractéristique commune. C'est le cas des formes commençant par une même chaîne de caractères. Par exemple, le groupe de formes « certif » comprend les mots « certifié », « certifier », « certificat(s) », « certification(s) », etc. Parfois, il peut y avoir du sens à travailler sur des groupes de formes plutôt que de considérer séparément les occurrences de ces différentes formes.

4.3. ANALYSE SÉMANTIQUE

Il s'agit ici de mettre en regard la fréquence observée des différentes formes avec la grille de référence établie en étape 1. L'analyse consiste à identifier et commenter, par référence à la grille :

- les idées, catégories, thèmes et termes qui figurent dans le corpus
  - Signification
  - Hiérarchie dans l'insistance
- les idées, catégories, thèmes et termes qui n'y figurent pas ;
- les idées, catégories, thèmes et termes qui constituent des originalités/innovations ;
- les représentations, explicites ou implicites ;

- les attitudes et positionnements ;
- les non-dits et arrières pensées ;
- les idéologies sous-jacentes ;
- les incohérences ; etc.

L'analyse s'effectue pour l'ensemble du corpus. Elle peut s'effectuer aussi, en outre, pour chaque partie du corpus, de façon à mettre en évidence l'éventuelle spécificité de telle ou telle partie du corpus, par exemple en termes de sur- ou sous-représentation de tel ou tel contenu sémantique par date de discours ou (type de) répondant.

Chapitre 5. Analyse classificatoire : discerner des groupes homogènes

L'analyse classificatoire s'utilise dans le cas où on cherche à répartir un ensemble de données en sous-groupes homogènes. On dispose d'observations relatives à un ensemble d'individus, et on cherche à répartir ces individus en groupes dont les membres sont plus proches les uns des autres qu'ils ne le sont du reste de la population. C'est le cas par exemple lorsqu'on cherche à distinguer des profils, des groupes de comportements, des groupes de performances, etc.

La problématique méthodologique de l'analyse classificatoire est différente de celle du calcul de quartiles, déciles ou centiles. En premier lieu, lorsqu'on décompose une série en quarts, dixièmes ou centièmes, l'affectation d'un individu à un intervalle s'effectue au regard d'une unique variable. Considérons par exemple les stagiaires inscrits dans un établissement de formation, on pourra bâtir des intervalles interquartiles au regard du score au test d'entrée, ou au regard du nombre d'heures de formation requises pour une remise à niveau, ou au regard de l'âge, etc. L'analyse classificatoire, elle, peut permettre de constituer des groupes homogènes au regard de plusieurs variables simultanément, réunissant au sein d'un même groupe des individus proches en termes à la fois d'âge, de besoin de formation et de performance au test d'entrée.

En deuxième lieu, dans le découpage en intervalles, les classes / intervalles sont, par définition, d'effectifs identiques. Dans l'analyse classificatoire, au contraire, la taille des groupes n'est pas prédéterminée. La taille d'un groupe dépend uniquement du nombre d'individus suffisamment proches pour pouvoir faire partie de ce groupe.

En troisième lieu, dans un intervalle, un individu situé à la borne supérieure est plus proche de l'individu situé à la borne inferieure de l'intervalle suivant qu'il ne l'est de l'individu situé à l'autre borne de l'intervalle auquel lui-même appartient. L'analyse classificatoire, au contraire, permet de n'inclure dans le groupe que les individus les plus proches entre eux au regard de toutes les variables d'intérêt.

La notion de distance est essentielle en analyse classificatoire puisque c'est la distance entre deux individus qui permet de décider si ces deux individus sont proches (et doivent faire partie du même groupe) ou non. Il existe plusieurs concepts de distance mais, quel que soit le concept choisi, la distance peut se mesurer pour toute variable quantitative. En revanche, il n'est pas possible de calculer une distance lorsque les variables au regard desquelles les individus sont caractérisés sont qualitatives. La classification repose alors sur l'identification de « classes latentes ». Quatre méthodes d'analyse classificatoire sont présentées dans

ce chapitre, correspondant chacune à une configuration particulière.

5.1. PARTITIONNEMENT UNIVARIÉ

La méthode du partitionnement univarié permet de traiter les cas dans lesquels on dispose de données sur une unique variable quantitative (X) pour plusieurs individus $i = 1, \cdots, n$, et où on cherche à répartir ces individus en groupes homogènes au regard de cette variable. Le partitionnement univarié permet d'effectuer cette classification. Le nombre de classes doit être défini par le chercheur, en fonction de la théorie dans le cadre de laquelle l'analyse s'inscrit.

EXEMPLE 5.1

On dispose de données sur la participation des ouvriers qualifiés aux formations financées par l'employeur dans l'Union européenne (Tableau 5.1). On considère que les pays doivent se répartir en quatre groupes : "Avancés", "Moyen-supérieur", "Moyen-inférieur", "En retard". On cherche à identifier la composition de chaque groupe.

Tableau 5.1.
Taux de participation des ouvriers qualifiés aux formations financées par l'employeur – 2011

Source : Eurostat (indicateur trng_aes_124)

Belgique	18,4 %
Bulgarie	35,7 %
République Tchèque	28,5 %
Danemark	42,3 %
Allemagne	35,4 %
Estonie	29,2 %
Espagne	27,5 %
France	34,8 %
Italie	23,7 %
Chypre	25,2 %
Lettonie	15,1 %
Lituanie	10,1 %
Luxembourg	55,6 %
Hongrie	45,2 %
Malte	19,0 %
Pays-Bas	48,5 %
Autriche	30,6 %
Pologne	12,3 %
Portugal	31,4 %
Roumanie	3,9 %
Slovénie	19,6 %
Slovaquie	36,2 %
Finlande	36,3 %
Suède	48,0 %
Royaume-Uni	12,8 %

Les résultats du partitionnement apparaissent comme suit :

Classes			
1	2	3	4
Danemark	Bulgarie	République Tchèque	Belgique
Luxembourg	Allemagne	Estonie	Lettonie
Hongrie	France	Espagne	Lituanie
Pays-Bas	Slovaquie	Italie	Malte
Suède	Finlande	Chypre	Pologne
		Autriche	Roumanie
		Portugal	Slovénie
			Royaume-Uni

Barycentres des classes	
Classe	Taux de participation
1	47,920
2	35,680
3	28,014
4	13,900

Objets centraux	
Classe	Taux de participation
1 (Suède)	48,000
2 (Bulgarie)	35,700
3 (République Tchèque)	28,500
4 (Royaume-Uni)	12,800

On constate que la taille des groupes est variable. Le groupe "Avancé" comprend 5 membres autour de la Suède avec 47,92% de participation en moyenne. Le groupe "Moyen-supérieur" comprend lui aussi 5 membres, autour de la Bulgarie avec 35,6% de participation en moyenne. Le groupe "Moyen-inférieur" comporte 7 membres, autour de la République tchèque avec 28,01% de participation en moyenne. Enfin le groupe "En retard" comporte lui 8 membres, autour du Royaume-Uni avec 13,9% de participation en moyenne.

5.2. MÉTHODE K-MEANS

La méthode k-means (ou méthode des nuées dynamiques) permet de traiter les cas dans lesquels on dispose d'observations relatives à plusieurs variables quantitatives $(X_1, X_2, , \cdots, X_\xi)$ pour plusieurs individus $(i = 1, 2, \cdots, n)$. On cherche à répartir ces individus en groupes homogènes au regard de l'ensemble de ces variables. La méthode permet d'effectuer cette classification. Le nombre de classes doit être défini par le chercheur, en fonction de la théorie dans le cadre de laquelle l'analyse s'inscrit.

EXEMPLE 5.2

On dispose de données sur la dépense par élève aux différents niveaux d'enseignement dans les pays de l'OCDE (Tableau 5.2) :
- Pré-primaire (« PP ») ;
- Primaire (« P ») ;
- Secondaire (« Sec. ») ;
- Supérieur (« Sup. »).

On considère que les pays doivent se répartir en trois groupes : "Avancés", "Moyen", "En retard". On cherche à identifier la composition de chaque groupe.

Tableau 5.2.
Dépense moyenne par élève aux différents niveaux d'enseignement dans les pays de l'OCDE (en équivalent USD) – 2010

Source : OCDE (2013), *Regards sur l'éducation* (indicateur T_B1.1a)

	PP	P	Sec.	Sup.
Australie	8 899	9 463	10 350	15 142
Autriche	8 893	10 244	12 551	15 007
Belgique	6 024	8 852	11 004	15 179
Chili	3 544	3 301	3 110	7 101
République Tchèque	4 247	4 120	6 546	7 635
Danemark	9 454	10 935	11 747	18 977
Estonie	2 533	5 140	6 444	6 501

	PP	P	Sec.	Sup.
Finlande	5 372	7 624	9 162	16 714
France	6 362	6 622	10 877	15 067
Hongrie	4 773	4 684	4 553	8 745
Islande	8 606	9 482	7 841	8 728
Irlande		8 384	11 380	16 008
Israël	3 910	5 758	5 616	10 730
Italie	7 177	8 296	8 607	9 580
Japon	5 550	8 353	9 957	16 015
Corée du sud	6 739	6 601	8 060	9 972
Luxembourg	20 958	21 240	17 633	
Mexique	2 280	2 331	2 632	7 872
Pays-Bas	7 664	7 954	11 838	17 161
Nouvelle Zélande	11 495	6 842	8 170	10 418
Norvège	6 610	12 255	13 852	18 512
Pologne	5 737	5 937	5 483	8 866
Portugal	5 977	5 922	8 882	10 578
Slovaquie	4 306	5 732	4 806	6 904
Slovénie	7 744	8 935	8 187	9 693
Espagne	6 685	7 291	9 608	13 373
Suède	6 582	9 987	10 185	19 562
Suisse	5 186	11 513	14 972	21 893
Turquie	2 490	1 860	2 470	
Royaume-Uni	7 047	9 369	10 452	15 862
États-Unis	10 020	11 193	12 464	25 576

Les résultats des calculs s'établissent comme suit :

1 (5 pays)	2 (9 pays)	3 (14 pays)
Danemark	Australie	Chili
Norvège	Autriche	République Tchèque
Suède	Belgique	Estonie
Suisse	Finlande	Hongrie
États-Unis	France	Islande
	Japon	Israël
	Pays-Bas	Italie
	Espagne	Corée du sud
	Royaume-Uni	Mexique
		Nouvelle Zélande
		Pologne
		Portugal
		Slovaquie
		Slovénie

Barycentres des classes				
Classe	PP	P	Sec.	Sup.
1	7570,40	11176,60	12644,00	20904,00
2	6944,00	8419,11	10644,33	15502,22
3	5647,71	5934,35	6352,64	8808,78

Objets centraux				
Classe	PP	P	Sec.	Sup.
1 (Danemark)	9454	10935	11747	18977
2 (Royaume-Uni)	7047	9369	10452	15862
3 (Pologne)	5737	5937	5483	8866

Le groupe de tête comprend cinq pays qui, autour du Danemark, ont en moyenne la dépense la plus élevée tous niveaux d'éducation pris en compte. Le groupe intermédiaire comprend 9 pays autour du Royaume-Uni. Le groupe de queue réunit les 14 autres pays autour de la Pologne.

5.3. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

Comme dans la méthode k-means, on dispose d'observations relatives à plusieurs variables quantitatives $(X_1, X_2, , \cdots, X_\xi)$ pour plusieurs individus $(i = 1, 2, \cdots, n)$. On cherche à répartir ces individus en groupes homogènes au regard de l'ensemble de ces variables. Mais à la différence de la méthode k-means, la classification ascendante hiérarchique (CAH) laisse au chercheur le choix de fixer ou non *a priori* le nombre de groupes. La méthode permet en effet, en l'absence de contrainte théorique, de déterminer par calcul le nombre de groupes qui permet d'assurer la plus grande homogénéité / proximité intra-groupe.

EXEMPLE 5.3

On dispose des notes au bac des 20 candidats d'un lycée pour trois disciplines : français, mathématiques et sport (Tableau 5.3). On cherche à identifier des profils de candidats sans déterminer *a priori* le nombre de profils à distinguer.

Tableau 5.3
Notes au bac en français, maths et sport de 20 candidats du lycée Lambda

	Français	Maths	Sport
Candidat-e 1	13	3	19
Candidat-e 2	14	13	19
Candidat-e 3	13	0	14
Candidat-e 4	9	1	19
Candidat-e 5	11	17	19
Candidat-e 6	12	2	14
Candidat-e 7	16	3	14
Candidat-e 8	10	18	14
Candidat-e 9	12	2	19
Candidat-e 10	13	19	19
Candidat-e 11	8	14	14
Candidat-e 12	12	18	17
Candidat-e 13	15	6	14
Candidat-e 14	16	4	18
Candidat-e 15	12	11	17
Candidat-e 16	16	11	19
Candidat-e 17	11	11	16

	Français	Maths	Sport
Candidat-e 18	17	0	17
Candidat-e 19	14	7	16
Candidat-e 20	12	15	16

Résultats des calculs

Profil 1	Profil 2	Profil 3
Candidat-e 1	Candidat-e 3	Candidat-e 8
Candidat-e 2	Candidat-e 6	Candidat-e 11
Candidat-e 4	Candidat-e 7	Candidat-e 12
Candidat-e 5	Candidat-e 13	Candidat-e 15
Candidat-e 9	Candidat-e 14	Candidat-e 17
Candidat-e 10	Candidat-e 18	Candidat-e 20
Candidat-e 16	Candidat-e 19	

Barycentres des classes			
Classe	Français	Maths	Sport
1	12,57	9,42	19,00
2	14,71	3,14	15,28
3	10,83	14,50	15,66

Objets centraux			
Classe	Français	Maths	Sport
1 (Candidat-e 16)	16	11	19
2 (Candidat-e 7)	16	3	14
3 (Candidat-e 20)	12	15	16

Suivant cette méthode, trois profils distincts émergent. Le premier est celui de candidats caractérisés par une

note excellente en sport. Le profil 2 est celui de candidats avec une faiblesse marquée en maths. Le profil 3 regroupe des candidats moyens en français mais qui obtiennent une note d'au moins 14 dans au moins l'une des deux autres matières.

5.4. CLASSIFICATION EN CLASSES LATENTES

La classification en classes latentes permet de traiter les cas dans lesquels on dispose d'observations relatives à plusieurs variables qualitatives pour plusieurs individus. S'agissant de variables qualitatives, on ne peut calculer de distance entre individus. L'alternative est alors de calculer les covariations entre variables. La méthode repose sur le postulat que les variables qualitatives observées sont en fait liées à des variables qualitatives sous-jacentes, non directement observables, mais qui les déterminent. Par suite, les variables qualitatives observées qui varient de la même façon sont présumées appartenir à une même modalité d'une variable qualitative nominale sous-jacente (classe latente). Le chercheur peut, en fonction de son cadre théorique, fixer le nombre de classes latentes, ou identifier le nombre optimal au regard d'une batterie de critères.

EXEMPLE 5.4

On dispose de données sur les actions de soutien individualisé dispensées dans 2000 établissements scolaires, et on voudrait identifier des profils d'établissements dans ce secteur. Cinq domaines de soutien ont été définis en fonction du type de difficulté à traiter :

- A – Environnement social défavorisé
- B – Habitat dispersé
- C – Problèmes de santé
- D – Non-maitrise de la langue française
- E – Difficultés d'apprentissage

Pour chaque domaine, quatre modalités d'intervention ont été identifiées :

- 1 – Établissement seul
- 2 – Partenariat Établissement-Municipalité
- 3 – Partenariat Établissement-Associations
- 4 – Autre partenariat.

Le Tableau 5.4 donne une idée de la structure générale des données collectées. On fixe à trois le nombre de profils souhaités.

Tableau 5.4.
Modalités des interventions de soutien scolaire par domaine de soutien dans 2000 établissements

	A	B	C	D	E
Établissement 1	4	3	4	2	2
Établissement 2	4	4	2	2	3
Établissement 3	3	2	4	4	2
⋮	⋮	⋮	⋮	⋮	⋮
Établissement 1998	2	1	2	3	3
Établissement 1999	4	4	4	1	3
Établissement 2000	4	2	4	1	4

Résultats des calculs

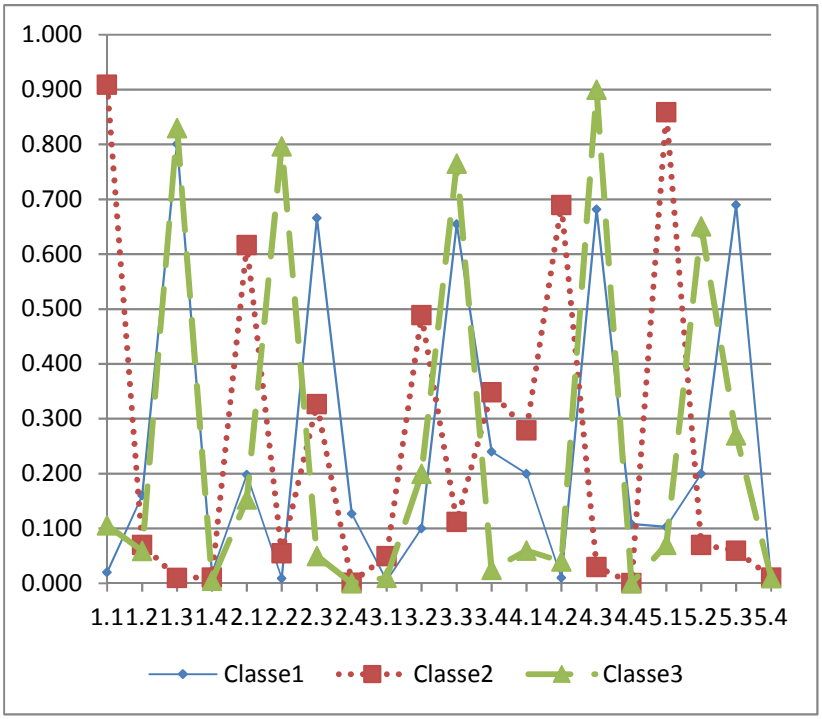
		Probabilités de recours aux modalités		
		Classe 1	Classe 2	Classe 3
Variables	Modalités			
Environnement social défavorisé	1.1	0,020	0,909	0,106
	1.2	0,160	0,070	0,059
	1.3	0,801	0,010	0,830
	1.4	0,019	0,011	0,005
Habitat dispersé	2.1	0,198	0,617	0,153
	2.2	0,009	0,055	0,797
	2.3	0,666	0,327	0,050
	2.4	0,127	0,002	0,000

		Probabilités de recours aux modalités		
		Classe 1	Classe 2	Classe 3
Variables	Modalités			
Problèmes de santé	3.1	0,005	0,050	0,010
	3.2	0,100	0,489	0,200
	3.3	0,655	0,112	0,765
	3.4	0,240	0,348	0,025
Non-maitrise de la langue française	4.1	0,200	0,279	0,060
	4.2	0,010	0,690	0,040
	4.3	0,682	0,030	0,900
	4.4	0,108	0,001	0,000
Difficultés d'apprentissage	5.1	0,103	0,859	0,070
	5.2	0,200	0,070	0,651
	5.3	0,690	0,060	0,270
	5.4	0,007	0,011	0,010

Le graphique du profil des classes (Graphique 5.1 ci-après) permet de visualiser les caractéristiques de chaque classe. On voit que les établissements de la classe 1 ont une forte probabilité (entre 65% et 83%) de privilégier le partenariat avec les associations (modalité 3) quel que soit le domaine de soutien. Les établissements de la classe 2 ont une très forte probabilité d'intervenir seuls (entre 61 et 90%) quand il s'agit d'environnement social défavorisé, d'habitat

dispersé et de difficultés d'apprentissage. Les établissements de la classe 3 ont une très forte

Graphique 5.1.
Profil des classes



probabilité (80 à 90%) d'intervenir en partenariat avec les municipalités lorsqu'il est question d'habitat dispersé et de difficultés d'apprentissage, et en partenariat avec les associations dans les trois autres domaines.

Chapitre 6. Introduction aux tests statistiques : les tendances apparentes sont-elles réellement significatives ?

6.1. PROBLÉMATIQUE

Imaginons que l'analyse des réponses à un sondage indique que 46% des répondants sont favorables à une option proposée, tandis que 40% y sont opposés. D'un strict point de vue arithmétique, il y a une majorité de réponses favorables. Mais d'un point de vue statistique, l'écart entre les deux groupes est-il significatif ? Peut-on être certain que tout ne s'est pas joue à si peu de participants près que la majorité aurait pu être dans l'autre camp ? Que représente un écart de six points si le nombre total de répondants était de 15 ? L'écart est-il suffisamment marqué pour qu'on puisse dire qu'il n'est pas juste dû au hasard ? Peut-on affirmer que le résultat serait allé dans le même sens si les circonstances avaient été légèrement différentes ?

C'est ce type de problématique que permettent d'approcher les tests statistiques. De façon générale, les tests statistiques permettent de vérifier si des écarts entre chiffres peuvent être considérés comme suffisants pour que les indications fournies par ces chiffres puissent elles-mêmes être considérées comme *probablement* solides et stables.

Les tests reposent sur des tables de fréquences minimales à partir desquelles des observations peuvent être considérées comme suffisamment nombreuses pour qu'il soit raisonnable de penser qu'elles sont significatives, probablement solides et stables, et non simplement dues au hasard et à la chance.

Il existe un grand nombre de tests statistiques, chacun applicable à des cas de figure spécifiques. Quelques-uns d'entre eux seulement sont habituellement utilisés pour l'analyse en éducation et formation. Six catégories de ces tests sont présentées dans les chapitres suivants de cet ouvrage : les tests de comparaison de variances ; les tests de comparaison de moyennes ; les tests de comparaison de proportions ; le test de Mood pour la comparaison de médianes ; les tests du Khi-carré ; et les tests d'équivalence. L'application de tests de comparaison des variances, moyennes et médianes et de tests d'équivalence est pertinente s'il y a du sens à calculer des variances, moyennes et médianes, ce qui présuppose au minimum que les variables soient quantitatives. L'application de tests de comparaison de proportions et de tests du Khi-carré nécessite que les variables soient catégorielles. On présente ici d'abord le principe général des tests statistiques.

6.2. LOGIQUE GÉNÉRALE DES TESTS STATISTIQUES

Un test consiste à confronter deux hypothèses mutuellement exclusives : une hypothèse H_0 (également appelée hypothèse nulle) et une hypothèse H_A (hypothèse alternative, souvent notée aussi H_1). Il s'agit de voir s'il est possible de rejeter H_0. Si H_0 peut être rejetée, c'est donc H_A qui est vraie et par conséquent acceptée. En ce sens, ces tests sont aussi appelés « tests d'hypothèses ».

L'hypothèse nulle postule en général une absence de différence entre les termes comparés (que ces termes soient des variances, ou des moyennes, ou des proportions, ou des médianes, etc.). *En revanche, la formulation de l'hypothèse alternative peut varier*. On distingue trois principaux cas de figure :

- dans le premier cas de figure, on cherche essentiellement à savoir s'il y a ou non une différence significative entre les termes comparés. L'hypothèse alternative postule la différence. Le test est dit *bilatéral*. Hypothèse nulle et hypothèse alternative sont alors formulées suivant le schéma :

$$\begin{cases} H_0 : les\ deux\ termes\ comparés\ sont\ égaux \\ H_A : les\ deux\ termes\ comparés\ ne\ sont\ pas\ égaux \end{cases}$$

- dans le deuxième cas de figure, on cherche à savoir si le premier terme comparé est *inférieur* au second. L'hypothèse alternative postule que le premier terme

comparé est inférieur au second. Le test est dit *unilatéral à gauche*. Hypothèse nulle et hypothèse alternative sont alors formulées suivant le schéma :

$$\begin{cases} H_0 : les\ deux\ termes\ comparés\ sont\ égaux \\ H_A : terme\ 1 < terme\ 2 \end{cases}$$

- dans le troisième cas de figure, on cherche à savoir si le premier terme comparé est *supérieur* au second. L'hypothèse alternative postule que le premier terme comparé est supérieur au second. Le test est dit *unilatéral à droite*. Hypothèse nulle et hypothèse alternative sont alors formulées suivant le schéma :

$$\begin{cases} H_0 : les\ deux\ termes\ comparés\ sont\ égaux \\ H_A : terme\ 1 > terme\ 2 \end{cases}$$

Dans le présent ouvrage, sauf mention contraire, les tests sont toujours bilatéraux.

La décision de rejeter ou non H_0 repose sur le calcul d'une « statistique de test », c'est-à-dire d'une valeur définie par une formule dans laquelle sont prises en compte les propriétés statistiques de l'échantillon analysé (moyenne, écart-type, médiane, etc.). Les propriétés prises en compte varient évidemment d'un test à l'autre puisque chaque test a sa propre statistique de test, adaptée à l'objectif poursuivi (par exemple comparer des proportions, ou comparer des distributions, etc.). De façon générale, quel que soit le test, il existe une table répertoriant les différentes « valeurs critiques » à partir desquelles le test peut être

considéré comme concluant. Exécuter un test sur un échantillon revient donc à calculer une statistique de test sur cet échantillon puis à rechercher dans la table du test si la valeur obtenue est suffisante par rapport à une valeur critique pour que le test puisse être considéré comme concluant.

La valeur critique à considérer dépend en premier lieu du nombre d'observations pris en compte lors du calcul de la statistique de test. En effet, on peut intuitivement comprendre que la valeur critique qui indique si un test est concluant n'est pas la même si l'échantillon dont proviennent les observations comprend 10 individus ou plutôt 1000. Les tables de tests listent donc les valeurs critiques en tenant compte des tailles d'échantillons. Dans la plupart des tables cependant, ce n'est pas simplement la taille d'échantillon qui est prise en compte, mais le nombre de « degrés de liberté », c'est-à-dire uniquement un nombre suffisant[54] d'observations, ainsi que, le cas échéant, le nombre d'échantillons. La manière de calculer le nombre de degrés de liberté dépend du test spécifique considéré[55].

---

[54] En effet, connaitre un certain nombre de valeurs de l'échantillon suffit à déterminer les autres.

[55] Par exemple, le nombre de degrés de liberté est égal :

- à $(n-1)$ s'il s'agit de tests *t* de Student pour échantillon unique ou pour échantillons appariés (n est la taille d'échantillon) ;

La bonne valeur critique à prendre en considération dépend aussi du degré de confiance[56] que le chercheur assigne à son estimation. Le degré de confiance est représenté ici par le « seuil de significativité », qui en est le complément à 1 :

$$Seuil\ de\ significativité = 1 - degré\ de\ confiance$$

Le seuil de significativité permet de trancher lors du choix entre hypothèse nulle H_0 et hypothèse alternative H_A. Lors de ce choix, en effet, le chercheur est exposé à deux types de risques. Le premier est de rejeter l'hypothèse nulle alors qu'elle est vraie : c'est l'« *erreur de première espèce* » (parfois appelée aussi « *erreur de type 1* »). Le second risque est au contraire de ne pas rejeter l'hypothèse nulle alors qu'elle devrait l'être : c'est l'« *erreur de seconde espèce* » (ou « *erreur de type 2* »).

- à $(n-2)$ s'il s'agit d'un test de significativité du coefficient de corrélation de Pearson ;
- à 1 s'il s'agit d'un test de McNemar ou d'un test de Khi-carré d'ajustement ;
- à $(l-1)\times(c-1)$ s'il s'agit d'un test de Khi-carré d'indépendance ou d'homogénéité (l et c sont respectivement le nombre de lignes et le nombre de colonnes) ;
- etc.

[56] Voir Encadré 2.1, page 93.

Pour chaque valeur de la statistique du test, il peut être calculé une probabilité p (ou encore « *p-value* » ou « *p-valeur* ») de commettre l'erreur de première espèce, c'est-à-dire de se tromper en rejetant H_0 alors qu'elle est vraie. Le seuil de significativité permet de décider si p est trop élevée pour que l'hypothèse nulle puisse être rejetée. On considère que p est trop élevée si p est supérieure ou égale au seuil de significativité, et dans ce cas l'hypothèse nulle ne peut pas être rejetée. Elle peut l'être sinon.

Il n'y a pas de règle absolue quant au niveau de seuil de significativité à adopter. Le chercheur peut décider d'adopter un seuil de significativité très faible, par exemple $\frac{1}{10\,000}$. Dans ce cas, on ne rejette pas l'hypothèse nulle tant que p n'est pas inférieure à 0,0001. Mais, ce faisant, on risque de commettre l'erreur de seconde espèce. Inversement, on pourrait adopter un seuil de significativité plus élevé, par exemple 33%, mais alors on augmente le risque de commettre l'erreur de première espèce (dans cet exemple en acceptant l'hypothèse alternative même si $p = 32{,}9\%$, ce qui représente près d'une chance sur 3 de se tromper ce faisant).

En recherche en éducation et formation, la pratique la plus fréquente est de fixer le seuil de significativité à 5%. C'est aussi le cas dans d'autres disciplines, par exemple en psychologie. Dans certaines disciplines cependant, par exemple en économie, il n'est pas rare

de rencontrer des seuils de significativité plus restrictifs (par exemple 1% ou 1‰). Dans le présent ouvrage, sauf mention contraire, le seuil de significativité adopté est toujours de 5%[57].

De façon générale, les tables de test indiquent les valeurs critiques pour différents seuils de significativité. La bonne valeur critique à considérer

[57] Il importe de garder à l'esprit l'intérêt que présente l'existence d'une pratique dominante en la matière. Dans une discipline ou un domaine de recherche donnés, la référence commune à un même niveau de seuil de significativité contribue à la comparabilité des résultats : il y a clairement une différence entre deux résultats de recherche si l'un conclut à une significativité au seuil de 5% et l'autre à une non-significativité au même seuil. La comparabilité est moins évidente si les résultats sont significatifs dans les deux recherches, mais au seuil de 5% dans la première, et au seuil de 10% dans la deuxième. Cela étant, il ne faut pas perdre de vue que le niveau de seuil de significativité retenu, quel qu'il soit, est une convention arbitraire. La pratique des sciences de l'éducation, ou de la psychologie, etc., auraient pu retenir un seuil de 4% ou un seuil de 6%, par exemple. En d'autres termes, le seuil de significativité ne doit pas être interprété comme une barrière départageant le « vrai » et le « faux ». Une différence non significative au seuil de 1% ne veut pas dire qu'il n'y a pas de différence du tout, et n'empêche pas une éventuelle significativité au seuil de 5% ou de 10%. Les seuils de significativité doivent plutôt être compris comme des outils indiquant ce qui est le plus plausible. Il faut donc veiller, à l'issue de tests, à interpréter ses résultats et à formuler ses conclusions de façon circonspecte et nuancée, et non pas excessivement péremptoire, tranchée et définitive.

pour juger si un test est concluant est donc, *in fine*, celle qui correspond à la fois au nombre approprié d'observations ou de degrés de liberté et au seuil de significativité adopté.

Chapitre 7. Les tests de comparaison de variances

La variance est généralement mobilisée dans l'analyse de problématiques impliquant la variabilité, la diversité, les inégalités, l'instabilité, la volatilité, par exemple. En arrière-plan figurent souvent des notions comme celles d'écart à la norme, égalité, équité, perturbation, ou risque, par exemple.

De façon générale, la comparaison de variances a lieu d'être lorsque l'on est en présence de groupes caractérisés chacun par un certain niveau de variance au regard d'un ou plusieurs caractères. On distingue trois principaux cas de figures : comparaison d'une variance à une référence ; comparaison de deux variances ; et comparaison de plus de deux variances.

7.1. Cas 1 – Comparaison d'une variance à une référence : test de conformité

Soit une variable X (par exemple les notes dans une matière, ou bien le nombre de jours d'absence des enseignants, ou bien le nombre d'heures de formation continue par salarié, etc.). On suppose qu'il existe un niveau de référence pour cette variance (par exemple une hypothèse théorique, ou une norme, ou un constat de fait, etc.). On dispose d'un échantillon pour lequel

on cherche à savoir si la variance de la variable observée dans l'échantillon est conforme à la référence.

On applique un *test de conformité de variance*[58]. L'hypothèse nulle du test est que la variance de l'échantillon est égale à la référence. Les logiciels statistiques indiquent directement la probabilité p associée à la valeur de test obtenue (p-value), c'est-à-dire encore la probabilité de commettre l'erreur de première espèce. L'hypothèse nulle peut être rejetée si p est inférieure au seuil de significativité de 0,05 : on peut considérer dans ce cas que la variance de l'échantillon est significativement différente de la référence.

Une condition préalable à l'utilisation du test de conformité de variance est que la distribution de

---

[58] La statistique du test est définie par :

$$statistique\ du\ test\ de\ conformité\ de\ variance = (n-1) \times \frac{\sigma^2}{\sigma^{2^*}}$$

où

n est la taille de l'échantillon ;

σ^2 est la variance observée dans l'échantillon ;

σ^{2^*} est la variance de référence.

Dans XLSTAT, le test est accessible via la commande *Tests paramétriques / Test de la variance pour un échantillon*.

l'échantillon soit gaussienne. On vérifie la normalité par un *test de normalité*. Il en existe plusieurs[59], que les logiciels statistiques permettent d'exécuter aisément. Dans un test de normalité, la probabilité p s'interprète comme suit : la distribution est normale si p est *supérieure*[60] à 0,05.

EXEMPLE 7.1

Dans le cadre d'une étude sur la diversité sociale à l'école, on mesure la diversité en termes de différents indicateurs, dont notamment la variance des revenus des parents. On dispose de données sur les revenus des parents dans un échantillon de 40 élèves (Tableau 7.1). On observe que l'écart-type des revenus mensuels des

---

[59] Par exemple les tests de Shapiro-Wilk, Anderson-Darling, Jarque-Bera, et Lilliefors, disponibles dans XLSTAT via la commande *Tests de normalité* sous l'onglet *Description des données*. Le test d'Agostino-Pearson est accessible notamment via la fonction DPTEST du « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9). Par exemple, pour une série figurant dans la plage A1:A15 d'une feuille de calcul Excel, on inscrit dans la cellule B1 la formule

=DPTEST(A1:A15)

et on appuie sur la touche Entrée. La p-value du test d'Agostino-Pearson s'inscrit en B1. La série suit une distribution gaussienne si la p-value est supérieure à 0,05.

[60] Supérieure car dans un test de normalité, H_0 postule que la distribution est normale.

parents s'établit à 1787 dans l'échantillon alors que, dans la région, l'écart-type des revenus mensuels de la population s'établit à 1500. On veut savoir si la variance des revenus mensuels des parents dans l'échantillon est significativement différente de la référence régionale.

Tableau 7.1.
Revenu mensuel du chef de famille pour un échantillon de 40 élèves (EUR)

	Revenu mensuel du chef de famille
Élève 1	1213
Élève 2	1368
Élève 3	4970
Élève 4	4519
Élève 5	2328
Élève 6	5922
Élève 7	3846
Élève 8	2027
Élève 9	3878
Élève 10	3287
Élève 11	509
Élève 12	3103
Élève 13	3744
Élève 14	2441
Élève 15	5140
Élève 16	1179
Élève 17	4320

	Revenu mensuel du chef de famille
Élève 18	5656
Élève 19	3249
Élève 20	1902
Élève 21	3331
Élève 22	2321
Élève 23	1760
Élève 24	4172
Élève 25	2496
Élève 26	1213
Élève 27	6287
Élève 28	1794
Élève 29	2885
Élève 30	2154
Élève 31	2325
Élève 32	423
Élève 33	5993
Élève 34	5985
Élève 35	649
Élève 36	5136
Élève 37	5814
Élève 38	1208
Élève 39	4109
Élève 40	6420

Vérification de la normalité

Tous les tests de normalité indiquent que la variable « Revenus mensuels du chef de famille » dans l'échantillon est normalement distribuée :

Résultats des tests de normalité

p			
Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
0,069	0,128	0,165	0,300

Test de conformité de la variance

La variance dans l'échantillon s'établit à 3 193 660, contre une variance de référence de 2 250 000. Le test de conformité de variance indique une probabilité $p = 0{,}087$. Donc, au seuil de significativité de 5%, on ne peut pas rejeter l'hypothèse nulle : il n'y a pas de différence significative entre la variance observée dans l'échantillon et la variance de référence.

7.2. CAS 2 – COMPARAISON DE DEUX VARIANCES

Soit une variable X (par exemple la distance école-domicile, ou bien le salaire du père, ou bien l'âge des élèves, etc.) observée dans deux groupes A et B distincts d'individus (par exemple d'élèves, d'enseignants, etc.) ; ou encore deux variables X_1 et X_2,

observées sur un même groupe. On suppose que le calcul des deux variances révèle une différence entre elles. On cherche à savoir si cette différence est ou non statistiquement significative.

On peut distinguer deux situations suivant que les variables sont normalement distribuées ou pas.

7.2.1. Les variables sont normalement distribuées

Chacune des variables X_1 et X_2 suit une distribution normale. Trois[61] principaux tests de comparaison des

[61] Un test alternatif pour comparer la variabilité de deux groupes normalement distribués est le test de comparaison de coefficients de variation (la notion de coefficient de variation est présentée page 44). Le test indique si la différence entre deux coefficients de variation est significative. L'hypothèse nulle du test postule l'égalité des coefficients de variation. Les coefficients de variation sont significativement différents si la p-value est inférieure au seuil de significativité. Le test est accessible via la fonction CV2TEST du « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9). Supposons par exemple qu'on veuille tester la différence entre coefficients de variation des séries X_1 et X_2. Il y a 30 observations dans chaque série. On inscrit les valeurs de la série X_1 dans la plage A1:A30 de la feuille Excel, puis les valeurs de X_2 dans la plage B1:B30. Puis on sélectionne la plage C1:D7 et on inscrit dans la fenêtre la formule

variances sont le plus souvent utilisés : les tests de Levene, Bartlett et F de Fisher[62]. Cependant le troisième, le test F de Fisher, n'est applicable que si les groupes / échantillons sont indépendants.

7.2.1.1. Test de Levene et test de Bartlett

Le principe consiste à calculer la statistique du test et à voir si elle indique que la différence des variances est significative. Pour le test de Levene comme pour le test de Bartlett, l'hypothèse nulle est que les différences de variances ne sont pas significatives[63]. Les logiciels statistiques indiquent directement la probabilité p de commettre l'erreur de première espèce. La probabilité p s'interprète comme suit : la différence des variances est statistiquement significative si p est inférieure à 0,05.

=CV2TEST(A1:A30,B1:B30,Vrai)

Puis on appuie en même temps sur les touches Ctrl Maj Entrée. Les résultats du test s'affichent. La p-value du test figure en cellule D5.

[62] Accessibles dans XLST via la commande *Tests paramétriques / Comparaison des variances de deux échantillons*.

[63] L'hypothèse nulle est $H_0 : \sigma_A^2 = \sigma_B^2$ (autrement dit les variances des deux échantillons sont égales) ; et l'hypothèse alternative du test bilatéral est $H_1 : \sigma_A^2 \neq \sigma_B^2$.

L'utilisation des tests de Levene et Bartlett nécessite que la condition de normalité soit remplie : chaque variable X doit être normalement distribuée[64]. L'indépendance des échantillons n'est pas requise.

EXEMPLE 7.2

Le nombre d'élèves en Centre de formation d'apprentis (CFA) est très variable d'une académie à l'autre. Le Tableau 7.2 montre ces effectifs en 1997 et 2012. On y voit que la variance du nombre d'apprentis a doublé pendant la période. On cherche à savoir si cette augmentation est significative.

Tableau 7.2.
Nombre d'élèves inscrits en CFA publics et privés – 1997 et 2012

Source : MENESR-DEPP,
http://www.education.gouv.fr/cid58535/series-chronologiques-de-donnees-statistiques-sur-le-systeme-educatif.html

	1997	2012
Aix-Marseille	18 050	18 472
Amiens	10 180	12 412
Besançon	7 719	10 228
Bordeaux	16 034	18 676

[64] Donc le résultat du test de normalité doit indiquer $p > 0{,}05$ puisque dans un test de normalité, H_0 postule que la distribution est normale.

	1997	2012
Caen	9 560	10 710
Clermont-Ferrand	7 988	8 963
Corse	1 250	2 080
Créteil	14 951	24 491
Dijon	10 826	10 777
Grenoble	13 519	20 330
Lille	15 856	22 589
Limoges	3 375	3 920
Lyon	15 089	22 908
Montpellier	10 968	16 227
Nancy-Metz	14 018	16 502
Nantes	25 064	29 435
Nice	10 303	12 387
Orléans-Tours	16 233	20 152
Paris	17 132	24 351
Poitiers	12 670	14 601
Reims	6 964	8 498
Rennes	15 363	18 348
Rouen	12 692	13 940
Strasbourg	12 546	15 511
Toulouse	14 071	17 963
Versailles	19 073	34 905
Guadeloupe	1 259	1 722
Guyane	210	702
Martinique	1 348	1 599
La Réunion	3 379	4 477
Variance	*36691227,1*	*72936200,5*

Tests de normalité

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
1997	0,230	0,155	0,360	0,825
2012	0,631	0,821	0,874	0,842

Tous les tests indiquent que les deux séries sont normalement distribuées.

Comparaison des variances

Résultats des tests de comparaison des variances

p	
Test de Levene	Test de Bartlett
0,082	0,069

Les deux tests indiquent une probabilité p supérieure à 5%. Il n'y a pas lieu de rejeter l'hypothèse nulle : comme l'indique le Tableau 7.2, la variance a doublé pendant la période, mais cette augmentation n'est pas statistiquement significative. La variabilité des effectifs académiques d'apprentis n'a donc pas significativement augmenté.

7.2.1.2. Test F de Fisher

Le test consiste à calculer la statistique de Fisher (notée F), et à voir si elle indique que la différence des variances est significative.

La statistique de Fisher se définit comme le rapport des deux variances à comparer :

$$F = \frac{\sigma_A^2}{\sigma_B^2} \quad , \quad o\text{ù} \quad \sigma_A > \sigma_B$$

La table de Fisher-Snedecor (voir Annexe 2) – ou les logiciels statistiques – indiquent les valeurs critiques. L'usage de la table nécessite de connaitre le nombre de degrés de liberté (DDL1 et DDL2) pris en compte dans l'estimation. La table distingue DDL1 (en colonne) et DDL2 (en ligne). DDL1 est le nombre de degrés de liberté correspondant au numérateur :

$$DDL1 = n_A - 1$$

où n_A est l'effectif de l'échantillon ayant la plus grande variance.

DDL2 est le nombre de degrés de liberté correspondant au dénominateur :

$$DDL2 = n_B - 1$$

où n_B est l'effectif de l'échantillon ayant la plus petite variance.

Dans le test F, l'hypothèse nulle est que les différences de variances ne sont pas significatives. La probabilité p s'interprète comme suit : la différence des variances est statistiquement significative si p est inférieure à 0,05. Dans la table, si la valeur de F calculée est inférieure à la valeur critique identifiée par l'intersection de DDL1 et DDL2, l'hypothèse nulle ne peut être rejetée.

L'utilisation du test F nécessite non seulement que les deux variables soient normalement distribuées, mais aussi que soit remplie la *condition d'indépendance des échantillons* : les deux groupes ne doivent avoir aucun lien entre eux. Par exemple, deux personnes susceptibles de s'influencer mutuellement ne peuvent faire partie l'une du groupe A et l'autre du groupe B.

EXEMPLE 7.3

Le nombre d'établissements publics dans le second degré varie beaucoup d'une académie à l'autre. On cherche à savoir si cette variabilité se retrouve aussi dans le secteur privé. Le Tableau 7.3 montre les chiffres pour 2012. Les variances des deux groupes y apparaissent très différentes. Mais cette différence est-elle statistiquement significative ?

Tableau 7.3.
Nombre d'établissements du second degré par académie à la rentrée 2012, secteur public et secteur privé

Source : Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche (2013), *Repères et références statistiques*, page 53.

	Public	Privé
Aix-Marseille	319	201
Amiens	262	96
Besançon	164	63
Bordeaux	391	179
Caen	216	102
Clermont-Ferrand	192	99
Corse	43	5
Créteil	523	141
Dijon	222	56
Grenoble	368	195
Lille	499	209
Limoges	120	24
Lyon	333	214
Montpellier	277	137
Nancy-Metz	336	113
Nantes	347	291
Nice	202	104
Orléans-Tours	324	106
Paris	218	175
Poitiers	239	75

	Public	Privé
Reims	196	57
Rennes	315	271
Rouen	245	69
Strasbourg	220	74
Toulouse	365	171
Versailles	625	219
Guadeloupe	68	26
Guyane	42	5
Martinique	66	14
Mayotte	27	1
La Réunion	118	11
Variance	*20 659,59*	*6 692,73*

Les deux groupes, établissements publics et établissements privés, sont indépendants puisqu'aucun établissement n'appartient simultanément aux deux groupes. On vérifie d'abord la normalité :

Tests de normalité des distributions d'établissements

	p			
	Test de Shapiro-Wilk	Test de Anderson-Darling	Test de Lilliefors	Test de Jarque-Bera
Public	0,316	0,464	0,818	0,536
Privé	0,129	0,212	0,375	0,433

Quel que soit le test appliqué et le groupe (public ou privé) considéré, la probabilité p est supérieure à 5%.

On en conclut que les deux séries sont normalement distribuées.

On vérifie ensuite la significativité de la différence des variances en appliquant le test F. La valeur calculée de F s'élève à 3,087 contre une valeur critique de 1,84 pour DDL1=DDL2=30, avec une probabilité p égale à 0,003 (ce qui signifie que le risque de se tromper en rejetant l'hypothèse nulle H_0 d'égalité des variances est de trois pour mille). On en conclut que la différence des variances est significative. Autrement dit, le nombre d'établissements d'une académie à l'autre est significativement plus stable dans le secteur privé qu'il ne l'est dans le secteur public.

7.2.2. Les variables ne sont pas normalement distribuées

En règle générale, lorsque les variables ne suivent pas une distribution gaussienne, on utilise des tests non-paramétriques. En effet, alors que la plupart des tests sont construits en faisant l'hypothèse que la distribution des données au sein des échantillons ou des populations-mères est d'un certain type, par exemple normale, les tests non-paramétriques ne font pas d'hypothèse sur la distribution des données. Ils peuvent

donc s'appliquer y compris lorsque la condition de normalité n'est pas remplie[65].

Lorsque les variables ne suivent pas une distribution gaussienne, on peut comparer leurs variances au moyen du test de Fligner-Killeen. Le test de Fligner-Killeen est un test non-paramétrique de comparaison des variances[66]. Dans sa version standardisée par centration, la statistique du test est calculée[67] par :

---

[65] En général, les tests non-paramétriques peuvent en outre s'appliquer à des séries de petite taille ($n < 30$), contrairement aux tests paramétriques dont l'utilisation nécessite le plus souvent des effectifs d'au moins trente observations.

[66] Voir en liste des références Fligner & Killeen (1976).

[67] Le test de Fligner-Killeen n'est pas systématiquement proposé par tous les logiciels statistiques. Il est cependant accessible via la fonction FKTEST du « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9). Supposons par exemple qu'on veuille tester l'homogénéité des variances de trois variables X_1, X_2, X_3. Il y a 30 observations dans chaque échantillon. On inscrit les valeurs de la variable X_1 dans la plage A1:A30 de la feuille Excel, puis les valeurs de X_2 dans la plage B1:B30, et enfin les valeurs de X_3 dans la plage C1:C30. Puis on sélectionne la cellule D1 et on y inscrit la formule

=FKTEST(A1:C30)

On appuie sur la touche Entrée. Dans la cellule D1 s'affiche la p-value du test de Fligner-Killeen. Il y a homogénéité des variances si la p-value est supérieure ou égale à 0,05. Il n'y a pas homogénéité dans le cas contraire.

$$FK = \frac{\sum_{j=1}^{k}\left[n_j \times \left(\overline{a_j} - \overline{a}\right)^2\right]}{\sigma^2}$$

où

- k est le nombre de groupes dont on veut comparer les variances ;
- n_j est la taille du groupe j ;
- $\overline{a_j}$ est la moyenne des valeurs standardisées centrées du groupe j ;
- $\overline{a}$ est la moyenne de l'ensemble des valeurs standardisées centrées tous groupes confondus ;
- σ^2 est la variance de l'ensemble des valeurs standardisées centrées tous groupes confondus.

À chaque valeur de FK correspond une probabilité p qui, si elle est inférieure à 0,05, indique que l'hypothèse nulle d'égalité des variances peut être rejetée.

EXEMPLE 7.4

On dispose (Tableau 7.4) de mesures du sentiment d'auto-efficacité de deux groupes d'étudiants, issus les uns de formations à pédagogie transmissive orientée vers l'acquisition de savoirs (cours magistral et examens sur table), et les autres de formations à pédagogie active orientée vers la construction de compétences (situations problèmes, apprentissage par projets et problèmes, études de cas, etc.). On veut comparer les deux groupes au regard de la variance du sentiment d'auto-efficacité.

Tableau 7.4.
Mesures du sentiment d'auto-efficacité de deux groupes d'étudiants formés l'un par pédagogie transmissive et l'autre par approche compétences

Pédagogie transmissive		Approche compétences	
Identifiant de l'étudiant	Degré de confiance exprimé par l'étudiant	Identifiant de l'étudiant	Degré de confiance exprimé par l'étudiant
1	1	31	27
2	19	32	96
3	6	33	89
4	4	34	15
5	20	35	88
6	25	36	15
7	5	37	40
8	2	38	99
9	11	39	25
10	11	40	22
11	44	41	25
12	31	42	98
13	43	43	94
14	31	44	30
15	44	45	69
16	48	46	46
17	54	47	99
18	15	48	69
19	1	49	80

Pédagogie transmissive		Approche compétences	
Identifiant de l'étudiant	Degré de confiance exprimé par l'étudiant	Identifiant de l'étudiant	Degré de confiance exprimé par l'étudiant
20	42	50	96
21	30	51	5
22	58	52	4
23	61	53	2
24	26	54	65
25	42	55	98
26	24	56	76
27	21	57	76
28	94	58	1
29	1	59	78
30	91	60	30
Variance	*1215,97*	*Variance*	*597,76*

On teste la normalité des distributions. Trois tests sur quatre indiquent que p est inférieure à 0,05 :

Tests de normalité des distributions

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Pédagogie transmissive	0,003	0,003	0,044	0,214
Approche compétences	0,004	0,004	0,048	0,223

On teste ensuite l'homogénéité des variances. Si, malgré la non-normalité des distributions, on applique les tests paramétriques de comparaison de variances, on obtient deux tests sur trois indiquant des variances homogènes :

Résultats des tests paramétriques de comparaison des variances

p		
Fisher	Levene	Bartlett
0,061	0,001	0,061

Au contraire, le test de Fligner-Killeen indique $p = 0{,}0089$, ce qui conduit à conclure que les variances sont en réalité non-homogènes.

7.3. CAS 3 – COMPARAISON DE PLUS DE 2 VARIANCES

Soit une variable X observée dans au moins trois groupes ; ou encore au moins trois variables X_1, X_2, X_3, observées sur un même groupe. On suppose que le calcul des variances révèle une différence entre elles. On cherche à savoir si cette différence est ou non statistiquement significative.

Si la condition de normalité n'est pas remplie, on applique un test de Fligner-Killeen dans les mêmes conditions que s'il s'agissait de comparer deux variances.

Si la condition de normalité est remplie, on applique un *test de comparaison des variances sur K échantillons* ($K \geq 3$). On utilise ici également le test de Levene et/ou le test de Bartlett, mais pas le test F de Fisher, qui ne permet de comparer des variances que prises deux à deux. Les tests de Levene et Bartlett s'utilisent ici suivant les mêmes principes que pour la comparaison de deux variances : dans les deux cas, il s'agit de calculer la statistique du test et de voir si elle indique que la différence des variances est significative. Dans les deux cas, l'hypothèse nulle est que les différences de variances ne sont pas significatives[68]. La probabilité p s'interprète comme précédemment : la différence des variances est statistiquement significative si p est inférieure à 0,05. Chacune des variables X doit être normalement distribuée.

[68] L'hypothèse nulle est $H_0 : \quad \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \cdots = \sigma_k^2$ (autrement dit, les variances des k échantillons sont égales) ; et l'hypothèse alternative est H_1 : il existe au moins un couple (i, j) d'échantillons tel que $\sigma_i^2 \neq \sigma_j^2$.

EXEMPLE 7.5

La dépense publique par étudiant est variable d'un pays à l'autre. On examine la dépense publique par étudiant aux niveaux licence et master dans les pays de l'OCDE pendant la période 2001-2005 (Tableau 7.5). On cherche à savoir si la différence de variance d'une année à l'autre est statistiquement significative.

Tableau 7.5.
Dépense publique par étudiant aux niveaux licence et master dans les pays de l'OCDE pendant la période 2001-2005 (en EUR)

Source : Eurostat (indicateur educ_bo_fi_ftot)

	2001	2002	2003	2004	2005
Belgique	10 237,9	10 480,5	10 007,4	9 621,1	10 117,4
Bulgarie	2 950,2	3 462,1	3 646,2	3 610	3 642,2
République Tchèque	5 086,7	5 311,6	5 914	5 583,4	5 624,3
Danemark	12 569,1	13 167,4	11 764,5	12 819,8	12 654,4
Allemagne	9 340	9 566	10 138,2	10 124,6	10 425,5
Irlande	8 492,7	8 366,7	7 940,4	8 509,6	8 855,5
Grèce	3 856	4 151,3	4 126,3	4 705,3	5 185,9
Espagne	6 577	6 941,9	7 520,1	7 871,1	8 534,8
France	8 679,2	9 116,8	8 789,3	8 870,7	9 301,5
Italie	7 276,1	6 978,6	7 087,1	6 415,6	6 785,6
Chypre	8 492,5	8 694,8	7 506,3	7 342	8 816,8
Lettonie	2 750,2	2 945	2 840,1	2 931,1	3 764,9
Lituanie	2 956,9	3 191,4	3 341,3	3 685,4	3 801,4
Malte	5 885,3	7 022,7	5 762,9	5 806,6	9 079,1

	2001	2002	2003	2004	2005
Pays-Bas	11 426,9	11 777,3	11 319,9	11 505	11 744,2
Autriche	9 639,4	10 828,2	11 017,6	11 891,4	12 813,4
Pologne	3 362	4 122,7	3 542,9	3 715,9	4 715,6
Portugal	4 220,2	3 983,5	4 428,8	4 651,7	6 244,4
Slovénie	7 383,8	6 216,1	5 803,9	6 241,8	7 080,5
Slovaquie	4 766,2	4 142,3	4 027	5 485,4	4 892,5
Finlande	7 831,7	9 689,2	9 811,1	10 525,4	10 390,2
Suède	13 210,9	13 448,7	13 534,2	13 775,2	13 489,7
Royaume-Uni	9 104,4	9 104,4	9 104,4	9 104,4	9 104,4
Variance	*9787795*	*10788422*	*10020209*	*10146385*	*10468749*

Tests de normalité

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
2001	0,353	0,540	0,793	0,613
2002	0,201	0,301	0,168	0,531
2003	0,290	0,363	0,411	0,519
2004	0,382	0,518	0,486	0,512
2005	0,265	0,407	0,562	0,558

Tous les tests indiquent que toutes les variables sont normalement distribuées.

Comparaison des variances

Les tests de Levene et Bartlett indiquent tous les deux une probabilité $p = 0{,}99$ (il y a 99% de chances de se tromper en rejetant l'hypothèse nulle d'égalité des variances). On peut donc conclure que les différences de variances ne sont pas statistiquement significatives.

7.4. TABLEAU RÉCAPITULATIF DES TESTS DE COMPARAISON DE VARIANCES

Le Tableau 7.6 ci-après récapitule les principaux tests de comparaison de variances pertinents suivant le cas de figure.

Tableau 7.6.
Principaux tests de comparaison de variances par cas de figure

Cas de figure	Condition de normalité	Test(s) applicable(s)
Comparaison de la variance d'un échantillon à une référence (*test de conformité*)	Doit impérativement être remplie pour que le test puisse être appliqué	Test de conformité de variance

Cas de figure	Condition de normalité	Test(s) applicable(s)
Comparaison de deux variances	Si la condition est remplie	*Si les échantillons sont appariés* : Test de Levene Test de Bartlett
		Si les échantillons sont indépendants : Test de Levene Test de Bartlett Test de Fisher
	Si la condition n'est pas remplie	Test de Fligner-Killeen
Comparaison de plus de deux variances	Si la condition est remplie	Test de Levene Test de Bartlett
	Si la condition n'est pas remplie	Test de Fligner-Killeen

Chapitre 8. Les tests de comparaison de moyennes

Les problématiques sous-jacentes à l'étude des moyennes sont variées elles aussi : égalité, équité, inégalités, mais aussi changement et évaluation. La comparaison de moyennes permet de savoir s'il y a eu, par exemple, changement significatif au cours du temps (du nombre d'élèves par enseignant, du coût par élève, etc.). Dans la perspective de l'évaluation, elle permet, par exemple, de mesurer l'efficacité pédagogique avant et après la mise en place d'un enseignement ou d'une réforme éducative.

D'un point de vue de pratique statistique, on distingue trois principaux cas de figure. Dans le premier cas, on compare la moyenne d'un groupe à une moyenne de référence. Dans le deuxième cas, la comparaison porte sur deux (ou plus) moyennes d'un même groupe, par exemple lorsqu'on étudie la différence de moyennes pour deux caractères distincts du même groupe (les résultats en maths et en français d'un même groupe d'élèves, par exemple) ; ou encore lorsqu'on examine la moyenne d'un même caractère du même groupe à deux moments différents (par exemple la moyenne avant et après que le groupe ait suivi un enseignement de remédiation). Dans le troisième cas enfin, ce sont des groupes indépendants que l'on compare.

8.1. CAS 1 – COMPARAISON D'UNE MOYENNE À UNE RÉFÉRENCE : TESTS DE CONFORMITÉ

Soit une variable *X* observée dans un groupe. Le cas de figure étudié ici est celui dans lequel la moyenne du groupe pour cette variable diffère d'une valeur de référence (par exemple une norme ou une valeur théorique). Par exemple, la moyenne ($\overline{x}$) des notes dans une matière donnée (variable *X*) des élèves d'une classe (ou d'un établissement, ou d'une académie, etc.) diffère d'un niveau de référence ($x^{\star}$). La question est de savoir si l'écart entre cette moyenne et la référence ($\overline{x} - x^{*}$) est statistiquement significatif.

Deux approches sont possibles suivant que l'on connait ou non la variance de la population-mère dont le groupe est tiré.

8.1.1. Situation 1 – On ignore la variance de la population-mère

L'approche pertinente dans cette situation consiste à appliquer un *test t de Student pour comparaison d'une moyenne à une référence*[69]. Le test consiste à calculer

[69] Accessible dans XLSTAT via la commande *Tests paramétriques / Tests t et z pour un échantillon*.

la statistique de Student (notée *t*)[70], et à voir si elle indique que l'écart de la moyenne à la référence est significatif.

Dans ce test, l'hypothèse nulle H_0 postule que la moyenne observée est égale à la valeur de référence ($\bar{x} = x^*$). Les logiciels statistiques indiquent la probabilité p associée à la valeur de t obtenue. La probabilité p s'interprète comme suit : l'écart de la moyenne à la référence est statistiquement significatif si p est inférieure à 0,05.

---

[70] La statistique de Student pour tester la conformité d'une moyenne d'échantillon à une valeur de référence est définie par :

$$t = \frac{\bar{x} - x^*}{\frac{\sigma}{\sqrt{n}}}$$

où

$\bar{x}$ est la moyenne de l'échantillon ;

x^* est la valeur de référence à laquelle est comparée la moyenne de l'échantillon ;

σ est l'écart-type de l'échantillon ;

n est la taille de l'échantillon.

On évalue la significativité de l'écart de la moyenne à la référence en comparant la valeur calculée de *t* (en valeur absolue) avec la valeur critique correspondant au seuil de significativité choisi pour $DDL = n - 1$ sur une table de Student (voir Annexe 3).

L'utilisation du test t nécessite que la condition de normalité soit remplie : la variable X étudiée doit être normalement distribuée[71].

EXEMPLE 8.1

On examine le nombre d'élèves par classe en lycées professionnels publics dans 21 (sur 26) académies métropolitaines en 2012-2013 (Tableau 8.1). Par comparaison, le nombre d'élèves par classe dans l'ensemble de l'enseignement secondaire public s'établissait à 21,9 en moyenne nationale la même année. On ignore la variance de l'effectif par classe en lycée professionnel public pour l'ensemble des académies métropolitaines. On veut savoir si la moyenne pour les lycées professionnels métropolitains est significativement différente de la situation nationale du second degré public.

[71] Rappelons que dans un test de normalité, la probabilité p s'interprète comme suit : la distribution est normale si p est *supérieure* à 0,05.

Tableau 8.1.
Nombre moyen d'élèves par classe en lycées professionnels publics dans 21 (sur 26) académies métropolitaines en 2012-2013

Source : Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche (2013) *Repères et références statistiques sur les enseignements, la formation et la recherche*, page 313 http://cache.media.education.gouv.fr/file/2013/49/9/DEPP-RERS-2013_266499.pdf.

Académie	Nombre moyen d'élèves par classe
Aix-Marseille	15,00
Amiens	16,80
Besançon	16,00
Bordeaux	15,70
Caen	16,20
Clermont-Ferrand	16,20
Dijon	15,50
Grenoble	15,80
Lille	16,40
Limoges	14,90
Montpellier	17,10
Nancy-Metz	15,90
Nantes	15,80
Orléans-Tours	16,50
Paris	15,40
Poitiers	15,70

Académie	Nombre moyen d'élèves par classe
Reims	15,80
Rennes	15,30
Strasbourg	15,00
Toulouse	15,80
Versailles	15,90
Moyenne	*15,84*

Tests de normalité

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Effectif par classe en lycée professionnel public en France métropolitaine	0,666	0,551	0,509	0,837

Tous les tests indiquent que la distribution suit une loi normale.

Comparaison de la moyenne à la référence

On ignore la variance de la population-mère, donc on utilise le *test t de conformité de moyenne*. La valeur calculée de la statistique du test s'élève à $-48{,}72$ pour une valeur critique de 2,086 au seuil de significativité de 5% avec DDL=20 ($DDL = n - 1$ dans un test *t* pour échantillon unique). La probabilité associée est $p < 0{,}0001$ (le risque de se tromper en rejetant l'hypothèse nulle d'égalité à la référence est inférieur à 0,01%). On en conclut que l'écart entre l'effectif en lycée professionnel public métropolitain et l'effectif moyen national en second degré public est significatif.

8.1.2. Situation 2 – On connaît la variance de la population-mère

L'outil pertinent dans cette situation est le *test z de conformité de moyenne*[72]. Le test consiste à calculer la statistique z et à voir si elle indique que l'écart de la moyenne à la référence est significatif[73].

---

[72] Accessible dans XLSTAT via la commande *Tests paramétriques / Tests t et z pour un échantillon*.

[73] La statistique z pour tester la conformité d'une moyenne d'échantillon à une valeur de référence est calculée par :

L'hypothèse nulle H_0 du test est que la moyenne observée est égale à la valeur de référence ($\bar{x} = x^*$). Les logiciels statistiques indiquent la probabilité p associée à la valeur de z obtenue. La probabilité p s'interprète comme suit : l'écart de la moyenne à la référence est statistiquement significatif si p est inférieure à 0,05.

L'utilisation du test z nécessite que la condition de normalité soit remplie : la variable X étudiée doit être normalement distribuée.

EXEMPLE 8.2

On examine le nombre moyen d'heures d'enseignement par semaine et par élève dans l'enseignement secondaire public :

$$\frac{Nombre\ hebdomadaire\ d'heures\ d'enseignement}{Nombre\ d'élèves}$$

$$z = \frac{\bar{x} - x^*}{\frac{\sigma_P}{\sqrt{n}}}$$

où

- $\bar{x}$ est la moyenne de l'échantillon ;
- x^* est la valeur de référence à laquelle est comparée la moyenne de l'échantillon ;
- σ_P est l'écart-type de la *population-mère* ;
- n est la taille de l'échantillon.

En 2012-2013, le chiffre était de 1,18 dans les collèges de France métropolitaine (Tableau 8.2) contre 1,36 en moyenne pour l'ensemble du second degré au niveau national. Cet écart est-il significatif ?

Tableau 8.2.
Nombre moyen d'heures d'enseignement par semaine et par élève dans les collèges publics de France métropolitaine en 2012-2013

Source : Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche (2013) *Repères et références statistiques sur les enseignements, la formation et la recherche*, page 313 http://cache.media.education.gouv.fr/file/2013/49/9/DEPP-RERS-2013_266499.pdf

Académie	Nombre d'heures
Aix-Marseille	1,18
Amiens	1,20
Besançon	1,18
Bordeaux	1,14
Caen	1,15
Clermont-Ferrand	1,20
Corse	1,25
Créteil	1,21
Dijon	1,19
Grenoble	1,14
Lille	1,23
Limoges	1,16

Académie	Nombre d'heures
Lyon	1,16
Montpellier	1,15
Nancy-Metz	1,21
Nantes	1,17
Nice	1,16
Orléans-Tours	1,16
Paris	1,18
Poitiers	1,17
Reims	1,18
Rennes	1,14
Rouen	1,19
Strasbourg	1,18
Toulouse	1,15
Versailles	1,16
Moyenne	*1,18*
Variance	*0,000753402*

Tests de normalité

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Nombre d'heures	0,094	0,166	0,166	0,244

Tous les tests indiquent que la distribution suit une loi normale.

Comparaison de la moyenne à la référence

On connaît la variance de la population-mère (calculable ici puisque l'échantillon est exhaustif), donc on utilise le *test z de conformité de moyenne*. La valeur calculée de la statistique de test s'élève à $-34{,}08$ contre une valeur critique de 1,96 au seuil de significativité de 5%. Le test indique une probabilité $p < 0{,}0001$ (le risque de se tromper en rejetant l'hypothèse nulle d'égalité à la référence est inférieur à 0,01%). On en conclut que le nombre moyen d'heures d'enseignement par semaine et par élève en collège public métropolitain est significativement inférieur à la moyenne nationale de l'enseignement secondaire.

8.2. CAS 2 – COMPARAISON DE DEUX (OU PLUS) MOYENNES D'UN MÊME GROUPE OU DE GROUPES APPARIÉS

Les tests diffèrent suivant que le nombre de moyennes à comparer est égal à deux ou supérieur à deux.

8.2.1. Comparaison de deux moyennes d'un même groupe ou de groupes appariés

Soit une variable X observée pour un même groupe à deux moments différents, ou pour deux groupes appariés (c'est-à-dire non indépendants), ou encore deux variables X_1 et X_2 d'un même groupe observées au même instant. Dans tous ces cas, les séries des deux variables sont de même effectif n. On suppose que le calcul des moyennes révèle une différence. On cherche à savoir si cette différence est ou non statistiquement significative.

Deux tests de comparaison des moyennes d'échantillons appariés sont applicables, le test *t* et le test *z*[74]. Tous deux reposent sur l'examen des différences entre variables. Dans chaque cas, le point de départ consiste à calculer, individu par individu, l'écart entre les deux variables. En résulte la série des différences entre variables, base de calcul du test. Le choix entre test *t* et test *z* dépend du point de savoir si l'on connait ou pas la variance de la série des différences pour l'ensemble de la population-mère. Le test *z* s'applique si le chercheur connait cette variance

[74] Accessibles dans XLSTAT via la commande *Tests paramétriques / Tests t et z pour deux échantillons*.

(ce qui, en principe, permet de gagner en précision), le test *t* s'il ne la connait pas[75].

8.2.1.1. Situation 1 – On ne connait pas la variance de la série des différences pour l'ensemble de la population-mère

On applique le *test t de Student pour échantillons appariés*[76]. L'hypothèse nulle est que les moyennes des

---

[75] En pratique toutefois, il est rare que cette variance soit connue, de sorte que c'est le test *t* qui est, en réalité, le plus souvent appliqué.

[76] La logique du test implique de calculer d'abord, pour chaque individu de l'échantillon, l'écart entre la valeur dans la première série et la valeur dans la seconde série (par exemple, si les séries concernent les notes avant et après remédiation, il s'agit de calculer pour chaque élève l'écart entre note avant et note après). En résulte une nouvelle série, la série des différences, à partir de laquelle est définie la statistique *t* :

$$t = \frac{\bar{x} - \Delta}{\frac{\sigma}{\sqrt{n}}}$$

où

$\bar{x}$ est la moyenne de la série des différences ;

Δ est la différence moyenne de référence, à laquelle est comparée la moyenne de la série des différences. C'est la différence moyenne attendue dans l'hypothèse nulle. Donc en général $\Delta = 0$;

σ est l'écart-type de la *série des différences* ;

deux variables sont égales. Les logiciels statistiques indiquent la probabilité p associée à la valeur de t obtenue. La probabilité p s'interprète comme suit : la différence des moyennes est statistiquement significative si p est inférieure à 0,05.

L'utilisation du test t pour échantillons appariés nécessite que la condition de normalité soit remplie : chaque variable X étudiée doit être normalement distribuée. Lorsque la condition de normalité n'est pas remplie, l'alternative consiste à utiliser des tests non-paramétriques de « *comparaison des distributions d'échantillons* ». S'agissant ici d'échantillons appariés, deux tests sont possibles : le *test du signe* et le *test de Wilcoxon signé*[77]. Les tests de comparaison d'échantillons ne comparent pas directement les moyennes des échantillons, mais plutôt la structure des données dans les deux échantillons, et indiquent la probabilité p que les éventuelles différences de structure soient significatives. Dans les deux tests, l'hypothèse nulle est que les échantillons suivent la même loi de distribution. Une différence de

---

n est la taille de la série des différences.

On évalue la significativité de l'écart entre moyennes en comparant la valeur calculée de t (en valeur absolue) avec la valeur critique correspondant au seuil de significativité choisi pour $DDL = n - 1$ sur une table de Student (voir Annexe 3).

[77] Accessibles dans XLSTAT via la commande *Tests non paramétriques / Comparaison de deux échantillons*.

distributions suggère qu'une différence observée entre moyennes pourrait être significative.

EXEMPLE 8.3.
Test *t* de comparaison de moyennes pour échantillons appariés

On examine la dépense moyenne d'éducation par élève ou étudiant dans les premier et second degrés et dans l'enseignement supérieur. La dépense moyenne d'éducation est définie par le rapport

$$\frac{dépense\ totale\ d'éducation}{nombre\ d'élèves\ et\ étudiants}$$

La dépense totale d'éducation comprend l'ensemble des dépenses consenties pour l'éducation par l'État, les collectivités publiques, les établissements d'enseignement publics et privés, et les usagers de l'école (infrastructures, personnels, fonctionnement, matériels scolaires, transports, cours particuliers, etc.). On dispose des données pour la période 1980-1998 (Tableau 8.3). La dépense moyenne varie suivant le niveau d'enseignement considéré. Elle est la plus faible dans l'enseignement du premier degré, chaque année et en moyenne annuelle. On cherche à savoir si l'écart entre dépense moyenne par élève dans le premier degré et dépense moyenne tous secteurs confondus est statistiquement significatif. La période examinée est un échantillon, on ignore la variance de la dépense

moyenne dans la population-mère que constituerait un ensemble plus complet de périodes.

Tableau 8.3.
Dépense moyenne d'éducation par élève ou étudiant pour les trois niveaux d'enseignement en France métropolitaine – 1980-1998 – En Euros 2012

Source : MENESR-DEPP,
http://www.education.gouv.fr/cid58535/series-chronologiques-de-donnees-statistiques-sur-le-systeme-educatif.html

	Premier degré	Second degré	Supérieur	Ensemble
1980	3 050	6 090	7 760	4 600
1981	3 200	6 260	7 800	4 770
1982	3 400	6 540	7 990	5 030
1983	3 430	6 530	7 930	5 070
1984	3 560	6 530	7 990	5 170
1985	3 620	6 520	8 300	5 240
1986	3 510	6 500	8 280	5 190
1987	3 560	6 530	8 310	5 240
1988	3 650	6 710	8 340	5 370
1989	3 740	6 950	8 210	5 520
1990	3 790	7 280	8 500	5 720
1991	3 930	7 540	8 620	5 940
1992	4 080	7 970	8 710	6 240
1993	4 240	8 190	8 710	6 440
1994	4 370	8 320	8 620	6 560

	Premier degré	Second degré	Supérieur	Ensemble
1995	4 490	8 530	8 740	6 720
1996	4 570	8 650	8 860	6 820
1997	4 700	8 830	9 080	6 980
1998	4 880	9 020	9 310	7 170
Moyenne annuelle	*3 883*	*7 342*	*8 424*	*5 778*

Tests de normalité

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Premier degré	0,407	0,302	0,328	0,552
Ensemble	0,141	0,106	0,161	0,448

Tous les tests indiquent que les deux séries sont normalement distribuées.

Test t de comparaison des moyennes

On ignore les variances dans les populations-mères, donc on applique un *test t pour échantillons appariés*. La valeur calculée de la statistique du test s'élève à $-29{,}01$ contre une valeur critique de 2,10 pour 18

DDL[78] au seuil de significativité de 5%. La probabilité p associée est inférieure à 0,0001. Il y a donc moins de 0,01% de risque de se tromper en rejetant l'hypothèse nulle H_0. On peut conclure que la dépense annuelle moyenne d'éducation par élève dans le premier degré était, entre 1980 et 1998, significativement inférieure à la moyenne de l'ensemble du système éducatif.

EXEMPLE 8.4.
Test du signe et test de Wilcoxon signé

Dans le cadre d'une préparation à un concours, on compare les résultats obtenus par les étudiants lors de deux concours blancs à quatre mois d'intervalle. Le Tableau 8.4 présente les notes d'un échantillon de trente étudiants, on ignore la variance des notes dans la population-mère.

Tableau 8.4.
Notes d'étudiants à deux concours blancs

Identifiant étudiant	Concours 1	Concours 2
1	6	10
2	1	20
3	16	19
4	7	8
5	12	19
6	15	8
7	19	13

[78] $DDL = n - 1$ dans un test *t* pour deux échantillons appariés.

Identifiant étudiant	Concours 1	Concours 2
8	16	20
9	15	10
10	8	9
11	20	10
12	16	20
13	19	19
14	14	11
15	18	17
16	1	14
17	15	20
18	14	17
19	14	18
20	17	20
21	5	4
22	17	4
23	4	3
24	16	16
25	15	20
26	17	17
27	15	17
28	2	2
29	20	18
30	2	11
Moyenne	*12,53*	*13,8*

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Concours blanc 1	0,003	0,003	0,002	0,239
Concours blanc 2	0,001	0,000	< 0,0001	0,152

Trois tests sur quatre indiquent qu'aucune des deux séries n'est normalement distribuée. On applique donc le test du signe et le test de Wilcoxon signé pour comparer les deux groupes de notes.

Résultats des tests de comparaison des distributions

p	
Test du signe	Test de Wilcoxon signé
0,327	0,242

Les deux tests indiquent qu'on ne peut pas rejeter l'hypothèse nulle de similarité des distributions. Rien ne suggère que la différence de moyennes observée est significative.

8.2.1.2. Situation 2 – On connaît la variance de la série des différences pour l'ensemble de la population-mère

On applique le *test z pour échantillons appariés*[79]. Le test consiste à calculer la statistique z et à voir si elle indique que la différence des moyennes est significative.

L'hypothèse nulle est que les moyennes sont égales. Les logiciels statistiques indiquent les valeurs de z à partir desquelles la différence des moyennes peut être considérée comme significative. La probabilité p s'interprète comme suit : la différence des moyennes est statistiquement significative si p est inférieure à 0,05.

L'utilisation du test z pour échantillons appariés nécessite que la condition de normalité soit remplie. Lorsque ce n'est pas le cas, on utilise plutôt les tests

---

[79] La statistique *z* est définie par :

$$z = \frac{\bar{x} - \Delta}{\frac{\sigma_P}{\sqrt{n}}}$$

où

$\bar{x}$ est la moyenne de la série des différences (dans l'échantillon) ;

Δ est la différence moyenne de reference, attendue dans l'hypothèse nulle (en général $\Delta = 0$) ;

σ_P est l'écart-type de la série des différences *pour l'ensemble de la population-mère* ;

n est la taille de la série des différences (dans l'échantillon).

non-paramétriques de comparaison des distributions : le *test du signe* et le *test de Wilcoxon signé*, comme dans le cas de distribution non-gaussienne de deux échantillons appariés dont on ignore la variance.

EXEMPLE 8.5

On dispose des notes obtenues par des étudiants de classe préparatoire à deux concours d'entrée en école d'ingénieur (Tableau 8.5). On veut savoir si l'écart de moyenne entre les deux concours est significatif. L'échantillon est exhaustif.

Tableau 8.5.
Notes des étudiants d'une classe préparatoire à deux concours d'entrée en école d'ingénieur

Identifiant étudiant	Concours 1	Concours 2
1	3	15
2	2	3
3	15	1
4	19	6
5	7	17
6	0	3
7	11	8
8	13	7
9	20	18
10	19	2
11	16	5
12	5	11

Identifiant étudiant	Concours 1	Concours 2
13	0	10
14	16	19
15	11	5
16	18	5
17	18	19
18	3	9
19	9	10
20	7	9
21	17	6
22	18	20
23	8	5
24	2	15
25	15	8
26	6	11
27	0	0
28	15	5
29	12	6
30	12	15
Moyenne	*10,56*	*9,1*

Tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Concours 1	0,033	0,063	0,082	0,302
Concours 2	0,065	0,052	0,153	0,359

Trois tests sur quatre indiquent que les deux distributions sont normales.

Test z des différences de moyennes pour échantillons appariés

L'échantillon étant exhaustif, la variance de la série des différences est la même pour l'échantillon et la population-mère. La valeur calculée de la statistique z s'élève à 0,98 contre une valeur critique de 1,96 au seuil de significativité de 5%. La probabilité associée est $p = 0{,}324$. On peut donc conclure que la différence des moyennes n'est pas significative.

8.2.2. Comparaison de trois moyennes (ou plus) d'un même groupe ou de groupes appariés

Si le nombre de moyennes à comparer est supérieur à deux, on effectue une analyse de variance[80]. Alternativement, on peut appliquer un *test de Friedman pour K échantillons*[81] ($K > 2$). Le test de Friedman est non-paramétrique. Il ne compare pas directement les moyennes, mais indique dans quelle mesure toutes les valeurs relatives à un même individu occupent le même rang dans les différentes séries comparées. Si le rang des valeurs diffère significativement, les différentes

[80] Voir chapitre 15.

[81] Accessible dans XLSTAT via la commande *Tests non paramétriques / Comparaison de k échantillons*.

séries sont considérées appartenir à des populations différentes, et les éventuelles différences de moyennes observées peuvent être considérées comme significatives. L'hypothèse nulle du test est que les séries comparées proviennent de la même population. L'hypothèse alternative est acceptée si la probabilité p de se tromper en rejetant l'hypothèse nulle est inférieure à 5%.

EXEMPLE 8.6

On dispose de données sur le nombre d'étudiants dans les tranches d'âge 50-54 ans, 55-59 ans et 60-64 ans dans l'enseignement supérieur dans 16 pays de l'Union européenne en 2016 (Tableau 8.6). On examine la significativité des différences entre les moyennes des tranches d'âge.

Tableau 8.6.
Nombre d'étudiants dans les tranches d'âge 50-54 ans, 55-59 ans et 60-64 ans dans l'enseignement supérieur dans 16 pays de l'Union européenne en 2016

Source : Eurostat, série educ_uoe_enrt02, http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=educ_uoe_enrt02&lang=en

	50-54 ans	55-59 ans	60-64 ans
Autriche	5739	2603	1217
Bulgarie	1672	609	169
Chypre	568	183	42
Croatie	442	131	21

	50-54 ans	55-59 ans	60-64 ans
Espagne	25934	11638	4133
Finlande	6789	3316	1331
Hongrie	3175	1029	231
Irlande	3279	1764	725
Italie	15286	7303	2604
Lettonie	739	343	91
Lituanie	524	156	20
Malte	168	81	43
Pays-Bas	12533	6478	2501
Pologne	9826	3932	1286
Portugal	4422	2164	773
Suède	9851	5026	1929
Moyenne	*6309*	*2922*	*1070*

Trois tests de normalité sur quatre indiquent que les séries ne suivent pas une distribution gaussienne :

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
50-54 ans	0,002	0,006	0,074	0,001
55-59 ans	0,002	0,005	0,058	0,009
60-64 ans	0,002	0,003	0,045	0,027

On applique donc le *test de Friedman pour K échantillons*. Les résultats indiquent une probabilité p inférieure à 0,0001. Donc les différences entre effectifs moyens des trois tranches d'âge peuvent être considérées comme significatives.

8.3. CAS 3 – COMPARAISON DE DEUX (OU PLUS) MOYENNES DE GROUPES INDÉPENDANTS

La procédure diffère suivant que le nombre de moyennes à comparer est égal ou supérieur à deux.

8.3.1. Comparaison de deux moyennes de groupes indépendants

Soit une variable X observée pour deux groupes indépendants A et B. On suppose que le calcul des moyennes révèle une différence. On cherche à savoir si cette différence est ou non statistiquement significative.

Le choix du test à appliquer dépend du point de savoir si l'on connaît ou non la variance de la variable pour chacune des populations-mères.

8.3.1.1. Situation 1 – On ignore la variance de la variable pour chacune des deux populations-mères

On applique un *test t de Student pour échantillons indépendants*[82]. La statistique du test diffère de celle du test pour échantillons appariés[83], mais les principes

[82] Accessible dans XLSTAT via la commande *Tests paramétriques / Tests t et z pour deux échantillons*.

[83] La statistique *t* est définie ici par :

$$t = \frac{\overline{x_A} - \overline{x_B} - \Delta}{\sqrt{\sigma^2 \times \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

avec

$$\sigma^2 = \frac{[(n_A - 1) \times c_A^2] + [(n_B - 1) \times c_B^2]}{n_A + n_B - 2}$$

où

$\overline{x_A}$ et $\overline{x_B}$ sont respectivement les moyennes de la variable pour le groupe A et le groupe B ;

Δ est la différence entre moyennes attendue dans l'hypothèse nulle (donc en général $\Delta = 0$) ;

σ^2 est la variance commune aux deux groupes (le test est en effet basé sur l'hypothèse que les deux groupes ont la même variance) ;

n_A et n_B sont respectivement les effectifs des groupes A et B ;

c_A^2 est le carré de la somme des valeurs de la série A ;

c_B^2 est le carré de la somme des valeurs de la série B.

On évalue la significativité de l'écart entre moyennes en comparant la valeur calculée de *t* (en valeur absolue) avec la

d'application restent les mêmes. L'hypothèse nulle est que la différence de moyennes est nulle. La probabilité p s'interprète de la même façon que dans le test apparié : la différence des moyennes est statistiquement significative si p est inférieure à 0,05.

L'utilisation du test *t* pour échantillons indépendants nécessite que deux conditions préalables soient remplies : (1) chaque variable X étudiée doit être normalement distribuée ; et (2) les variances doivent être « homogènes ». Les tests de Bartlett, Fisher et Levene permettent de vérifier cette seconde condition[84].

Lorsque la condition de normalité n'est pas remplie, on applique plutôt un *test U de Mann-Whitney*. Le test U de Mann-Whitney[85] est non-paramétrique. Il ne compare pas les moyennes mais indique si les structures des deux échantillons sont significativement différentes. L'hypothèse nulle du test est que la

---

valeur critique correspondant au seuil de significativité choisi pour $DDL = n_A + n_B - 2$ sur une table de Student (voir Annexe 3).

[84] Rappelons que l'hypothèse nulle des tests de comparaison de variances postule que les variances sont homogènes, donc la condition d'homogénéité des variances est remplie si $p > 5\%$.

[85] Accessible dans XLSTAT via la commande *Tests non paramétriques / Comparaison de deux échantillons*.

différence de position des échantillons est égale à 0. $p < 0{,}05$ indique une différence significative. Lorsque c'est le cas, on peut considérer que les différences de moyennes observées sont significatives.

Lorsque la condition de normalité est remplie mais pas la condition d'homogénéité des variances, on applique un *test t de Welch*[86]. Le test *t* de Welch est l'adaptation du test *t* de Student pour le cas où les variances des deux variables ne sont pas égales[87].

---

[86] La statistique *t* de Welch est définie par :

$$t = \frac{\overline{x_A} - \overline{x_B} - \Delta}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

où

$\overline{x_A}$ et $\overline{x_B}$ sont respectivement les moyennes de la variable pour le groupe A et le groupe B ;

Δ est la différence entre moyennes attendue dans l'hypothèse nulle (donc en général $\Delta = 0$) ;

σ_A^2 et σ_B^2 sont respectivement les variances de la variable pour le groupe A et le groupe B;

n_A et n_B sont respectivement les effectifs des groupes A et B.

[87] Certains logiciels appliquent directement le test *t de Welch* à la place du *test t de Student* lorsque les variances des variables ne sont pas égales. C'est le cas de XLSTAT.

EXEMPLE 8.7

Le taux de réussite au baccalauréat est variable d'une académie à l'autre. En juin 2014, la médiane des taux de réussite au bac technologique s'établissait à 91,3% (Tableau 8.7). Le nombre moyen d'admis par académie s'établissait à 4 172 dans les académies ayant un taux de réussite inférieur à la médiane, et à 4 338 dans les autres. On veut savoir si la différence de nombre moyen d'admis entre ces deux groupes d'académies est statistiquement significative.

Tableau 8.7.
Nombre d'admis au baccalauréat technologique – 2014

Source : MENESR-DEPP et MAAF, Résultats par série du baccalauréat technologique (session de juin 2014 : résultats provisoires), http://www.education.gouv.fr/cid56455/resultats-du-baccalaureat-session-de-juin-2014.html

Académies	Admis	Taux de réussite
Mayotte	749	58,8
Guyane	524	77,1
Amiens	3 694	86,9
La Réunion	2 400	88,0
Aix-Marseille	5 723	88,5
Créteil	9 284	88,8
Rouen	3 680	89,1
Versailles	11 304	89,1
Paris	3 439	89,3

Académies	Admis	Taux de réussite
Reims	2 470	89,4
Orléans-Tours	4 836	90,1
Nice	3 855	90,2
Dijon	2 913	90,4
Guadeloupe	1 171	90,6
Lyon	6 542	90,9
Nombre moyen d'admis par académie	*4 172*	
Limoges	1 265	91,3
Martinique	949	91,3
Besançon	2 215	91,4
Lille	8 433	91,7
Montpellier	5 377	91,7
Strasbourg	3 553	91,7
Toulouse	5 484	91,8
Clermont-Ferrand	2 238	91,9
Nancy-Metz	4 954	91,9
Bordeaux	5 726	92,1
Caen	2 624	92,5
Poitiers	3 140	93,2
Grenoble	6 057	93,8
Nantes	6 905	94,1
Rennes	6 874	94,5
Corse	537	95,2
Nombre moyen d'admis par académie	*4 338*	

Tests de normalité

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Académies à taux de réussite inférieur à la médiane	0,101	0,120	0,079	0,255
Académies à taux de réussite supérieur à la médiane	0,721	0,613	0,624	0,703

Tous les tests indiquent que les deux distributions suivent une loi normale.

Tests d'homogénéité des variances

Résultats des tests d'homogénéité des variances

p		
Fisher	Levene	Bartlett
0,355	0,842	0,355

Le test de Levene indique que les variances sont homogènes, ce que confirment les deux autres tests. On peut utiliser ici le test de Fisher car il n'y a que deux

moyennes à comparer et les échantillons sont indépendants.

Test t des différences de moyennes de deux échantillons indépendants

La valeur calculée de la statistique du test s'élève à $-0{,}16$ contre une valeur critique de 2,04 pour 28 DDL ($DDL = n_A + n_B - 2$; la médiane est exclue de la comparaison). La probabilité associée est $p = 0{,}868$. On peut donc conclure que la différence de nombre moyen d'admis entre les deux groupes d'académies n'est pas significative.

EXEMPLE 8.8

On dispose de données sur le nombre de diplômes en Sciences appliquées et en Sciences humaines délivrés par les établissements d'enseignement supérieur du Québec en 2015 (Tableau 8.8). On cherche à savoir si la différence en termes de nombre moyen de diplômes délivrés par établissement est significative.

Tableau 8.8.
Nombre de diplômes en Sciences appliquées et en Sciences humaines délivrés par les établissements d'enseignement supérieur du Québec en 2015

Source : Gouvernement du Québec, http://www.education.gouv.qc.ca/references/indicateurs-et-statistiques/statistiques/diplomation/

	Sciences appliquées	Sciences humaines
École de technologie supérieure	4690	–
École des hautes études commerciales de Montréal	74	–
École nationale d'administration publique	7	–
École Polytechnique de Montréal	1982	14
Institut national de la recherche scientifique	11	24
Télé-université	52	189
Université Bishop's	17	184
Université Concordia	1545	1867
Université de Montréal	879	4253
Université de Sherbrooke	1033	988
U. du Québec à Chicoutimi	451	256
U. du Québec à Montréal	530	2504
U. du Québec à Rimouski	47	105
Université du Québec à Trois-Rivières	258	755
Université du Québec en Abitibi-Témiscamingue	21	204
U. du Québec en Outaouais	48	415
Université Laval	2177	2242
Université McGill	1437	1704
Moyenne	*847,72*	*1046,93*

Trois des quatre tests de normalité indiquent qu'aucune des deux séries ne suit une distribution gaussienne :

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Sciences appliquées	0,000	0,000	0,006	< 0,0001
Sciences humaines	0,005	0,006	0,033	0,092

On applique donc un *test U de Mann-Whitney*. Le test indique une probabilité $p = 0{,}396$. On ne peut donc rejeter l'hypothèse nulle suivant laquelle la différence de position entre les deux échantillons est égale à 0. On ne dispose pas d'élément suggérant que la différence de moyennes est significative.

8.3.1.2. Situation 2 – On connaît la variance de la variable pour chacune des deux populations-mères

La variance de la variable, commune aux deux populations-mères, est connue. On effectue un *test z pour échantillons indépendants*[88]. L'hypothèse nulle du test est que les moyennes sont égales. La différence des moyennes est statistiquement significative si p est inférieure à 0,05. Chacune des séries de la variable doit être normalement distribuée, et les variances des deux séries doivent être homogènes. Le *test U de Mann-Whitney* s'applique si la condition de normalité n'est

---

[88] La variance de la variable, commune aux deux populations-mères, est connue. Elle est notée σ_P^2 et remplace σ^2 de sorte que la statistique *z* est définie par :

$$z = \frac{\overline{x_A} - \overline{x_B} - \Delta}{\sqrt{\sigma_P^2 \times \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

où

$\overline{x_A}$ et $\overline{x_B}$ sont respectivement les moyennes de la variable pour le groupe A et le groupe B ;

Δ est la différence entre moyennes attendue dans l'hypothèse nulle (donc en général $\Delta = 0$) ;

n_A et n_B sont respectivement les effectifs des groupes A et B.

Le test est accessible dans XLSTAT via la commande *Tests paramétriques / Tests t et z pour deux échantillons*.

pas remplie. Le *test z de Welch*[89] remplace le test z si les variances des populations-mères ne sont pas égales.

EXEMPLE 8.9

On examine la répartition par discipline des étudiants étrangers originaires d'Europe dans les universités françaises en 2012-2013 (Tableau 8.9). On compare la moyenne par nationalité pour 12 nationalités. La moyenne s'établit à 400,5 en Sciences économiques – AES et à 515 en Sciences – Staps. La variance des variables dans les populations-mères est connue. On cherche à savoir si cette différence de moyennes est statistiquement significative.

---

[89] La statistique z de Welch est définie par :

$$z = \frac{\overline{x_A} - \overline{x_B} - \Delta}{\sqrt{\frac{\sigma_{PA}^2}{n_A} + \frac{\sigma_{PB}^2}{n_B}}}$$

où

$\overline{x_A}$ et $\overline{x_B}$ sont respectivement les moyennes de la variable pour le groupe A et le groupe B ;

Δ est la différence entre moyennes attendue dans l'hypothèse nulle (donc en général $\Delta = 0$) ;

σ_{PA}^2 et σ_{PB}^2 sont les variances (connues et inégales) de la variable pour les populations-mères des groupes A et B ;

n_A et n_B sont respectivement les effectifs des groupes A et B.

Tableau 8.9.
Répartition par discipline des étudiants étrangers originaires d'Europe dans les universités françaises en 2012-2013

Source : MENESR (2013), *Repères et références statistiques sur les enseignements, la formation et la recherche*, page 203
http://cache.media.education.gouv.fr/file/2013/49/9/DEPP-RERS-2013_266499.pdf

	SE et AES	Sciences et Staps
Allemagne	750	891
Italie	441	1138
Espagne	388	801
Roumanie	475	537
Royaume-Uni	177	379
Belgique	315	479
Portugal	400	456
Pologne	228	350
Grèce	90	325
Bulgarie	358	170
Luxembourg	281	160
Russie	903	494
Moyenne par nationalité	*400,5*	*515*
Variances des populations-mères	*52865,36*	*85806,72*

Tests de normalité

Résultats des tests de normalité

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
SE et AES	0,218	0,175	0,170	0,400
Sciences et Staps	0,227	0,199	0,112	0,502

Les tests indiquent que les deux distributions sont normales.

Test d'homogénéité des variances

Résultats des tests d'homogénéité des variances

p		
Fisher	Levene	Bartlett
0,435	0,426	0,434

Les variances sont homogènes.

Comparaison des moyennes

La valeur calculée de la statistique z s'élève à -1,06 contre une valeur critique de 1,96 au seuil de significativité de 5%, avec une probabilité $p = 0{,}287$ de se tromper en rejetant l'hypothèse nulle. On peut donc conclure que la différence entre filières en termes de nombre moyen d'étudiants étrangers par nationalité n'est pas statistiquement significative.

8.3.2. Comparaison de trois (ou plus) moyennes de groupes indépendants

Si le nombre de moyennes à comparer est supérieur à deux, on effectue une analyse de variance[90]. Alternativement, on peut appliquer un *test de Kruskal-Wallis pour K échantillons*[91] ($K > 2$). Le test de Kruskal-Wallis est non-paramétrique. Il ne compare pas directement les moyennes, mais indique si les K échantillons peuvent être considérés comme provenant de populations ayant des caractéristiques identiques en termes de rang des observations. L'hypothèse nulle du test est que les séries comparées ont les mêmes caractéristiques. $p < 0{,}05$ indique des différences significatives.

EXEMPLE 8.10

On dispose de données sur le nombre d'admis au Brevet de technicien supérieur (Spécialités de la production) dans les académies de Paris, Créteil et Versailles en 2016 (Tableau 8.10). On cherche à savoir si les différences entre académies en termes de nombre moyens d'admis par filière sont significatives.

[90] Voir chapitre 15.

[91] Accessible dans XLSTAT via la commande *Tests non paramétriques / Comparaison de k échantillons*.

Tableau 8.10.
Nombre d'admis au Brevet de Technicien Supérieur et au BTSA (BTS agricole) – Spécialités de la production, académies de Créteil, Paris, Versailles (2016)

Source : Ministère de l'éducation nationale, Études et statistiques de la DEPP, Base de données de référence, http://www.education.gouv.fr/pid25496/etudes-et-statistiques-de-la-depp.html

	Créteil	Paris	Versailles
Technologies industrielles fondamentales	135	243	207
Technologies de commandes des transformations industrielles	228	116	256
Spécialités pluri technologiques des transformations	–	12	15
Agro-alimentaire, alimentation, cuisine	18	77	43
Transformations chimiques et apparentées	59	187	65
Métallurgie	19	8	–
Matériaux de construction, verre	–	16	–

	Créteil	Paris	Versailles
Plasturgie, matériaux composites	–	–	4
Energie, génie climatique	102	91	72
Spécialités pluri technologiques génie civil, construction…	201	249	99
Mines & carrières, génie civil, topographie	80	56	53
Bâtiment : construction	7	10	3
Bâtiment : finitions	5	26	–
Travail du bois	25	11	–
Textile	–	45	–
Habillement	16	145	25
Cuirs et peaux	–	17	–
Spécialités pluri technologiques mécanique-électricité	280	194	397
Moteurs, mécanique auto	86	–	141
Aéronautique & spatiale	–	1	43
Structures métalliques	33		37
Électricité, électronique	314	247	355
Nombre moyen d'admis par filière	*101*	*92*	*113*

Le test de Kruskal-Wallis indique une probabilité $p = 0{,}848$. On ne peut donc rejeter l'hypothèse nulle. Il n'y a pas de raison de considérer les différences observées entre académies en matière de nombre moyen d'admis par filière comme significatives.

8.4. TABLEAU RÉCAPITULATIF DES TESTS DE COMPARAISON DE MOYENNES

Le Tableau 8.11 ci-après récapitule les principaux tests de comparaison de moyennes pertinents suivant le cas de figure.

Tableau 8.11.
Principaux tests de comparaison de moyennes par cas de figure

Cas de figure	Conditions de normalité et de variance	Test(s) applicable(s)
Comparaison de la moyenne d'un échantillon à une référence (*test de conformité*)	La condition de normalité doit impérativement être remplie pour que les tests puissent être appliqués	*Si la variance de la population-mère est connue* : Test *z* de conformité
		Si la variance de la population-mère n'est pas connue : Test *t de Student* de conformité

<table>
<tr><th>Cas de figure</th><th colspan="2">Conditions de normalité et de variance</th><th>Test(s) applicable(s)</th></tr>
<tr><td rowspan="4">Comparaison de deux moyennes</td><td rowspan="4">Si la condition de normalité est remplie</td><td rowspan="2">Si la variance des populations-mère est connue</td><td>Si les échantillons sont appariés : Test z pour échantillons appariés</td></tr>
<tr><td>Si les échantillons sont indépendants
- Les variances sont homogènes : Test z pour échantillons indépendants
- Les variances sont inégales : Test z de Welch</td></tr>
<tr><td rowspan="2">Si la variance des populations-mères n’est pas connue</td><td>Si les échantillons sont appariés : Test t de Student pour échantillons appariés</td></tr>
<tr><td>Si les échantillons sont indépendants
- Variances homogènes : Test t de Student pour échantillons indépendants
- Variances inégales : Test t de Welch</td></tr>
</table>

Cas de figure	Conditions de normalité et de variance	Test(s) applicable(s)
	Si la condition de normalité n'est pas remplie	*Si les échantillons sont appariés* : - Test du signe - Test de Wilcoxon signé
		Si les échantillons sont indépendants : Test U de Mann-Whitney
Comparaison de plus de deux moyennes	Aucune condition de normalité	*Échantillons appariés* : - Test de Friedman pour k échantillons appariés (k>2) - Analyse de variance
		Échantillons indépendants : - Test de Kruskal-Wallis pour k échantillons indépendants (k>2) - Analyse de variance

Chapitre 9. Les tests de comparaison de proportions

Soit des variables qualitatives comprenant chacune au moins deux modalités. On cherche à comparer une proportion de modalité (c'est-à-dire la proportion dans laquelle une modalité est représentée) à une ou plusieurs autres. On distingue trois principaux cas de figure : comparaison à une référence, comparaison de deux proportions, ou comparaison de trois proportions ou plus. La méthode générale consiste à effectuer des tests de comparaison de proportions (ou « tests de proportions » ou « tests de fréquence »).

9.1. CAS 1 – COMPARAISON D'UNE PROPORTION OBSERVÉE À UNE PROPORTION THÉORIQUE

On observe sur un échantillon une certaine proportion d'une propriété, par exemple le taux de succès au bac dans un lycée, ou le pourcentage de redoublants dans une classe, etc. On cherche à savoir si cette proportion est significativement différente d'une norme.

On applique un *test z de conformité d'une proportion*[92] ou *test z pour une proportion*.

[92] La statistique du *test z de conformité d'une proportion* est définie par :

L'hypothèse nulle du test est que la proportion observée n'est pas différente de la référence. Il y a une différence significative si $p < 5\%$.

L'utilisation du test z de conformité d'une proportion nécessite qu'une condition préalable soit remplie : l'effectif de l'échantillon doit être au moins égal à 30.

EXEMPLE 9.1

En 2013, 391 467 garçons se sont présentés au diplôme national du Brevet en France, et 319 932 ont été admis, soit un taux de réussite de 81,7%. Le taux de réussite pour l'ensemble des candidats (garçons et filles) était de 84,5% (source : RERS 2013, page 239). On cherche à savoir si le taux de réussite des garçons est significativement différent de la performance nationale.

$$z = \frac{\rho - \rho^*}{\sqrt{\frac{\rho^* \times (1 - \rho^*)}{n}}}$$

où

ρ est la proportion observée dans l'échantillon ;

ρ^* est la proportion de référence par rapport à laquelle la comparaison est effectuée ;

n est la taille de l'échantillon.

Dans XLSTAT, le test est accessible via la commande *Tests paramétriques / Tests pour une proportion*.

La condition relative à la taille de l'échantillon est manifestement remplie. La valeur calculée de la statistique z s'élève à -47,95 contre une valeur critique de 1,96 au seuil de significativité de 5%. Le test indique une probabilité $p < 0{,}0001$. On peut donc conclure que la différence entre taux de réussite des garçons et taux de réussite général est significative.

9.2. CAS 2 – COMPARAISON DE DEUX PROPORTIONS

On distingue suivant que les échantillons sont indépendants ou appariés.

9.2.1. Situation 1 – Comparaison de deux proportions sur échantillons indépendants

Soit deux échantillons indépendants A et B sur lesquels on mesure la proportion d'une propriété. On constate que la proportion diffère d'un échantillon à l'autre. On cherche à savoir si cette différence est significative.

On peut effectuer cette comparaison au moyen d'un *test du Khi-carré d'ajustement* ou d'un *test du Khi-carré*

d'homogénéité[93]. Une autre approche consiste à appliquer un *test z pour deux proportions*[94].

L'hypothèse nulle du test z pour deux proportions est qu'il n'y a pas de différence significative entre les deux proportions. Il y a une différence significative si $p < 5\%$.

La condition préalable pour que le test z de comparaison de deux proportions puisse être appliqué est que l'effectif de chaque échantillon soit au moins égal à 30.

---

[93] Voir chapitre 11.

[94] La statistique du *test z pour deux proportions* est définie par :

$$z = \frac{\rho_A - \rho_B - \Delta}{\sqrt{\rho \times (1-\rho) \times \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \quad , \quad avec \quad \rho = \frac{(n_A \times \rho_A) + (n_B \times \rho_B)}{n_A + n_B}$$

où

ρ_A et $\rho_B$sont les proportions observées dans les groupes A et B respectivement ;

Δ est la difference de proportions attendue dans l'hypothèse nulle (en général $\Delta = 0$) ;

ρ est la proportion commune aux deux groupes ;

n_A et n_B sont les effectifs des groupes A et B respectivement.

Dans XLSTAT, le test est accessible via la commande *Tests paramétriques / Tests pour deux proportions*.

EXEMPLE 9.2

En 2005, 1 791 Diplômes universitaires de technologie (DUT) en Mesures physiques ont été délivrés, dont 20,5% à des femmes. En 2011, les femmes représentaient 22,5% des 1 852 diplômés (source : RERS 2013, page 259). On cherche à savoir si la part des femmes dans l'effectif annuel des lauréats de ce diplôme est significativement différente en 2011 par rapport à 2005.

La condition relative aux tailles d'échantillon est remplie. On applique le test z pour deux proportions. La valeur calculée de la statistique de test s'élève à -1,43 contre une valeur critique de 1,96 au seuil de significativité de 5%. Le test indique $p = 0{,}153$. Il n'y a donc pas lieu de rejeter l'hypothèse nulle, et on peut conclure que la part des femmes dans l'effectif annuel des lauréats de ce diplôme n'est pas significativement différente en 2011 par rapport à 2005.

9.2.2. Situation 2 – Comparaison de deux proportions sur échantillons appariés

Soit une propriété (par exemple l'obtention d'une note supérieure à la moyenne) dont on observe la proportion dans deux échantillons non-indépendants (par exemple, parmi les élèves d'une même classe, deux groupes distingués en fonction de l'âge ou du genre). Le

raisonnement est le même si on considère un unique échantillon à deux moments différents, par exemple une classe avant et après un enseignement de remédiation. On constate que la proportion de la propriété diffère d'un échantillon à l'autre ou d'une période à l'autre. On cherche à savoir si cette différence est statistiquement significative.

On peut effectuer cette comparaison au moyen d'un *test du Khi-carré d'ajustement* ou d'un *test du Khi-carré d'homogénéité*[95]. Une autre approche, dont le champ d'application est plus restreint cependant, consiste à appliquer un *test de McNemar*[96]. Le test de McNemar[97]

[95] Voir chapitre 11.

[96] Accessible dans XLSTAT via la commande *Tests non paramétriques / Test de McNemar*.

[97] Le test de McNemar a été initialement conçu pour détecter l'effet ou l'absence d'effet d'un traitement en comparant la situation de patients avant et après l'administration de ce traitement. Les données avant et après traitement peuvent être présentées sous forme d'un *tableau de contingence* à deux lignes et deux colonnes :

est non-paramétrique. Il importe cependant de noter qu'il ne s'applique que si la variable est binaire (donc n'admettant que deux modalités, par exemple « a obtenu une note au moins égale à la moyenne » et « n'a pas obtenu une note au moins égale à la moyenne »).

		Après traitement : nombre de sujets en situation	
		Favorable	Défavorable
Avant traitement : nombre de sujets en situation	Favorable	A	B
	Défavorable	C	D

Le test compare le nombre de sujets dont la situation s'est améliorée (C) au nombre de sujets dont la situation s'est dégradée (B). On appelle B et C les « discordants ». On considère l'effet du traitement comme positif si C>B, négatif dans le cas contraire, et nul si B=C. Mais encore faut-il que l'éventuel effet soit significatif. La statistique du test de McNemar est définie par :

$$statistique\ du\ test\ de\ McNemar = \frac{(B-C)^2}{B+C}$$

Si les effectifs sont suffisamment grands (c'est-à-dire si $(B + C) > 25$), la valeur calculée de la statistique s'interprète directement à partir de la table du Khi-carré (Annexe 4). L'effet est significatif si la valeur calculée de la statistique de test est supérieure à la valeur critique pour DDL=1 au taux de significativité choisi.

L'hypothèse nulle du test postule que la différence de proportions n'est pas significative. La différence est significative si $p < 5\%$.

EXEMPLE 9.3

Le Tableau 9.1 présente les résultats d'évaluation d'un groupe d'élèves avant et après enseignement de remédiation. Le pourcentage d'élèves ayant obtenu une note au moins égale à la moyenne est passé de 45% à 55%. On veut savoir si cette progression est significative.

Tableau 9.1.
Résultats d'évaluation d'un groupe d'élèves avant et après enseignement de remédiation

	A obtenu une note au moins égale à la moyenne	
	Avant remédiation	Après remédiation
Élève 1	Oui	Oui
Élève 2	Oui	Oui
Élève 3	Oui	Non
Élève 4	Non	Non
Élève 5	Oui	Non
Élève 6	Non	Non
Élève 7	Non	Oui
Élève 8	Non	Non
Élève 9	Non	Oui

	A obtenu une note au moins égale à la moyenne	
	Avant remédiation	Après remédiation
Élève 10	Oui	Oui
Élève 11	Non	Non
Élève 12	Non	Oui
Élève 13	Non	Oui
Élève 14	Oui	Oui
Élève 15	Oui	Non
Élève 16	Non	Oui
Élève 17	Non	Non
Élève 18	Non	Non
Élève 19	Oui	Oui
Élève 20	Oui	Oui
% de Oui	*45%*	*55%*

Le tableau de contingence[98] indique :

		Après remédiation	
		Oui	Non
Avant remédiation	Oui	6	3
	Non	5	6

[98] XLSTAT fournit les tableaux de contingence à partir de la commande *Créer un tableau de contingence* sous l'onglet *Préparation des données*.

Donc l'augmentation de la proportion d'élèves ayant une note au moins égale à la moyenne résulte d'une amélioration de la situation de cinq élèves et d'une détérioration de la situation de trois élèves.

La p-value du test de McNemar s'établit à 0,72. Donc on peut conclure que le cours de remédiation n'a pas eu d'effet statistiquement significatif, et que la différence entre pourcentages d'élèves ayant obtenu une note au moins égale à la moyenne avant et après remédiation n'est pas non plus significative.

Exemple 9.4.

On dispose de données sur les résultats en langues de 30 élèves avant et après redoublement (Tableau 9.2). On veut savoir si le redoublement a été efficace du point de vue particulier du niveau en langues.

Tableau 9.2.
Résultats au test de langues avant et après redoublement (0=Non satisfaisant ; 1=Satisfaisant)

Identifiant	Avant	Après
1	0	0
2	0	0
3	0	0
4	0	1
5	0	1
6	0	0
7	0	1

Identifiant	Avant	Après
8	0	1
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	1
16	0	0
17	0	0
18	1	1
19	1	1
20	0	1
21	0	0
22	0	0
23	0	1
24	0	0
25	1	0
26	0	1
27	0	0
28	0	1
29	0	0
30	0	0

Le tableau de contingence s'établit comme suit :

		Après redoublement	
		Satisfaisant	Non-satisfaisant
Avant redoublement	Satisfaisant	2	1
	Non-satisfaisant	9	18

Le nombre d'élèves dont la situation s'est améliorée (9) est supérieur au nombre d'élèves dont la situation s'est détériorée (1). La p-value du test de McNemar s'établit à 0,02. Donc l'évolution constatée est significative. On peut donc conclure que du point de vue du niveau en langues, le redoublement a eu un effet significativement positif.

9.3. CAS 3 – COMPARAISON DE PLUS DE DEUX PROPORTIONS

Soit différentes modalités d'une variable. On veut comparer les proportions de ces différentes modalités dans plusieurs échantillons (E_1, E_2, etc.). On peut effectuer cette comparaison au moyen d'un *test du Khi-*

carré d'homogénéité[99]. Une autre approche, dont le champ d'application est plus restreint cependant, consiste à appliquer un *test Q de Cochran*[100]. Le test Q de Cochran s'applique au cas dans lequel un échantillon est soumis à différents traitements avec seulement deux

[99] Voir chapitre 11.

[100] La statistique du test est définie par :

$$Q = T \times (T-1)\frac{\sum_{t=1}^{T}\left(C^t - \frac{S}{T}\right)^2}{\sum_{i=1}^{n}[L_i \times (T-L_i)]} \quad , \quad avec \quad S = \sum_{i=1}^{n}\sum_{t=1}^{T} x_i^t$$

où

T est le nombre de traitements ;

C^t est le total de la colonne t ;

x_i^t est la valeur de la variable pour l'individu i dans le traitement t ;

S est la somme des valeurs de la variable pour tous les individus à tous les traitements ;

n est le nombre d'individus soumis aux traitements ;

L_i est la somme des valeurs de la variable pour l'individu i au cours de l'ensemble des traitements : $L_i = \sum_{t=1}^{T} x_i^t$

La valeur calculée de la statistique Q s'interprète à partir de la table du Khi-carré (voir Annexe 4). Les proportions sont significativement différentes si la valeur calculée de la statistique de test est supérieure à la valeur critique pour DDL=(T-1) au taux de significativité choisi.

Le test Q de Cochran est accessible dans XLSTAT via la commande *Tests non paramétriques / Test Q de Cochran*.

modalités de réponse possibles (codées 0 ou 1) par individu et traitement, par exemple :

	Traitement 1	Traitement 2	…	Traitement T	Total
Individu 1	0	1	…	0	L_1
Individu 2	1	0	…	1	L_2
⋮	⋮	⋮	⋮	⋮	⋮
Individu n	0	1	…	0	L_n
Total	C^1	C^2	…	C^T	S

En termes plus généraux, le cas est celui dans lequel une même variable *binaire* est observée dans différents échantillons *de même taille*. L'objectif du test est de comparer la proportion de la modalité 1 dans les différents traitements / échantillons.

Le test Q de Cochran est non-paramétrique. L'hypothèse nulle postule l'égalité des proportions, l'hypothèse alternative qu'il y a au moins deux proportions différentes l'une de l'autre. La différence est significative si $p < 5\%$.

EXEMPLE 9.5

On examine les résultats d'une classe de Terminale aux trois bacs blancs réalisés pendant une année. On cherche à savoir si le pourcentage de mentions Bien et Très Bien varie significativement d'un examen blanc à l'autre.

Tableau 9.3.
Pourcentages de mentions au cours de trois bacs blancs

	A obtenu la mention Bien / Très Bien		
	Bac blanc 1	Bac blanc 2	Bac blanc 3
Élève 1	Oui	Non	Non
Élève 2	Non	Non	Non
Élève 3	Non	Non	Non
Élève 4	Oui	Oui	Oui
Élève 5	Oui	Non	Non
Élève 6	Non	Non	Non
Élève 7	Oui	Non	Non
Élève 8	Non	Non	Non
Élève 9	Oui	Oui	Oui
Élève 10	Non	Non	Non
Élève 11	Oui	Non	Oui
Élève 12	Non	Non	Non
Élève 13	Non	Non	Non
Élève 14	Non	Non	Non
Élève 15	Non	Oui	Oui
Élève 16	Oui	Non	Non
Élève 17	Non	Non	Non

	A obtenu la mention Bien / Très Bien		
	Bac blanc 1	Bac blanc 2	Bac blanc 3
Élève 18	Oui	Oui	Oui
Élève 19	Non	Non	Non
Élève 20	Non	Non	Oui

La proportion de mentions s'établit respectivement à 40% (examen 1), 20% (examen 2) et 30% (examen 3). La valeur calculée de la statistique du test s'élève à 3,43 contre une valeur critique de 5,99 pour DDL=2 au seuil de significativité de 5%. Le test indique aussi une probabilité $p = 0{,}180$. On en conclut que le pourcentage de mentions Bien ou Très Bien n'est pas significativement différent d'un examen blanc à l'autre.

9.4. TABLEAU RÉCAPITULATIF DES TESTS DE COMPARAISON DE PROPORTIONS

Le Tableau 9.4 ci-après récapitule les principaux tests de comparaison de proportions pertinents suivant le cas de figure.

Tableau 9.4.
Principaux tests de comparaison de proportions par cas de figure

Cas de figure	Conditions d'utilisation	Test(s) applicable(s)
Comparaison d'une proportion observée à une proportion théorique (*test de conformité*)	L'échantillon doit comporter au moins trente individus	Test *z* pour une proportion
Comparaison de deux proportions	Cas général	- Khi-carré d'homogénéité - Khi-carré d'ajustement
	- Échantillons appariés ; et - les variables sont binaires	Test de McNemar
	- Échantillons indépendants ; et - chaque échantillon comporte au moins trente individus	Test *z* pour deux proportions

Cas de figure	Conditions d'utilisation	Test(s) applicable(s)
Comparaison de plus de deux proportions	Cas général	Khi-carré d'homogénéité
	Les variables sont binaires	Test Q de Cochran

Chapitre 10. Test de comparaison de médianes

L'objectif ici est de comparer les médianes de plusieurs séries, que ces séries soient relatives à une ou à plusieurs variables d'échantillons appariés ou indépendants. On utilise à cette fin le *test de Mood*[101]. Le test de Mood est non-paramétrique.

---

[101] Soit *K* échantillons 1 à *K* de tailles égales ou variables n_1 à n_K. Pour appliquer le test de Mood, on commence par déterminer la médiane ζ de l'ensemble que constituent les *K* échantillons réunis. On dresse ensuite un tableau de contingence qui indique, pour chaque échantillon *k*, le nombre de valeurs supérieures ou inférieures à la médiane d'ensemble ζ :

	Échantillons			
	1	2	…	K
Nombre de valeurs supérieures à la médiane d'ensemble	A_1	A_2	…	A_K
Nombre de valeurs inférieures ou égales à la médiane d'ensemble	$n_1 - A_1$	$n_2 - A_2$	…	$n_K - A_K$

La statistique de Mood est définie par

$$M = 4 \times \sum_{k=1}^{K} \frac{\left(A_k - \frac{n_k}{2}\right)^2}{n_k}$$

L'hypothèse nulle du test postule l'égalité des médianes, l'hypothèse alternative étant que l'une au moins des médianes est différente des autres. Une probabilité $p < 0{,}05$ indique que l'hypothèse nulle peut être rejetée.

EXEMPLE 10.1

On compare les médianes des notes obtenues par une classe dans cinq matières (Tableau 10.1) :
- Français (F) ;
- Histoire-Géographie (HG) ;
- Maths ;
- Anglais ;
- Sport.

---

où n_k est la taille de l'échantillon *k*.

La valeur calculée de la statistique M de Mood s'interprète à partir de la table du Khi-carré (voir Annexe 4). Il existe au moins un couple de médianes significativement différentes si la valeur calculée de la statistique de test est supérieure à la valeur critique pour DDL=(K-1) au taux de significativité choisi.

Le test de Mood est accessible dans XLSTAT via la commande *Tests non paramétriques / Test de Mood*.

Tableau 10.1
Notes, moyennes et médianes d'une classe dans cinq matières

	F	HG	Maths	Anglais	Sport
Élève 1	9	15	0	15	14
Élève 2	3	14	19	6	15
Élève 3	5	7	20	6	12
Élève 4	6	5	2	15	11
Élève 5	5	19	14	8	8
Élève 6	18	8	3	7	14
Élève 7	8	4	19	12	10
Élève 8	16	9	19	15	10
Élève 9	6	15	0	13	15
Élève 10	16	3	15	12	15
Élève 11	4	8	18	10	16
Élève 12	13	7	11	7	13
Élève 13	12	7	19	11	9
Élève 14	8	5	12	9	13
Élève 15	12	16	10	7	15
Élève 16	9	16	16	13	16
Élève 17	12	13	14	11	12
Élève 18	13	11	10	8	13
Élève 19	7	4	0	6	10
Élève 20	17	4	2	10	16
Élève 21	17	3	8	15	16
Élève 22	15	13	3	11	14
Élève 23	9	10	0	9	16
Élève 24	7	12	11	8	13

	F	HG	Maths	Anglais	Sport
Élève 25	18	7	6	9	14
Élève 26	3	3	12	14	16
Élève 27	6	4	19	11	8
Élève 28	16	7	7	13	10
Élève 29	18	10	19	6	9
Élève 30	9	3	5	6	8
Moyenne	*10.56*	*8.73*	*10.43*	*10.1*	*12.7*
Médiane	*9*	*7.5*	*11*	*10*	*13*

La valeur calculée de la statistique de Mood s'élève à 8,533 contre une valeur critique de 9,48 pour DDL=4 au seuil de significativité de 5%. Le test indique aussi une probabilité $p = 0{,}078$. On ne peut donc rejeter l'hypothèse nulle. Les différences observées entre médianes ne sont pas significatives.

Chapitre 11. Les tests du Khi-carré

Soit des variables qualitatives. Il n'y aurait évidemment pas de sens à en calculer les moyennes ou variances. On peut en revanche en analyser et en comparer les structures, c'est-à-dire les proportions dans lesquelles les différentes catégories et combinaisons de catégories y sont représentées. C'est ce que permettent de faire les tests du Khi-carré.

On distingue trois types de tests du Khi-carré : le Khi-carré d'ajustement, le Khi-carré d'indépendance, et le Khi-carré d'homogénéité.

Le Khi-carré d'ajustement permet de vérifier s'il y a adéquation d'une distribution observée à une distribution de référence (on parle habituellement de « distribution théorique »). Le Khi-carré d'indépendance permet de vérifier s'il existe une relation entre deux variables qualitatives observées sur un même échantillon. Le Khi-carré d'homogénéité permet de vérifier si la structure des modalités d'une variable donnée – c'est-à-dire la part relative de chaque modalité dans l'effectif de l'échantillon – est identique dans plusieurs échantillons.

11.1. PRINCIPES GÉNÉRAUX DES TESTS DU KHI-CARRÉ

De façon générale, les tests du Khi-carré reposent sur l'analyse de la structure des échantillons au regard des variables. On distingue la structure observée et la structure théorique. La structure est représentée par un *tableau de contingence*. La structure observée d'un échantillon est donc représentée par un *tableau de contingence observé* :

Tableau de contingence observé

		Effectifs observés
Variable	Modalité 1	O_1
	Modalité 2	O_2
	⋮	⋮
	Modalité M	O_M
	Total	n

De même, la structure théorique de l'échantillon est représentée par un *tableau de contingence théorique*, dans lequel les effectifs théoriques remplacent les effectifs observés :

Tableau de contingence théorique

		Effectifs théoriques
Variable	Modalité 1	T_1
	Modalité 2	T_2
	⋮	⋮
	Modalité M	T_M
	Total	n

La statistique du Khi-carré est définie de façon générale par :

$$\chi^2 = \sum_{m=1}^{M} \frac{(O_m - T_m)^2}{T_m}$$

où
O_m est l'effectif observé pour la modalité m ; et
T_m est l'effectif théorique pour la modalité m.

On voit que le χ^2 est nul si tous les écarts entre effectifs observés et effectifs théoriques sont nuls, c'est-à-dire lorsque structure observée et structure théorique sont identiques. Au contraire, plus les écarts entre effectifs observés et effectifs théoriques sont importants, plus le χ^2 est élevé. L'idée est qu'au-delà d'un certain seuil, l'écart entre effectifs observés et effectifs théoriques devient significatif. Le test permet donc de comparer de cette façon la structure observée à la structure théorique. L'hypothèse nulle du test est que les

structures sont identiques. La table du Khi-carré indique les seuils au-dessus desquels la valeur de χ^2 calculée exprime un écart significatif.

La table du Khi-carré se présente comme suit (une version plus complète figure en Annexe 4) :

	Seuils de significativité		
	0,05	0,01	0,001
DDL			
1	3,84	6,64	10,83
2	5,99	9,21	13,82
3	7,82	11,35	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46
⋮	⋮	⋮	⋮

Le nombre de degrés de liberté se détermine comme suit :

- $DDL = nombre\ de\ catégories - 1$ quand le tableau de contingence se réduit à une seule ligne ou à une seule colonne ; et
- $DDL = (nombre\ de\ lignes - 1) \times (nombre\ de\ colonnes - 1)$ dans tous les autres cas, c'est-à-dire lorsque le tableau comprend au moins deux lignes et au moins deux colonnes.

L'hypothèse nulle peut être rejetée si la valeur de χ^2 obtenue est supérieure à la valeur critique correspondant au nombre de DDL et au seuil de significativité retenu.

Une condition empirique nécessaire à l'utilisation du test du Khi-carré est que chaque modalité i ait un effectif *théorique* au moins égal à 5. Le Khi-carré est en effet sensible aux petits effectifs : des effectifs théoriques faibles peuvent entrainer une surestimation de la valeur de χ^2 obtenu, ce qui, aux abords de la valeur critique, peut conduire à rejeter de façon erronée l'hypothèse nulle.

Si les trois formes de tests du Khi-carré obéissent à ces principes généraux, chaque forme se distingue cependant non seulement par son objectif et par la formulation de l'hypothèse nulle du test, mais aussi par la manière d'organiser les tableaux de contingence et de définir les effectifs théoriques.

11.2. TEST DU KHI-CARRÉ D'AJUSTEMENT

Dans l'absolu, le test du Khi-carré d'ajustement[102] (ou « test d'ajustement multinomial ») peut être utilisé pour vérifier l'adéquation à toute sorte de distribution : adéquation à une distribution empirique donnée, adéquation à une loi de probabilité, à une famille de lois, etc. L'hypothèse nulle du test est qu'il n'y pas de différence entre la distribution observée et la distribution théorique. Le tableau de contingence observé est de la forme :

Tableau de contingence observé

	Effectifs observés
Modalité 1	O_1
Modalité 2	O_2
⋮	⋮
Modalité M	O_M
Total	n

Les effectifs théoriques sont déterminés à partir de la distribution de référence. Cette dernière est représentée par la liste $(\rho_1, \rho_2, \cdots, \rho_M)$ des M fréquences théoriques

[102] Accessible dans XLSTAT via la commande *Tests paramétriques / Test d'ajustement multinomial*.

des M modalités de la variable[103], avec $\left(\sum_{m=1}^{M} \rho_m\right) = 1$.

Par conséquent, chaque effectif théorique T_m d'une modalité m est défini par :

$$T_m = \rho_m \times n$$

où
ρ_m est la fréquence théorique de la modalité m ;
n est l'effectif total observé.

Le tableau de contingence théorique est donc de la forme :

Tableau de contingence théorique

	Effectifs théoriques
Modalité 1	$T_1 = \rho_1 \times n$
Modalité 2	$T_2 = \rho_2 \times n$
⋮	⋮
Modalité M	$T_M = \rho_M \times n$
Total	n

On peut alors calculer la valeur du χ^2 conformément aux principes généraux :

[103] Voir exemple 11.1.

$$\chi^2 = \sum_{m=1}^{M} \frac{(O_m - T_m)^2}{T_m}$$

où O_m et T_m sont respectivement l'effectif observé et l'effectif théorique pour la modalité m.

On compare ensuite la valeur de χ^2 calculée à la table. Le tableau de contingence n'ayant qu'une seule colonne, le nombre de DDL est égal au nombre de catégories – 1, donc ici $DDL = M - 1$. L'hypothèse nulle peut être rejetée si la valeur de χ^2 calculée est supérieure à la valeur critique correspondant au nombre de DDL et au seuil de significativité retenu.

EXEMPLE 11.1.
Adéquation entre structure sociale d'un établissement et structure sociale du pays

Le Tableau 11.1 présente la structure sociale des élèves d'un établissement. On cherche à savoir si la structure sociale de l'établissement diffère significativement de la répartition socioprofessionnelle de la population active dans le pays, telle que décrite par le Tableau 11.2.

Tableau 11.1.
Répartition des élèves d'un établissement en fonction de la catégorie socioprofessionnelle du chef de famille

Catégorie socioprofessionnelle du chef de famille	Nombre d'élèves correspondant à la catégorie dans l'établissement (proportion du total)
Agriculteurs exploitants	98 (3,01 %)
Artisans, commerçants et chefs d'entreprise	910 (27,99 %)
Cadres et professions intellectuelles supérieures	423 (13,01 %)
Professions Intermédiaires	1007 (30,97 %)
Employés	586 (18,02 %)
Ouvriers	196 (6,02 %)
Autres	31 (0,95 %)
Total	*3251 (100 %)*

Tableau 11.2.
Répartition de la population active par catégories socioprofessionnelle, France, 2016. Source : Insee, enquête emploi

Catégorie socioprofessionnelle	Part dans la population active (r)
Agriculteurs exploitants	1,8 %
Artisans, commerçants et chefs d'entreprise	6,6 %
Cadres et professions intellectuelles supérieures	17,8 %
Professions Intermédiaires	25,8 %
Employés	27,4 %
Ouvriers	20,3 %
Autres	0,3 %
Total	*100 %*

Tableau de contingence observé

	Effectif observé
Agriculteurs exploitants	98
Artisans – Commerçants – CE	910
Cadres et P.I.S.	423
Professions Intermédiaires	1007
Employés	586
Ouvriers	196
Autres	31
Total	*3251*

Tableau de contingence théorique

	Effectif théorique
Agriculteurs exploitants	1,8 % x 3251 = 58,518
Artisans, commerçants et chefs d'entreprise	6,6 % x 3251 = 214,566
Cadres et professions intellectuelles supérieures	17,8 % x 3251 = 578,678
Professions Intermédiaires	25,8 % x 3251 = 838,758
Employés	27,4 % x 3251 = 890,774
Ouvriers	20,3 % x 3251 = 659,953
Autres	0,3 % x 3251 = 9,753
Total	*3251*

On note que pour chaque modalité, l'effectif théorique est au moins égal à 5.

Calcul du Khi-carré

	$O_m - T_m$	$(O_m - T_m)^2$	$\frac{(O_m - T_m)^2}{T_m}$
Agriculteurs exploitants	39,48	1558,82	26,63
Artisans, commerçants et	695,43	483628,4	2253,98

	$O_m - T_m$	$(O_m - T_m)^2$	$\frac{(O_m - T_m)^2}{T_m}$
chefs d'entreprise			
Cadres et professions intellectuelles supérieures	-155,68	24235,64	41,88
Professions Intermédiaires	168,24	28305,37	33,74
Employés	-304,77	92887,19	104,27
Ouvriers	-463,95	215252,4	326,16
Autres	21,25	451,43	46,28
		$Total = \chi^2 =$	*2832,97*

La table du Khi-carré indique que la valeur de Khi-carré calculée (2832,97) est supérieure à la valeur critique de 12,59 pour 6 degrés de liberté au seuil de significativité de 5%. On peut donc rejeter l'hypothèse nulle et conclure que la structure sociale des élèves de l'établissement est significativement différente de la structure socioprofessionnelle de la population active du pays.

EXEMPLE 11.2.
Évolution du sentiment de compétence en gestion de groupes difficiles chez 116 enseignants après un stage de formation

On dispose des données suivantes qui indiquent le sentiment subjectif d'enseignants en matière de gestion de groupes difficiles. Les données sont recueillies avant et après un stage de formation sur ce thème.

	Sentiment d'incompétence	Sentiment de compétence		
		Faible	Bonne	Forte
Distribution de l'effectif avant le stage	17	28	45	26
Distribution de l'effectif après le stage	13	32	36	35

Calcul du Khi-carré

On décide de retenir comme distribution théorique l'une des deux distributions, par exemple la distribution d'avant stage. On note que pour chaque modalité, l'effectif théorique est au moins égal à 5. Les calculs s'établissent comme suit :

	Sentiment d'incompétence	Sentiment de compétence		
		Faible	Bonne	Forte
$O_m - T_m$	-4,00	4,00	-9,00	9,00
$(O_m - T_m)^2$	16,00	16,00	81,00	81,00
$\frac{(O_m - T_m)^2}{T_m}$	0,941	0,571	1,800	3,115
$\chi^2 = 6{,}428$				

La table du Khi-carré indique que la valeur de Khi-carré calculée (6,42) est inférieure à la valeur critique de 7,82 pour DDL=3 au seuil de significativité de 5%. H_0 ne peut pas être rejetée. On en conclut que la différence entre les deux distributions n'est pas significative.

EXEMPLE 11.3.
Comparaison des pratiques de formation continue chez les enseignants du primaire dans deux pays

L'enquête internationale TALIS 2013 de l'OCDE sur les enseignants, l'enseignement et l'apprentissage fournit notamment des données sur les pratiques de

formation continue des enseignants. Le Tableau 11.3 montre la popularité de cinq thèmes de formation chez les enseignants de collège au cours de l'année précédant l'enquête en France et en Corée du sud :
- Discipline d'appartenance (DA) ;
- Pratiques d'évaluation (PE) ;
- TICE ;
- Gestion de classe (GC) ;
- Administration scolaire (AS).

Tableau 11.3.
Thèmes de formation continue pour 3432 enseignants de collège en France et en Corée du sud

Source : OCDE, Enquête TALIS 2013, série BTGINTT2, indicateur TT2G22, items A1, D1, E1, F1 et G1.

	DA	PE	TICE	GC	AS
Corée du Sud	979	591	663	777	422
France	1050	1021	799	473	89

La question est de savoir si la différence dans la répartition des enseignants entre thèmes est significative.

Calcul du Khi-carré

On décide de retenir l'une des deux distributions comme distribution théorique, en l'occurrence la ligne

France. On note que chaque modalité a un effectif théorique au moins égal à 5. Les calculs s'établissent comme suit :

	DA	PE	TICE	GC	AS
$O_m - T_m$	-71	-430	-136	304	333
$(o_m - T_m)^2$	5041	184900	18496	92416	110889
$\frac{(o_m - T_m)^2}{T_m}$	4,80	181,10	23,15	195,38	1245,94
$\chi^2 = 1650{,}37$					

Au seuil de significativité de 5%, le Khi-carré calculé (1650,37) est supérieur à la valeur critique pour 4 DDL (soit 9,49). Donc la différence entre parts respectives des thèmes de formation continue dans les deux pays est significative.

11.3. TEST DU KHI-CARRÉ D'INDÉPENDANCE

Soit un échantillon d'individus caractérisés par deux variables qualitatives (par exemple le genre de l'apprenant et la filière choisie, ou encore la formation de l'enseignant et le style d'enseignement, etc.). On cherche à savoir s'il existe un lien statistiquement significatif entre les variables ou au moins entre certaines de leurs modalités.

On applique un *test du Khi-carré d'indépendance*[104]. L'objectif du test est de vérifier l'existence d'une relation[105]. On dit qu'il y a « *association* » (ou « *attraction* »)[106] lorsqu'il est mis en évidence un lien entre variables (ou modalités de variables), et « *indépendance* » sinon.

Dans le test du Khi-carré d'indépendance, l'hypothèse nulle postule que les variables (ou modalités de variables) n'ont pas de liens entre elles. Le tableau de contingence observé est de la forme :

---

[104] Accessible dans XLSTAT via la commande *Tests de corrélation/association – Tests sur les tableaux de contingence*.

[105] Mais le Khi-carré ne mesure pas l'*intensité* d'une relation éventuellement mise en évidence. Les méthodes pour mesurer l'intensité d'une relation sont présentées au chapitre 13.

[106] Les variables étant ici qualitatives, on n'utilise pas le terme « corrélation », qui s'applique plutôt aux relations entre variables quantitatives.

Tableau de contingence observé

		Variable 2				
		Modalité 2.1	Modalité 2.2	⋯	Modalité 2.ψ	Total
Variable 1	Modalité 1.1	O_{11}	O_{12}	…	$O_{1\psi}$	$\mu_{1.1}$
	Modalité 1.2	O_{21}	O_{22}	…	$O_{2\psi}$	$\mu_{1.2}$
	⋮	⋮	⋮	⋮	⋮	⋮
	Modalité 1.ξ	$O_{\xi 1}$	$O_{\xi 2}$	…	$O_{\xi\psi}$	$\mu_{1.\xi}$
	Total	$\mu_{2.1}$	$\mu_{2.2}$	…	$\mu_{2.\psi}$	n

Les effectifs théoriques sont habituellement calculés à partir des *sommes marginales*, c'est-à-dire à partir des marges (ligne total et colonne total) du tableau de contingence. Chaque effectif théorique T_{lc} correspondant à l'intersection entre ligne l et colonne c est alors défini par :

$$T_{lc} = \frac{\mu_{1.l} \times \mu_{2.c}}{n}$$

où

- $\mu_{1.l}$ est le total des effectifs de la ligne l de la variable 1 ;
- $\mu_{2.c}$ est le total des effectifs de la colonne c de la variable 2 ; et
- n est l'effectif total observé.

Par exemple :

$$T_{11} = \frac{\mu_{1.1} \times \mu_{2.1}}{n} \quad ; \quad T_{12} = \frac{\mu_{1.1} \times \mu_{2.2}}{n} \quad ; \quad T_{21} = \frac{\mu_{1.2} \times \mu_{2.1}}{n} \quad ; \quad etc.$$

Une approche alternative cependant consiste à attribuer le même effectif théorique à tous les couples de modalités. Dans ce cas, quel que soit l'effectif observé, l'effectif théorique correspondant est donné par :

$$Effectif\ théorique = \frac{n}{\xi * \psi}$$

où ξ et ψ sont respectivement le nombre de lignes et le nombre de colonnes dans le tableau de contingence observé. L'idée derrière cette approche alternative est que des effectifs théoriques constants suggèrent une absence de relation entre variables.

Une fois les effectifs théoriques déterminés, le calcul et l'interprétation du χ^2 s'effectuent suivant les règles générales[107].

[107] Voir section 11.1.

EXEMPLE 11.4
Relation entre niveau d'études et statut d'emploi chez les enseignants

L'enquête internationale TALIS 2013 de l'OCDE sur les enseignants, l'enseignement et l'apprentissage fournit notamment des données sur le niveau d'études et le statut d'emploi des enseignants. Le Tableau 11.4 montre la répartition des enseignants de lycée en fonction de leur niveau d'études et de la pérennité de leur emploi à Singapour.

Tableau 11.4.
Répartition des enseignants de lycée en fonction de leur niveau d'études et de la pérennité du contrat de travail à Singapour : emploi permanent ou contrat à durée déterminée (CDD)

Source : OCDE, Enquête TALIS 2013, série CTGINTT2, indicateurs TT2G06 et TT2G10.

		Pérennité du contrat de travail		
		Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Niveau d'études	Inférieur au premier cycle du supérieur	15	1	5
	Premier cycle du supérieur – Professionnel	84	9	11
	Premier cycle du supérieur – Général	2788	147	56
	Deuxième cycle du supérieur	7	4	0

Calcul du Khi-carré selon la méthode des effectifs théoriques égaux

Les calculs s'établissent comme suit :

Effectifs observés			
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Inférieur au premier cycle du supérieur	15	1	5
Premier cycle du supérieur – Professionnel	84	9	11
Premier cycle du supérieur – Général	2788	147	56
Deuxième cycle du supérieur	7	4	0

Effectifs théoriques égaux			
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Inférieur au premier cycle du supérieur	260,58	260,58	260,58
Premier cycle du supérieur – Professionnel	260,58	260,58	260,58
Premier cycle du supérieur – Général	260,58	260,58	260,58
Deuxième cycle du supérieur	260,58	260,58	260,58

Écarts entre effectifs observés et effectifs théoriques			
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Inférieur au premier cycle du supérieur	-245,58	-259,58	-255,58
Premier cycle du supérieur – Professionnel	-176,58	-251,58	-249,58
Premier cycle du supérieur – Général	2527,41	-113,58	-204,58
Deuxième cycle du supérieur	-253,58	-256,58	-260,58

Carrés des écarts entre effectifs			
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Inférieur au premier cycle du supérieur	60311,17	67383,50	65322,84
Premier cycle du supérieur – Professionnel	31181,67	63294,17	62291,84

Premier cycle du supérieur – Général	6387835,0	12901,17	41854,34
Deuxième cycle du supérieur	64304,50	65835,00	67903,67

Carrés rapportés aux effectifs théoriques			
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Inférieur au premier cycle du supérieur	231,44	258,58	250,67
Premier cycle du supérieur – Professionnel	119,66	242,89	239,04
Premier cycle du supérieur – Général	24513,59	49,50	160,61
Deuxième cycle du supérieur	246,77	252,64	260,58
$\chi^2 = 26826{,}039$			

On note que chaque modalité a un effectif théorique au moins égal à 5.

La valeur du Khi-carré obtenue (26826,039) est supérieure à la valeur critique pour 6 DDL au seuil de 5% (12,59), donc il y a association entre le niveau d'études et la pérennité du contrat de travail. La nature de la relation peut être précisée au moyen du « *tableau de significativité par case* », que fournissent certains logiciels, et qui permet d'identifier les cases / cellules pour lesquelles l'effectif observé est significativement supérieur ou inférieur à l'effectif théorique. Par exemple ici :

	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Inférieur au premier cycle du supérieur	<	(<)	>
Premier cycle du supérieur – Professionnel	<	(>)	>
Premier cycle du supérieur – Général	>	<	<
Deuxième cycle du supérieur	<	>	<

Les signes entre parenthèses indiquent que les écarts entre effectifs théoriques et effectifs observés ne sont pas significatifs. Les signes qui ne sont pas entre parenthèses indiquent que les valeurs observées sont significativement supérieures ou inferieures aux effectifs théoriques au seuil de significativité de 5%. On peut ainsi préciser la relation entre niveau d'études et pérennité du contrat de travail :

- l'emploi permanent est significativement plus fréquent chez les enseignants ayant un niveau d'études égal au Premier cycle du supérieur – Général ;
- le CDD de plus d'un an est significativement plus fréquent chez les enseignants ayant un niveau d'études égal au Deuxième cycle du supérieur ;
- le CDD d'un an au plus est significativement plus fréquent chez les enseignants dont le niveau d'études est Inférieur au premier cycle du supérieur ou égal au Premier cycle du supérieur – Professionnel.

Calcul du Khi-carré à partir des sommes marginales

Effectifs observés				
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus	Total
Inférieur au premier cycle du supérieur	15	1	5	21
Premier cycle du supérieur – Pro.	84	9	11	104
Premier cycle du supérieur – Général	2788	147	56	2991
Deuxième cycle du supérieur	7	4	0	11
Total	2894	161	72	3127

Effectifs théoriques				
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus	Total
Inférieur au premier cycle du supérieur	19,43	1,0812	0,48	21
Premier cycle du supérieur – Pro.	96,25	5,35	2,39	104
Premier cycle du supérieur – Général	2768,13	153,99	68,86	2991
Deuxième cycle du supérieur	10,18	0,56	0,25	11
Total	2894	161	72	3127

Écarts entre effectifs observés et effectifs théoriques			
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Inférieur au premier cycle du supérieur	-4,43	-0,08	4,51
Premier cycle du supérieur – Professionnel	-12,25	3,64	8,60
Premier cycle du supérieur – Général	19,86	-6,99	-12,86
Deuxième cycle du supérieur	-3,18	3,43	-0,253

Carrés des écarts entre effectifs			
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus
Inférieur au premier cycle du supérieur	19,67	0,01	20,40
Premier cycle du supérieur – Professionnel	150,08	13,29	74,05
Premier cycle du supérieur – Général	394,67	48,97	165,60
Deuxième cycle du supérieur	10,11	11,79	0,06

Carrés rapportés aux effectifs théoriques				
	Emploi permanent	CDD de plus d'un an	CDD d'un an au plus	Total
Inférieur au premier cycle du supérieur	1,01	0,01	42,19	43,20
Premier cycle du supérieur – Pro.	1,56	2,48	30,92	34,97
Premier cycle du supérieur – Général	0,14	0,32	2,40	2,87
Deuxième cycle du supérieur	0,99	20,82	0,25	22,06
$\chi^2 = 103{,}09$				

On constate que la valeur du Khi-carré obtenue diffère, mais la conclusion reste la même : il y a un lien significatif entre le niveau d'études de l'enseignant et la pérennité de son contrat de travail.

On peut observer cependant que les effectifs théoriques par classe ne sont pas tous ici au moins égaux à 5, ce qui peut remettre en cause la validité du test. Une méthode usuelle pour éliminer ce risque consiste à regrouper des catégories dont les effectifs sont trop faibles, par exemple ici regrouper les deux modalités de CDD. Le tableau de contingence devient :

Effectifs observés		
	Emploi permanent	CDD
Inférieur au premier cycle du supérieur	15	6
Premier cycle du supérieur – Professionnel	84	20
Premier cycle du supérieur – Général	2788	203
Deuxième cycle du supérieur	7	4

La valeur du Khi-carré s'établit à 49,75 et continue donc à indiquer une association significative au seuil de 5% (valeur critique du Khi-carré = 7,81 pour DDL = 3).

Cependant, le regroupement n'a pas permis d'éliminer tous les effectifs théoriques inférieurs à 5 :

Effectifs théoriques		
	Emploi permanent	CDD
Inférieur au premier cycle du supérieur	19,43	1,56
Premier cycle du supérieur – Professionnel	96,25	7,74
Premier cycle du supérieur – Général	2768,13	222,86
Deuxième cycle du supérieur	10,18	0,82

Le résultat du test par la méthode des sommes marginales n'est donc pas ici pleinement convaincant.

EXEMPLE 11.5

On dispose de données sur l'appartenance catégorielle des électeurs (étudiants, enseignants, personnels administratifs) et leur choix de vote lors d'une élection à la présidence d'une université (Tableau 11.5). On se demande si / dans quelle mesure la catégorie d'appartenance a été un facteur explicatif du vote.

Tableau 11.5.
Répartition des voix par catégorie d'électeur

	Candidat A	Candidat B	Candidat C
Nombre de voix chez les Enseignants	40	50	60
Nombre de voix chez les Administratifs	35	55	75
Nombre de voix chez les Étudiants	15	45	75
Total	*90*	*150*	*210*

Calcul du Khi-carré

Les calculs s'établissent comme suit :

Effectifs observés				
	Candidat A	Candidat B	Candidat C	Total
Nombre de voix chez les Enseignants	40	50	60	150
Nombre de voix chez les Administratifs	35	55	75	165
Nombre de voix chez les Étudiants	15	45	75	135
Total	90	150	210	450

Effectifs théoriques (sommes marginales)				
	Candidat A	Candidat B	Candidat C	Total
Nombre de voix chez les Enseignants	30	50	70	150
Nombre de voix chez les Administratifs	33	55	77	165
Nombre de voix chez les Étudiants	27	45	63	135
Total	90	150	210	450

Écarts entre effectifs observés et effectifs théoriques			
	Candidat A	Candidat B	Candidat C
Nombre de voix chez les Enseignants	10	0	-10
Nombre de voix chez les Administratifs	2	0	-2
Nombre de voix chez les Étudiants	-12	0	12

Carrés des écarts entre effectifs			
	Candidat A	Candidat B	Candidat C
Nombre de voix chez les Enseignants	100	0	100
Nombre de voix chez les Administratifs	4	0	4
Nombre de voix chez les Étudiants	144	0	144

Carrés rapportés aux effectifs théoriques				
	Candidat A	Candidat B	Candidat C	Total
Nombre de voix chez les Enseignants	3,33	0	1,42	4,76
Nombre de voix chez les Administratifs	0,12	0	0,051	0,17
Nombre de voix chez les Étudiants	5,33	0	2,28	7,61
$\chi^2 = 12{,}554$				

On note que les effectifs théoriques par case sont tous au moins égaux à 5.

La valeur du Khi-carré obtenue (12,55) est supérieure à la valeur critique au seuil de 5% (9,49 pour 4 DDL), donc H_0 est rejetée. On peut conclure que la catégorie d'appartenance détermine le vote. Le tableau de significativité par case permet d'être plus précis :

	Candidat A	Candidat B	Candidat C
Nombre de voix chez les Enseignants	>	(=)	<
Nombre de voix chez les Administratifs	(>)	(=)	(<)
Nombre de voix chez les Étudiants	<	(=)	>

Les enseignants ont significativement vote plus pour le candidat A et moins pour le candidat C. Les étudiants ont significativement vote plus pour le candidat C et moins pour le candidat A. Aucune tendance significative n'émerge concernant le candidat B.

11.4. TEST DU KHI-CARRÉ D'HOMOGÉNÉITÉ

Le test du Khi-carré d'homogénéité[108] permet de vérifier si les proportions des modalités d'une variable donnée sont identiques dans plusieurs échantillons. On suppose donc *K* échantillons, de tailles non nécessairement égales, dans lesquels on compare les proportions des *m* modalités d'une même variable.

L'hypothèse nulle du test postule que la structure des modalités de la variable est identique quel que soit l'échantillon. Le tableau de contingence observé est de la forme :

Tableau de contingence observé

		Échantillons				
		E1	E2	…	E *K*	Total
Variable	Modalité 1	O_{11}	O_{12}	…	O_{1K}	μ_1
	Modalité 2	O_{21}	O_{22}	…	O_{2K}	μ_2
	⋮	⋮	⋮	⋮	⋮	⋮
	Modalité M	O_{M1}	O_{M2}	…	O_{MK}	μ_M
	Total	n_1	n_2	…	n_K	S

[108] Accessible dans XLSTAT via la commande *Tests paramétriques – Comparaison de k proportions*.

Les effectifs théoriques se calculent par méthode des sommes marginales. Chaque effectif théorique T_{mk} d'une modalité m dans un échantillon k est donc défini par :

$$T_{mk} = \frac{\mu_m \times n_k}{S}$$

où

μ_m est la somme des effectifs correspondant à la modalité *m* dans tous les échantillons ; et

n_k est la taille de l'échantillon *k* ;

S est la somme des effectifs des *K* échantillons.

Par exemple :

$$T_{11} = \frac{\mu_1 \times n_1}{S} \quad ; \quad T_{12} = \frac{\mu_1 \times n_2}{S} \quad ; \quad T_{21} = \frac{\mu_2 \times n_1}{S} \quad ; \quad etc.$$

L'analyse s'effectue modalité par modalité. Pour chaque modalité, on compare sa proportion dans les *k* échantillons.

EXEMPLE 11.6

On dispose de données de 2014 sur les jeunes d'âge scolaire mais en marge de l'école à quatre niveaux d'éducation (pré-primaire, primaire, collège et lycée) dans quatre pays (Tableau 11.6). On cherche à savoir si la part relative de chaque niveau d'éducation dans cette

population est significativement différente d'un pays à l'autre.

Tableau 11.6.
Nombre de jeunes d'âge scolaire mais en marge de l'école aux niveaux pré-primaire, primaire, collège et lycée en Belgique, aux Etats-Unis, en Espagne et en France en 2014

Source : Banque mondiale, Base de données, Statistiques de l'Éducation
http://databank.worldbank.org/data/reports.aspx?source=education-statistics-~-all-indicators#

	Belgique	Etats-Unis	Espagne	France
Pré-primaire	5318	937060	5874	19465
Primaire	6538	1470712	28642	22173
Collège	333	114566	1742	2012
Lycée	4486	943721	34582	5234
Total	*16675*	*3466059*	*70840*	*48884*

Pré-primaire

Le tableau des observations à soumettre au test se présente comme suit :

Libellés des échantillons	Effectifs	Taille des échantillons
	Pré-primaire	Total
Belgique	5318	16675
Etats-Unis	937060	3466059
Espagne	5874	70840
France	19465	48884

Les résultats des tests indiquent :

Échantillons	Proportion de la modalité	*p-value* du test du Khi-carré	Classes
Belgique	0,32	<0,0001	A
Etats-Unis	0,27		B
Espagne	0,08		C
France	0,40		D

La *p-value* du test du Khi-carré indique que la proportion de la modalité « Pré-primaire » diffère significativement dans au moins deux pays. Reste alors à déterminer quels pays exactement ont des proportions différentes. La procédure de Marascuilo permet, en comparant deux-à-deux les proportions, de tester si

elles sont ou non significativement différentes. Le test consiste à comparer l'écart entre proportions à une valeur critique. Deux proportions sont significativement différentes l'une de l'autre si l'écart entre elles est supérieur à la valeur critique (à un seuil de significativité donné). Si deux proportions sont différentes, elles appartiennent à des classes distinctes. Ici, la proportion de la modalité « Pré-primaire » diffère significativement d'un pays à l'autre.

Primaire

Tableau des observations

Libellés des échantillons	Effectifs	Taille des échantillons
	Primaire	Total
Belgique	6538	16675
Etats-Unis	1470712	3466059
Espagne	28642	70840
France	22173	48884

Résultats des tests

Échantillons	Proportion de la modalité	*p-value* du test du Khi-carré	Classes
Belgique	0,39	<0,0001	A
Etats-Unis	0,42		B
Espagne	0,40		C
France	0,45		D

Le test du Khi-carré indique que la proportion de la modalité « Primaire » diffère significativement dans au moins deux pays. La procédure de Marascuilo montre qu'elle diffère significativement d'un pays à l'autre.

Collège

Tableau des observations

Libellés des échantillons	Effectifs	Taille des échantillons
	Collège	Total
Belgique	333	16675
Etats-Unis	114566	3466059
Espagne	1742	70840
France	2012	48884

Résultats des tests

Échantillons	Proportion de la modalité	*p-value* du test du Khi-carré	Classes
Belgique	0,019	<0,0001	A
Etats-Unis	0,033		B
Espagne	0,024		C
France	0,041		D

Le test du Khi-carré indique que la proportion de la modalité « Collège » diffère significativement dans au moins deux pays. La procédure de Marascuilo montre qu'elle diffère significativement d'un pays à l'autre.

Lycée

Tableau des observations

Libellés des échantillons	Effectifs	Taille des échantillons
	Lycée	Total
Belgique	4486	16675
Etats-Unis	943721	3466059
Espagne	34582	70840
France	5234	48884

Résultats des tests

Échantillons	Proportion de la modalité	*p-value* du test du Khi-carré	Classes
Belgique	0,269	<0,0001	A
Etats-Unis	0,272		A
Espagne	0,488		B
France	0,107		C

Le test du Khi-carré indique que la proportion de la modalité « Lycée » diffère significativement dans au moins deux pays. La procédure de Marascuilo montre qu'elle diffère significativement entre l'ensemble Belgique et Etats-Unis d'une part, et d'autre part l'Espagne ou la France, ainsi qu'entre l'Espagne et la France.

Chapitre 12. Les tests d'équivalence : la moyenne d'un échantillon est-elle suffisamment proche d'une référence ou de la moyenne d'un autre échantillon ?

Les tests de comparaison de moyennes permettent de savoir si la moyenne d'un échantillon est significativement différente d'une référence ou de la moyenne d'un ou plusieurs autres échantillons. Lorsque la probabilité *p* est inférieure au seuil de significativité, on accepte l'hypothèse alternative de différence des moyennes. Dans le cas contraire, on conclut qu'on ne peut pas rejeter l'hypothèse nulle d'égalité entre moyennes. La problématique traitée dans ce chapitre est différente.

12.1. PROBLÉMATIQUE

Le cas de figure envisagé ici est celui dans lequel l'objectif de l'analyse est de montrer que les moyennes comparées peuvent être considérées comme équivalentes.

La notion d'équivalence signifie que le chercheur dispose, par exemple sur la base d'une théorie, d'une idée de l'écart qui peut être considéré comme acceptable entre moyennes. Lorsque l'écart effectivement observé entre moyennes reste dans les

limites de l'écart acceptable, les moyennes sont équivalentes. Elles ne le sont plus dans le cas contraire.

On pourrait imaginer que les tests de comparaison de moyennes pourraient permettre de juger si deux moyennes sont équivalentes. L'idée serait que lorsque l'hypothèse nulle ne peut être rejetée, cela signifierait que les moyennes sont proches. Mais le problème est que le fait que les moyennes ne sont pas significativement différentes au sens des tests de comparaison des moyennes ne signifie pas qu'elles ont le niveau de proximité requis par la théorie dans le cadre de laquelle s'inscrit le chercheur.

D'où la nécessité de recourir à un outil plus approprié, les tests d'équivalence, dont l'objet est précisément d'évaluer l'équivalence entre moyennes.

La problématique et la méthodologie des tests d'équivalence ont été à l'origine développées en recherche pharmaceutique, notamment pour la définition de classes d'équivalence précises entre dosages. En éducation et formation, la question de l'équivalence se pose par exemple lorsqu'il s'agit de juger de la substituabilité d'épreuves auxquels des testeurs (phase pilote de questionnaire) ou des candidats (lors d'examens) sont soumis à des périodes ou horaires différents. Un autre exemple est celui dans lequel des instruments édumétriques (questionnaires de

motivation ou d'évaluation par exemple) doivent pouvoir être clairement identifiés comme équivalents ou pas.

12.2. DÉMARCHE

La démarche générale des tests d'équivalence s'organise en trois étapes.

12.2.1. Définition de l'intervalle d'équivalence

La première étape consiste à définir le niveau de différence considéré comme acceptable et en deçà duquel les moyennes peuvent être considérées comme équivalentes. Par exemple, le chercheur peut décider que la moyenne au bac dans les années 2000 est équivalente à la moyenne dans les années soixante-dix si l'écart entre les deux n'excède pas un point. Donc si la moyenne dans les années soixante-dix s'est établie à 12/20, la moyenne des années 2000 lui est équivalente si elle se situe dans l'intervalle [11 ; 13]. Ce qui revient à dire que la *différence* entre moyennes équivalentes se situe dans l'intervalle [-1 ; +1]. De même, autre exemple, si le chercheur considère que, pour être équivalentes, les moyennes de deux types de répondants à un questionnaire ne doivent pas différer de plus de 5 points, l'intervalle qui délimite l'*écart* acceptable (« intervalle d'équivalence ») est [-5 ; +5]. Etc.

12.2.2. Hypothèses du test d'équivalence et procédure des tests unilatéraux

Les tests d'équivalence ont pour objectif de rejeter une hypothèse nulle et retenir une hypothèse alternative définies par :

$$\begin{cases} H_0 : (\overline{x_A} - \overline{x_B}) \leq Inf \;\; et/ou \;\; (\overline{x_A} - \overline{x_B}) \geq Sup \\ H_A : Inf < (\overline{x_A} - \overline{x_B}) < Sup \end{cases}$$

où

- $(\overline{x_A} - \overline{x_B})$ est la différence entre moyennes ;
- Inf la borne inférieure de l'intervalle d'équivalence ;
- Sup la borne supérieure de l'intervalle d'équivalence.

La deuxième étape consiste alors à tester :

- l'hypothèse que la différence observée soit significativement *supérieure* à la borne *inférieure* de l'intervalle d'équivalence ; et
- l'hypothèse que la différence observée soit significativement *inférieure* à la borne *supérieure* de l'intervalle d'équivalence.

On applique à cette fin deux tests *t* de Student unilatéraux de comparaison de moyennes[109] :

[109] Ce qui donne son nom à la procédure : *two one-sided tests – TOST*.

- un test unilatéral à droite sur la borne inférieure de l'intervalle d'équivalence, dont les hypothèses sont

$$\begin{cases} H_0 : (\overline{x_A} - \overline{x_B}) = Inf \\ H_A : (\overline{x_A} - \overline{x_B}) > Inf \end{cases}$$

- un test unilatéral à gauche sur la borne supérieure de l'intervalle d'équivalence, dont les hypothèses sont

$$\begin{cases} H_0 : (\overline{x_A} - \overline{x_B}) = Sup \\ H_A : (\overline{x_A} - \overline{x_B}) < Sup \end{cases}$$

12.2.3. Interprétation

La troisième étape consiste à conclure :

- on retient comme p-value du test d'équivalence, la p-value la plus élevée des deux tests unilatéraux (ou les deux si elles sont égales) ;
- une p-value la plus élevée supérieure au seuil de significativité signifie que l'hypothèse alternative du test unilatéral considéré n'est pas retenue, c'est-à-dire encore que soit la différence n'est pas supérieure à la borne inférieure de l'intervalle d'équivalence, soit elle n'est pas inférieure à la borne supérieure de l'intervalle d'équivalence. Dans les deux cas, la différence n'est pas comprise dans l'intervalle d'équivalence, donc il n'y a pas équivalence ;
- une p-value la plus élevée inférieure au seuil de significativité signifie que l'hypothèse alternative est retenue dans chacun des deux tests unilatéraux, c'est-

à-dire en d'autres termes que la différence se situe dans l'intervalle d'équivalence, donc qu'on peut conclure à l'équivalence.

Les tests d'équivalence diffèrent donc, dans leur logique et dans leur procédure, des tests de comparaison de moyennes (qu'ils font intervenir cependant), et en particulier dans la définition de l'hypothèse nulle : dans les tests de comparaison de moyennes, l'hypothèse nulle postule l'égalité des moyennes ; *c'est l'inverse dans les tests d'équivalence*.

Enfin, il importe de souligner que la procédure des tests d'équivalence s'appuyant sur l'usage de tests de Student, la fiabilité des tests nécessite que la condition de normalité des distributions soit préalablement remplie.

EXEMPLE 12.1

On considère deux questionnaires d'évaluation des enseignements par les étudiants. On cherche à savoir si ces deux questionnaires peuvent être considérés comme conduisant à des résultats équivalents. On les administre à deux échantillons aléatoires de la même population d'étudiants. Le Tableau 12.1 montre les scores obtenus dans chaque échantillon. On effectue un test d'équivalence. L'équivalence est définie par un écart maximum de deux points entre moyennes.

Tableau 12.1.
Scores à deux questionnaires d'évaluation des enseignements par les étudiants

Identifiant étudiant-e	Questionnaire 1 administré à l'échantillon A	Identifiant étudiant-e	Questionnaire 2 administré à l'échantillon B
A1	80	B1	69
A2	35	B2	37
A3	11	B3	17
A4	74	B4	85
A5	25	B5	28
A6	30	B6	61
A7	32	B7	23
A8	81	B8	68
A9	60	B9	21
A10	24	B10	28
A11	94	B11	56
A12	40	B12	62
A13	87	B13	77
A14	80	B14	77
A15	20	B15	93
A16	89	B16	59
A17	33	B17	62
A18	95	B18	47
A19	55	B19	71
A20	61	B20	42
Moyenne	*55,30*	*Moyenne*	*54,15*

Les deux échantillons atteignent des moyennes très proches : 55,30 et 54,15, soit une différence de 1,15. Il y a donc au départ une présomption d'équivalence.

Si on applique un test de comparaison de moyennes :

- les deux séries sont normalement distribuées

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Q 1	0,066	0,061	0,161	0,391
Q 2	0,385	0,360	0,443	0,580

- les variances sont homogènes. En effet, la valeur calculée du F de Fisher s'élève à 1,53 contre une valeur critique de 2,52 au seuil de 5% avec DDL1=DDL2=19, et p=0,35 ;

- la valeur calculée de la statistique t de Student pour deux échantillons indépendants s'élève à 0,14 contre une valeur critique de 2,024 pour 38 DDL au seuil de significativité de 5% ; p=0,88. Donc on ne peut rejeter l'hypothèse nulle d'égalité des moyennes.

Si on applique à présent un test d'équivalence, on pose d'abord les hypothèses du test :

$$\begin{cases} H_0 : différence \leq -2 \ \ et/ou \ \ différence \geq 2 \\ H_A : -2 < différence < 2 \end{cases}$$

Les résultats des tests unilatéraux indiquent :

	Écart	t	Valeur critique de *t*	DDL	Seuil de signif. α	p
« Supérieur » (Différence supérieure à la borne inférieure : test unilatéral à droite sur la borne inférieure)	1,15	0,39	1,68	38	0,05	0,34
« Inférieur » (Différence inférieure à la borne supérieure : test unilatéral à gauche sur la borne supérieure)	1,15	-0,10	-1,68	38	0,05	0,45

La probabilité *p* la plus élevée (0,45) est supérieure au seuil de significativité, donc on ne peut rejeter l'hypothèse nulle du test d'équivalence. On conclut donc que les deux moyennes ne peuvent être considérées comme équivalentes au sens adopté pour cette analyse. On voit que cette conclusion n'est pas la même que celle de non-différence à laquelle conduisait le test *t* de comparaison des moyennes.

Chapitre 13. Corrélation et association : mesurer l'intensité d'un lien entre deux variables

Différentes techniques comme le calcul de covariance[110] ou l'application d'un test de Khi-carré d'indépendance[111] peuvent conduire à déceler l'existence d'un lien entre deux variables, qu'il s'agisse de variables quantitatives ou qualitatives. La question est alors de pousser plus avant l'exploration afin de préciser la nature de ce lien. On procède donc à des calculs de « corrélation » si les deux variables sont quantitatives, et à des mesures « d'association » si les deux variables sont qualitatives

Quatre méthodes de calcul de corrélations sont présentées dans ce chapitre : les coefficients de corrélation de Pearson, de Spearman, de Kendall, et le coefficient de corrélation bisérielle ponctuelle. Chacune de ces méthodes permet (a) de mesurer l'intensité de la relation, et (b) de dire si l'éventuelle relation est positive (quand les valeurs d'une variable augmentent, les valeurs de l'autre variable augmentent aussi) ou négative (les valeurs des deux variables varient en sens opposés).

[110] Voir chapitre 1.

[111] Voir chapitre 11.

L'usage du coefficient de Pearson est pertinent lorsque la relation analysée est linéaire et les variables normalement distribuées. L'usage des coefficients de Spearman ou Kendall est pertinent lorsque les relations sont non-linéaires mais monotones. La corrélation bisérielle ponctuelle s'utilise lorsque les deux variables dont on étudie la relation sont l'une qualitative nominale binaire et l'autre quantitative.

Lorsque les deux variables sont qualitatives, on applique des mesures d'association. La méthode pertinente dépend (a) du caractère ordinal ou nominal de la variable ; (b) de la taille du tableau de contingence ; et (c) de la portée générale ou locale de la mesure.

Il importe de ne pas se méprendre sur le rôle des mesures de corrélation et d'association : les mesures de corrélation et d'association visent simplement à quantifier l'intensité d'éventuels liens entre variables, elles n'ont ni pour objet ni pour effet d'établir des *causalités*. Ce n'est pas parce que deux variables sont liées que l'une est la cause de l'autre : par exemple, elles peuvent très bien être toutes les deux causées par une troisième variable non prise en compte dans l'analyse.

Les sections 13.1 et 13.2 présentent le coefficient de Pearson puis son utilisation pour mesurer la corrélation entre plus de deux variables. Les sections 13.3 et 13.4

présentent les coefficients de Spearman et Kendall. La section 13.5 présente la corrélation bisérielle ponctuelle. La section 13.6 présente les coefficients de mesure d'association.

13.1. COEFFICIENT DE CORRÉLATION LINÉAIRE DE PEARSON

Le coefficient de corrélation linéaire de Pearson s'utilise dans le cas de relation linéaire entre deux variables normalement distribuées. Lorsque ces conditions ne sont pas remplies, l'usage du coefficient peut conduire à des conclusions erronées. Pour la même raison, il importe également que les séries aient été expurgées de leurs valeurs aberrantes[112]. Pour deux variables quantitatives X et Y, le coefficient de Pearson est défini comme le rapport entre la covariance des deux variables et le produit de leurs écart-types :

$$r = \frac{Cov_{X,Y}}{\sigma_X \times \sigma_Y} = \frac{\frac{\sum[(x - \bar{x}) \times (y - \bar{y})]}{n}}{\sigma_X \times \sigma_Y}$$

Le r de Pearson permet de mesurer l'intensité de la relation linéaire, et son sens, positif ou négatif. Le r de Pearson peut prendre toute valeur comprise dans l'intervalle $[-1\,;1]$. La corrélation est considérée

[112] Voir chapitre 3.

comme fortement négative si elle est proche de -1, fortement positive si elle est proche de 1, et nulle si elle est égale à 0.

Enfin, la constatation de l'existence d'une corrélation ne suffit pas, même si cette corrélation est forte. Encore faut-il savoir si la corrélation est significative. Le test de significativité s'appuie sur la table du r de Pearson (Annexe 6[113]). La table indique les seuils (valeurs critiques) au-dessus desquels une corrélation peut être jugée significative. L'hypothèse nulle du test postule l'absence de relation entre variables ; l'hypothèse nulle peut être rejetée si la valeur calculée du r de Pearson est supérieure à la valeur critique indiquée dans la table pour le nombre de degrés de liberté et au taux de significativité considérés. Le nombre de degrés de liberté est égal à $(n-2)$ où n est la taille de l'échantillon, c'est-à-dire le nombre de paires (x, y) analysées. La probabilité p indique le risque de se tromper en rejetant l'hypothèse nulle.

[113] Une version plus complète de la table de Pearson pour test bilatéral peut être consultée par exemple sur le site Real Statistics à l'adresse :
http://www.real-statistics.com/statistics-tables/pearsons-correlation-table/.

EXEMPLE 13.1

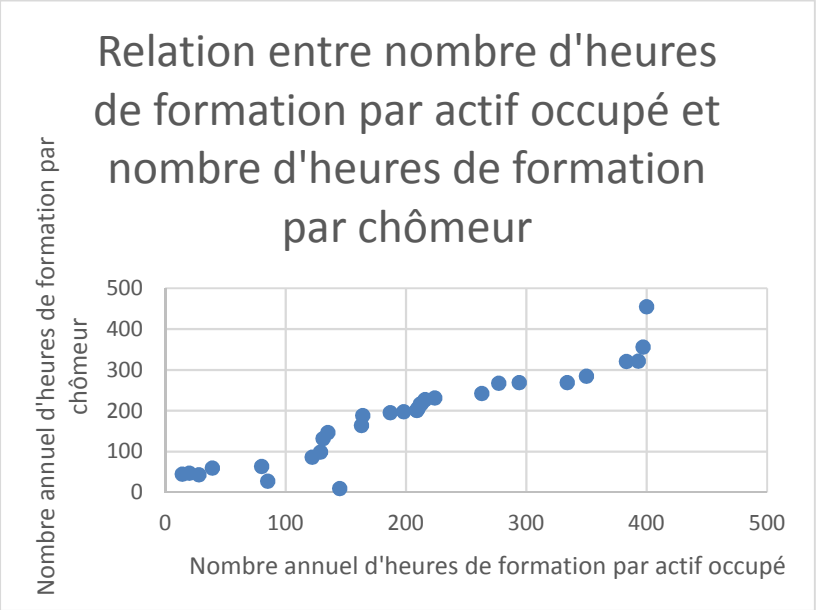
Le Tableau 13.1 indique la durée de la formation reçue en moyenne par chaque stagiaire de la formation continue dans 31 circonscriptions territoriales (Départements). Les données distinguent suivant que le bénéficiaire était en emploi ou au chômage. On cherche à savoir si dans ces 31 Départements, la durée de formation des chômeurs est corrélée à celle des actifs occupés.

Tableau 13.1.
Nombre annuel d'heures de formation par stagiaire (occupé ou chômeur)

Identifiant du Département	Occupés	Chômeurs
1	20	48
2	209	201
3	224	232
4	212	217
5	383	321
6	277	268
7	400	455
8	28	44
9	334	269
10	397	356
11	187	196
12	294	269
13	145	10
14	210	207

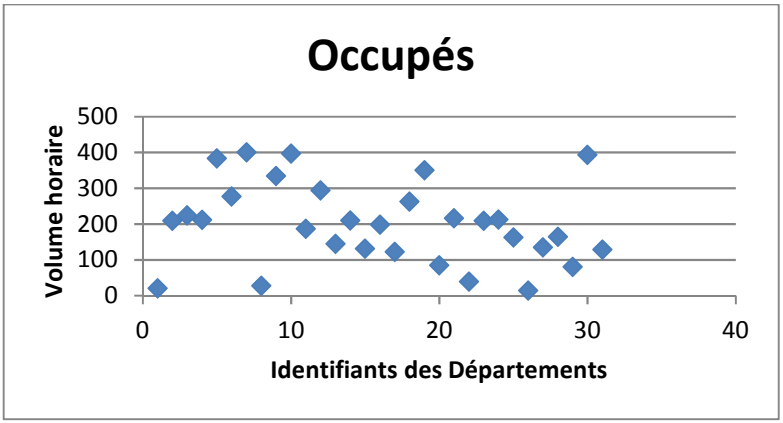
Identifiant du Département	Occupés	Chômeurs
15	131	132
16	198	198
17	122	87
18	263	243
19	350	285
20	85	28
21	216	228
22	39	60
23	209	202
24	213	218
25	163	164
26	14	45
27	135	147
28	164	189
29	80	64
30	393	322
31	129	99

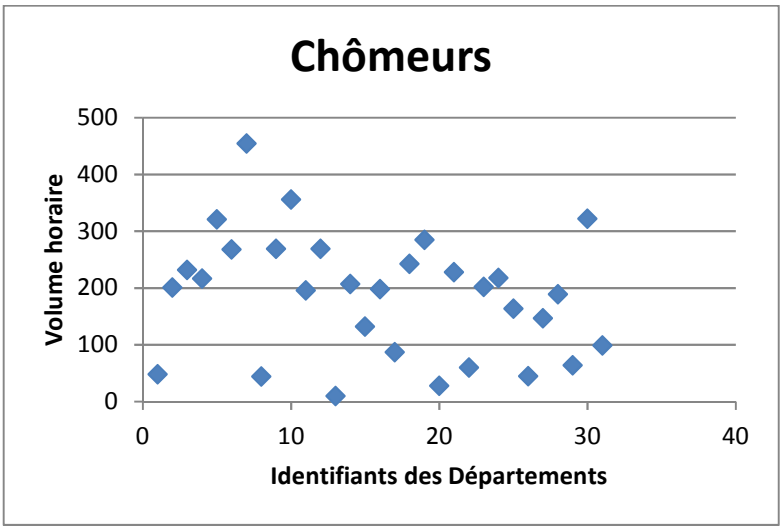
La première étape consiste à s'assurer que l'hypothèse d'une relation linéaire entre les deux variables est plausible, et que les séries ne comportent pas de valeurs aberrantes. On le vérifie en visualisant le nuage des points que constituent les couples (nombre d'heures / actif occupé ; nombre d'heures / chômeur) :



Le nuage de points suggère une tendance linéaire croissante (le nombre d'heures par chômeur tend à augmenter avec le nombre d'heures par actif occupé).

On recherche ensuite la présence de valeurs aberrantes. Les courbes de chacune des deux variables se présentent comme suit :





Le volume horaire par chômeur pour les Départements N°7 (455) et N°13 (10) semblent constituer des valeurs aberrantes. Pour vérifier ce diagnostic, on applique le

test de Grubbs[114]. Le test suppose que les données soit normalement distribuées. On vérifie donc d'abord la normalité :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Occupés	0,18	0,31	0,21	0,62
Chômeurs	0,35	0,31	0,31	0,83

[114] La statistique de Grubbs s'écrit :

$$\begin{cases} G_{max} = \dfrac{x_{max} - \bar{x}}{\sigma} \\ G_{min} = \dfrac{\bar{x} - x_{min}}{\sigma} \end{cases}$$

où

G_{max} , G_{min} représentent les valeurs calculées de la statistique suivant que l'extrémum testé est un maximum ou un minimum ;

x_{max} , x_{min} représentent les extrema de la série ;

$\bar{x}$, σ représentent respectivement la moyenne et l'écart-type de la série.

La valeur calculée de la statistique est comparée à la table de Grubbs. L'extrémum est une valeur aberrante si la valeur calculée de la statistique est supérieure à la valeur critique correspondant au nombre des observations et au taux de significativité choisi. Le test de Grubbs nécessite que la série soit normalement distribuée.

Dans XLSTAT, le test de Grubbs est accessible via la commande *Tests pour les valeurs extrêmes / Test de Grubbs*.

Tous les tests confirment que les variables sont normalement distribuées. Le test de Grubbs ensuite, appliqué à la série du volume horaire par chômeur (la série suspecte), n'indique pas cependant de valeur aberrante au seuil de significativité de 5% :

Valeur observée de la statistique de Grubbs	2,49
Valeur critique	2,92
p-value	0,27

Il peut donc être pertinent d'utiliser le coefficient de Pearson pour caractériser la relation.

La valeur calculée du coefficient de corrélation linéaire de Pearson s'élève à 0,94 avec une p-value inférieure à 0,0001. On peut donc conclure que la durée de formation des chômeurs est fortement corrélée à celle des actifs occupés dans les 31 départements considérés, et que cette corrélation est significative au seuil de 5%.

13.2. MESURE DE LA CORRÉLATION LINÉAIRE ENTRE PLUS DE DEUX VARIABLES QUANTITATIVES

La formule de Pearson ne s'applique que si le nombre de variables est égal à 2. Une méthode simple pour

mesurer la corrélation linéaire entre plus de deux variables quantitatives est celle de la *moyenne des intercorrélations*. Il s'agit alors de calculer la moyenne des corrélations entre variables prises deux à deux.

Soit un ensemble de K variables quantitatives $\{1, 2, \dots, K\}$ pour chacune desquelles on dispose de n observations récapitulées dans un *Tableau des observations* structuré comme suit :

	Variables					
	V1	V2	V3	V4	⋯	VK
Observation 1						
Observation 2						
Observation 3						
⋮						
Observation n						

On calcule d'abord les corrélations entre variables prises deux à deux, c'est-à-dire la corrélation entre V_1 et V_2, la corrélation entre V_1 et V_3, la corrélation entre V_1 et V_4, jusqu'à la corrélation entre V_1 et V_K, puis les corrélations entre V_2 et V_3, V_2 et V_4,jusqu'à la corrélation entre V_2 et V_K, et ainsi de suite. Les résultats peuvent être présentés sous la forme d'une *matrice des corrélations* (qui peut, du reste, être calculée directement par les logiciels à partir du tableau des observations) :

	V_1	V_2	V_3	V_4	$\cdots$	V_K
V_1	1					
V_2	$r_{V_1V_2}$	1				
V_3	$r_{V_1V_3}$	$r_{V_2V_3}$	1			
V_4	$r_{V_1V_4}$	$r_{V_2V_4}$	$r_{V_3V_4}$	1		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
V_K	$r_{V_1V_K}$	$r_{V_2V_K}$	$r_{V_3V_K}$	$r_{V_4V_K}$	$\cdots$	1

Dans cette matrice des corrélations, r_{XY} représente le coefficient de corrélation de Pearson entre les variables X et Y.

La corrélation entre l'ensemble des K variables est mesurée par la *moyenne des intercorrélations*, c'est-à-

dire la moyenne arithmétique des coefficients sous la diagonale (donc en excluant les 1) :

$$r_{V_1V_2\cdots V_K} = \bar{r} = \frac{r_{V_1V_2} + r_{V_1V_3} + \cdots + r_{V_{K-1}V_K}}{K \times \frac{K-1}{2}}$$

EXEMPLE 13.2

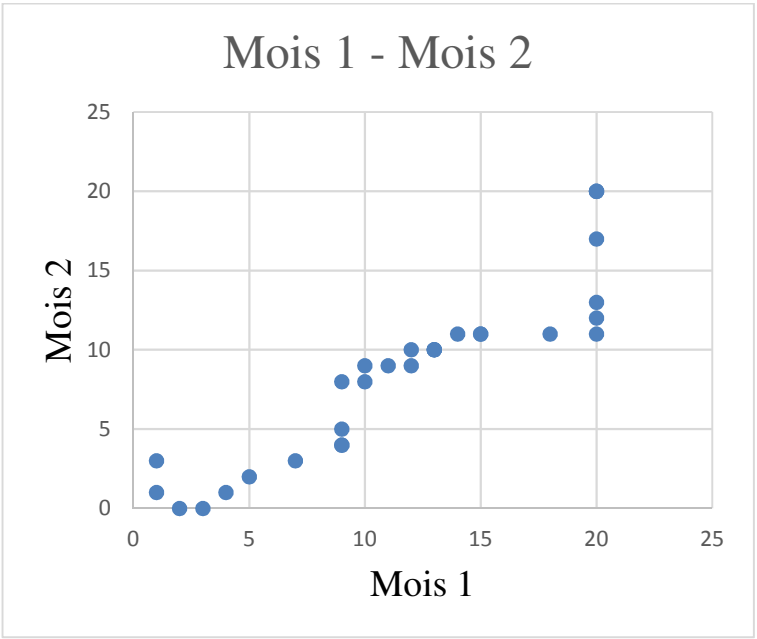
Le tableau suivant présente les notes obtenues par une classe dans une matière au cours de trois évaluations mensuelles successives. On cherche à savoir s'il y a corrélation entre ces notes.

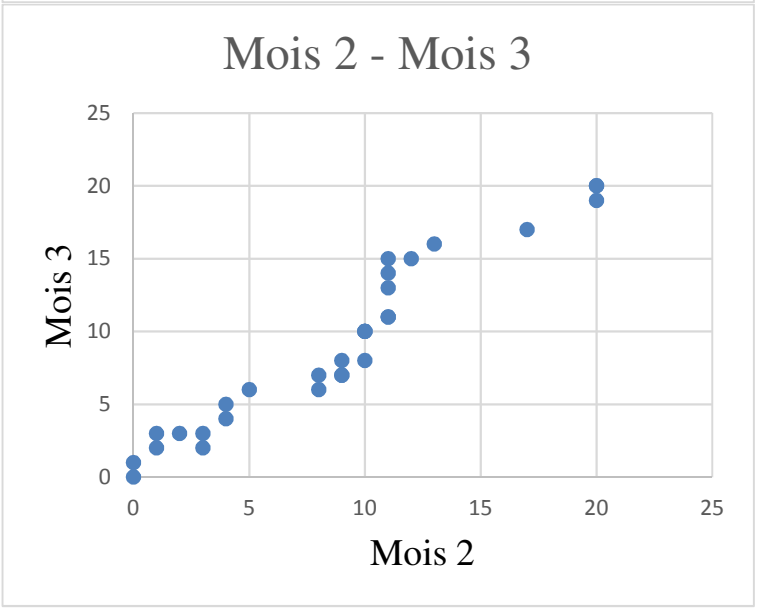
Tableau 13.2.
Notes obtenues par une classe au cours de trois évaluations mensuelles successives dans une même matière

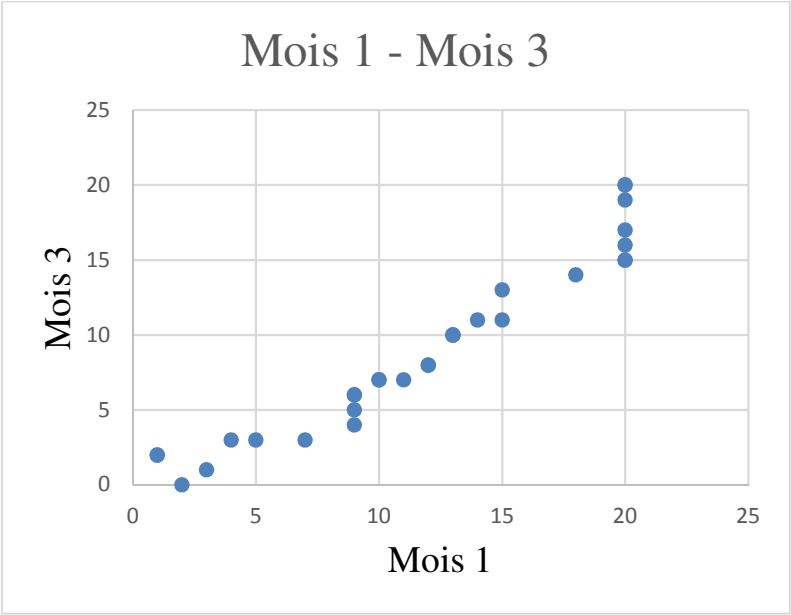
	Mois 1	Mois 2	Mois 3
Élève 1	9	5	6
Élève 2	10	9	7
Élève 3	11	9	7
Élève 4	10	8	7
Élève 5	1	3	2
Élève 6	20	20	19
Élève 7	20	13	16
Élève 8	15	11	13
Élève 9	9	4	4
Élève 10	1	1	2

	Mois 1	Mois 2	Mois 3
Élève 11	4	1	3
Élève 12	20	11	15
Élève 13	3	0	1
Élève 14	9	8	6
Élève 15	12	10	8
Élève 16	9	4	5
Élève 17	12	9	8
Élève 18	13	10	10
Élève 19	5	2	3
Élève 20	2	0	0
Élève 21	13	10	10
Élève 22	14	11	11
Élève 23	7	3	3
Élève 24	15	11	11
Élève 25	18	11	14
Élève 26	20	12	15
Élève 27	20	17	17
Élève 28	13	10	10
Élève 29	20	20	20
Élève 30	20	20	20

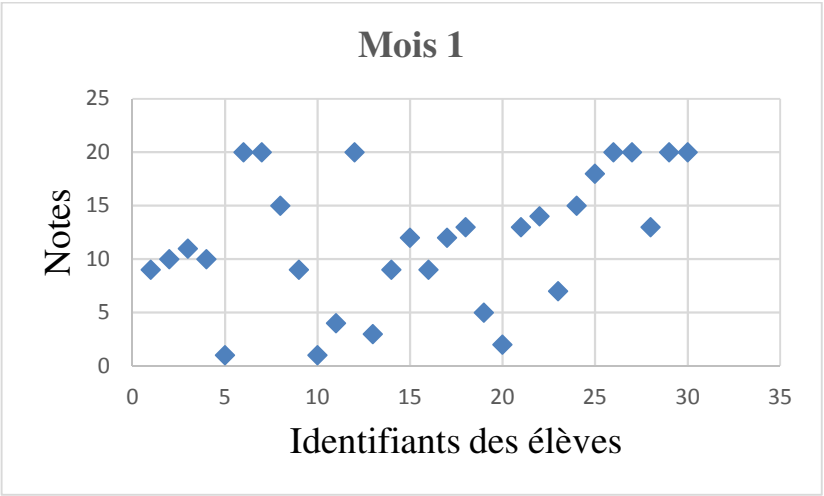
Les nuages de points des variables prises deux à deux suggèrent des relations linéaires croissantes :

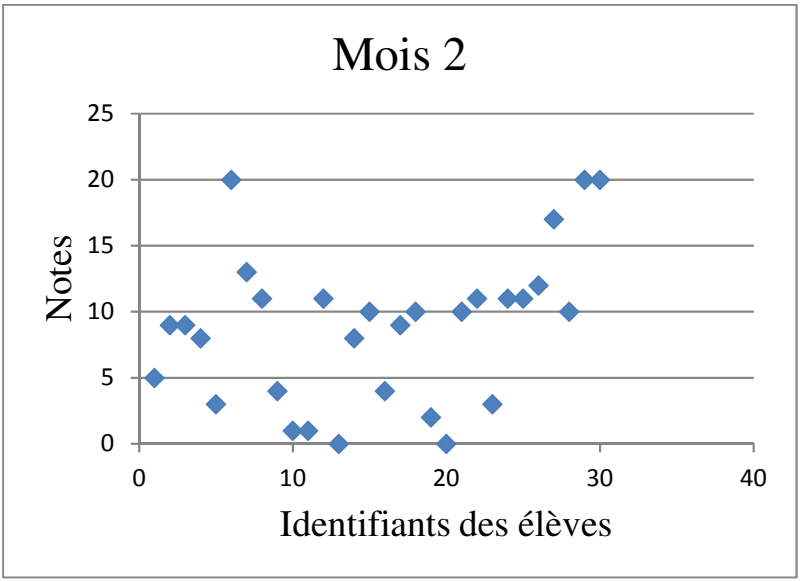

Mois 1 - Mois 2
Mois 2
Mois 1
0
5
10
15
20
25

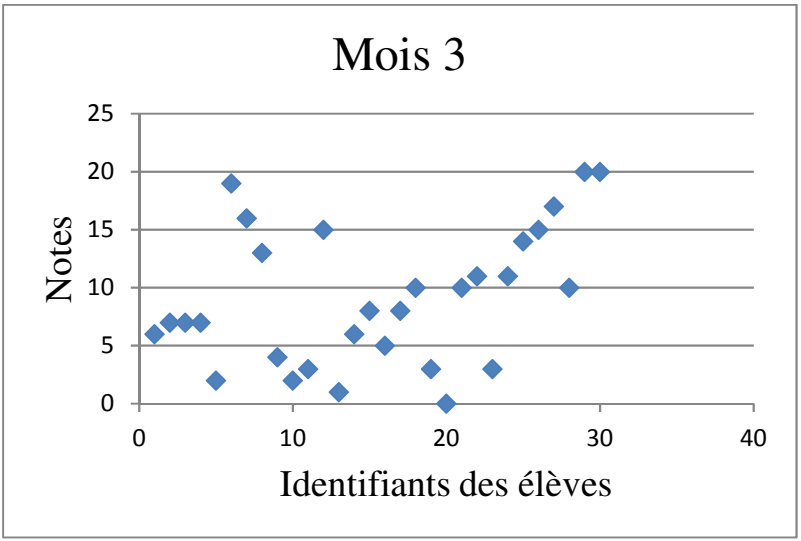

Mois 2 - Mois 3
Mois 3
Mois 2
0
5
10
15
20
25



On cherche ensuite à détecter la présence de valeurs aberrantes. On examine d'abord les courbes des séries :




Mois 2
Notes
25
20
15
10
5
0
0
10
20
30
40
Identifiants des élèves


Mois 3
Notes
25
20
15
10
5
0
0
10
20
30
40
Identifiants des élèves

Les courbes ne suggèrent pas *a priori* la présence de valeurs aberrantes. Il y a bien des valeurs extrêmes (0 et 20), mais elles ne constituent pas des cas isolés. On vérifie cependant le diagnostic au moyen du test de Grubbs. Au préalable, les tests de normalité indiquent que les trois distributions sont gaussiennes :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Mois 1	0,04	0,10	0,11	0,54
Mois 2	0,04	0,05	0,09	0,68
Mois 3	0,18	0,30	0,50	0,42

Le test de Grubbs signale une valeur aberrante :

	Mois 1	Mois 2	Mois 3
Valeur observée de la statistique de Grubbs	1,74	1,95	1,85
Valeur critique	2,90	2,90	2,90
p-value	<0,0001	0,67	0,26

Celle-ci apparaît être la note « 1 » de l'élève 5 pour le mois 1. Cette valeur est donc retirée de l'échantillon pour la suite de l'analyse.

La matrice des corrélations s'établit comme suit :

	Mois 1	Mois 2	Mois 3
Mois 1	1		
Mois 2	0,92	1	
Mois 3	0,96	0,96	1

Les probabilités p indiquent que les trois corrélations sont significatives au seuil de 5% :

	p		
	Mois 1	Mois 2	Mois 3
Mois 1			
Mois 2	< 0.0001		
Mois 3	< 0.0001	< 0.0001	

La moyenne des intercorrélations s'établit à :

$$\bar{r} = \frac{0{,}92 + 0{,}96 + 0{,}96}{3(3-1)/2} = 0{,}94$$

On peut ainsi estimer à 94% la corrélation moyenne entre les notes des trois évaluations.

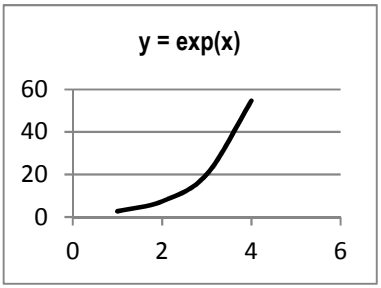
13.3. COEFFICIENT DE CORRÉLATION DE RANG DE SPEARMAN

Le coefficient de Spearman s'utilise lorsque la relation entre deux variables quantitatives, tout en étant monotone[115], n'est pas linéaire.

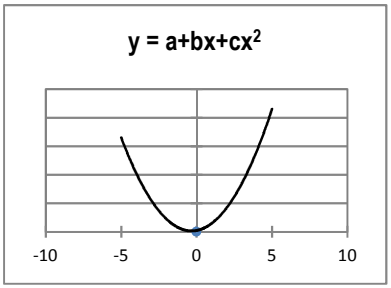
Le coefficient de corrélation de Spearman est un coefficient de *corrélation de rang*. Soit deux variables

[115] Une relation est monotone lorsqu'elle est toujours soit croissante, soit décroissante, c'est-à-dire qu'elle n'admet ni maximum ni minimum. Une relation linéaire est monotone. Une relation non-linéaire peut être monotone ou non-monotone. Dans le cas de relations non-monotones, le coefficient de Spearman ne s'applique pas.

Exemple de relation non-linéaire monotone



Exemple de relation non-linéaire non-monotone



X et Y. La corrélation de rang signifie qu'on examine le rang – et non pas la valeur – des observations x_i et y_i pour chaque individu / unité statistique *i* observé(e).

Par exemple, soit les individus A, B, C et D en lignes du tableau ci-après, et les valeurs x_i et y_i associées à ces individus pour les variables X et Y dans les colonnes 1 et 2 du tableau. Les rangs (par ordre croissant) R_{x_i} et R_{y_i} des valeurs x_i et y_i s'établissent comme l'indiquent les colonnes 3 et 4 :

	Variable X x_i	Variable Y y_i	Rang R_{x_i} dans la distribution X	Rang R_{y_i} dans la distribution Y
	(1)	(2)	(3)	(4)
A	27	7	2	1
B	58	10	3	2
C	12	15	1	3
D	100	20	4	4

Le coefficient de Spearman est défini[116] par :

$$\rho = \frac{cov_{R_X,R_Y}}{\sigma_{R_X} \times \sigma_{R_Y}}$$

où R_X et R_Y représentent respectivement les rangs dans la variable *X* et dans la variable *Y*.

On peut aussi calculer le coefficient de Spearman en appliquant la formule :

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} \left(R_{x_i} - R_{y_i}\right)^2}{n \times (n^2 - 1)}$$

où

- $i = 1, 2, \cdots, N$ sont les individus (ou unité statistiques) observés ;
- n, l'effectif de la série ; et
- R_{x_i} est le rang de l'individu i dans la série X.

Le ρ de Spearman s'interprète suivant les mêmes principes que le r de Pearson : il est compris entre -1 et 1 traduisant une corrélation négative ou positive, faible ou forte, ou une absence de corrélation quand $\rho = 0$.

[116] La définition du coefficient de Spearman suit celle du coefficient de Pearson, à la différence près que le coefficient de Spearman s'applique aux rangs des observations et non aux observations elles-mêmes.

La corrélation doit encore être significative. Le test de significativité s'appuie sur la table du ρ de Spearman[117] et obéit au même principe que pour le test du r de Pearson : l'hypothèse nulle du test postule l'absence de relation entre variables ; l'hypothèse nulle peut être rejetée si la valeur calculée du ρ de Spearman est supérieure à la valeur critique indiquée dans la table pour le nombre d'observations et au taux de significativité considérés ; et la probabilité p indique le risque de se tromper en rejetant l'hypothèse nulle.

Par exemple, dans le tableau ci-après, la valeur du ρ est de -0,8 suggérant une relation non-linéaire forte et négative :

	x_i	y_i	R_{x_i}	R_{y_i}	$R_{x_i}-R_{y_i}$	$\left(R_{x_i}-R_{y_i}\right)^2$
	(1)	(2)	(3)	(4)	(5)	(6)
A	27	20	2	4	-2	4
B	58	10	3	2	1	1
C	12	15	1	3	-2	4
D	100	7	4	1	3	9
					$\sum\left(R_{x_i}-R_{y_i}\right)^2$ = 18	
					ρ = -0,8	

[117] Consultable par exemple sur le site Real Statistics à l'adresse : http://www.real-statistics.com/statistics-tables/spearmans-rho-table/.

Mais $p = 0{,}33$, ce qui indique que cette corrélation n'est pas significative au seuil de significativité de 5%.

EXEMPLE 13.3

On dispose de données sur les financements reçus en 10 ans par 100 établissements de formation professionnelle au titre de la contribution financière des entreprises à la formation d'apprentis (Tableau 13.3). La taille des établissements aussi bien que les montants des contributions perçues sont variables. On cherche à savoir si le volume des financements perçus est corrélé à la taille de l'établissement de formation.

Tableau 13.3.
Financements perçus par 100 établissements de formation professionnelle

Identifiant de l'établissement	Taille de l'établissement	Montant des contributions financières perçues (EUR)
1	249	4372
2	351	66
3	246	1836
4	249	1034
5	281	4480

Identifiant de l'établissement	Taille de l'établissement	Montant des contributions financières perçues (EUR)
6	493	1305
7	382	1144
8	368	1875
9	268	2165
10	218	501
11	387	4468
12	342	3242
13	218	4154
14	223	240
15	237	1259
16	716	17910
17	920	86652
18	957	81545
19	517	85234
20	670	37645
21	926	82688
22	882	67893
23	730	91720
24	905	51682
25	634	73977
26	744	7446

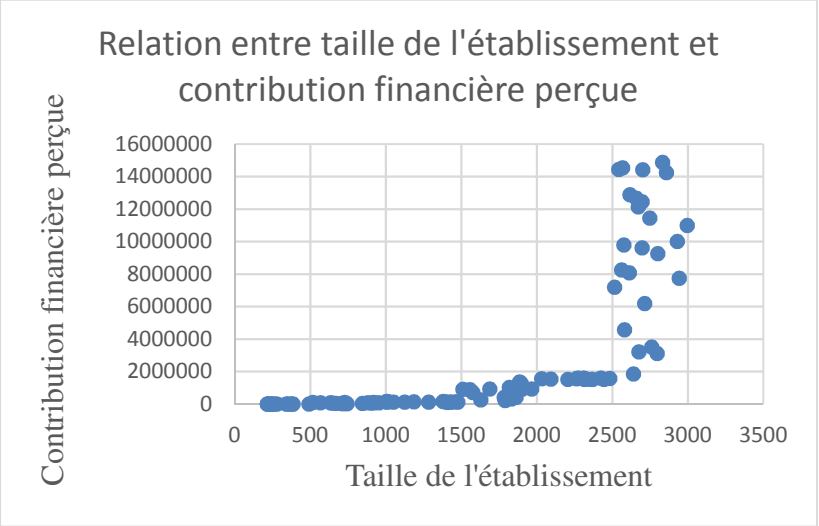
Identifiant de l'établissement	Taille de l'établissement	Montant des contributions financières perçues (EUR)
27	648	44541
28	709	29750
29	568	67587
30	846	38484
31	1432	109794
32	1007	110198
33	1397	126792
34	1005	149560
35	1012	149294
36	1284	118021
37	1013	141032
38	1380	148040
39	1477	110412
40	1052	105771
41	1187	132418
42	1462	124907
43	1125	112228
44	1426	117846
45	1410	101947
46	1689	927499
47	1897	1280895
48	1887	1344651

Identifiant de l'établissement	Taille de l'établissement	Montant des contributions financières perçues (EUR)
49	1818	1018229
50	1882	1088761
51	1864	412053
52	1967	927828
53	1557	878195
54	1791	203780
55	1629	253382
56	1512	907376
57	1834	301742
58	1784	417687
59	1577	694998
60	1901	893777
61	2272	1590573
62	2370	1508857
63	2306	1574904
64	2282	1579697
65	2032	1547309
66	2483	1575568
67	2330	1537139
68	2317	1513850
69	2206	1511822
70	2445	1506252

Identifiant de l'établissement	Taille de l'établissement	Montant des contributions financières perçues (EUR)
71	2360	1516791
72	2263	1551407
73	2093	1536748
74	2425	1582609
75	2312	1585562
76	2702	14406804
77	2676	3203012
78	2698	12451697
79	2931	9997477
80	2715	6172898
81	2577	9794601
82	2641	1846217
83	2858	14233877
84	2612	8078740
85	2659	12663322
86	2797	3109504
87	2515	7184602
88	2672	12123152
89	2801	9245019
90	2543	14438521
91	2562	8253491

Identifiant de l'établissement	Taille de l'établissement	Montant des contributions financières perçues (EUR)
92	2833	14870147
93	2697	9614900
94	2944	7736625
95	2748	11448791
96	2618	12876737
97	2569	14527242
98	2582	4565262
99	2761	3504334
100	2997	10980577

Le nuage de points des deux variables suggère une relation monotone mais non linéaire :



La relation n'étant manifestement pas linéaire, on utilise le coefficient de Spearman. Le calcul du coefficient de Spearman indique $\rho = 0{,}969$ avec $p < 0{,}0001$. On peut donc conclure qu'il y a une très forte corrélation entre la taille de l'établissement et le volume de financement perçu.

On peut comparer ce résultat avec celui que l'on aurait obtenu en appliquant le coefficient de corrélation linéaire de Pearson : $r = 0{,}698$ avec $p < 0{,}0001$. Le résultat aurait effectivement montré l'existence d'une corrélation, mais l'intensité de celle-ci aurait été considérablement sous-estimée.

13.4. COEFFICIENT DE CORRÉLATION DE RANG DE KENDALL

Le coefficient de corrélation de Kendall est aussi un coefficient de corrélation de rang. Il constitue une alternative au coefficient de Spearman et s'utilise dans le même cas de figure, c'est-à-dire en présence d'une relation monotone mais non linéaire entre deux variables quantitatives.

La définition du coefficient de Kendall repose sur les notions de paires concordantes et discordantes. Soit

deux variables X et Y et les observations (x_i, y_i) pour chaque individu i, le coefficient de Kendall s'écrit :

$$\tau = \frac{n_c - n_d}{n \times \frac{n-1}{2}}$$

où
n_c représente le nombre de paires d'observations concordantes dans les deux séries ;
n_d, le nombre de paires discordantes dans les deux séries ; et
n, l'effectif de la série.

On appelle *paire d'observations* l'ensemble des deux couples d'observations $\{(x_i, y_i) ; (x_j, y_j)\}$ pour deux individus i et j. Pour n individus observés, le nombre total de paires d'observations est égal à $n(n-1)/2$. C'est le dénominateur du τ de Kendall. On dit qu'une paire d'observations est *concordante* si la hiérarchie des rangs des individus i et j est la même dans les deux variables, c'est-à-dire si :

$$\begin{cases} R_{x_i} < R_{x_j} \ et \ R_{y_i} < R_{y_j} \\ \qquad ou \\ R_{x_i} > R_{x_j} \ et \ R_{y_i} > R_{y_j} \end{cases}$$

Une paire d'observations est *discordante* si la hiérarchie des rangs est inversée dans la deuxième variable, c'est-à-dire si :

$$\begin{cases} R_{x_i} < R_{x_j} \ mais \ R_{y_i} > R_{y_j} \\ \qquad\qquad ou \\ R_{x_i} > R_{x_j} \ mais \ R_{y_i} < R_{y_j} \end{cases}$$

Une paire d'observations n'est ni concordante ni discordante si les rangs de i et j sont égaux dans X ou dans Y, c'est-à-dire si :

$$R_{x_i} = R_{x_j} \ ou \ R_{y_i} = R_{y_j}$$

Le τ de Kendall peut prendre toute valeur entre -1 et 1. Il est égal à 1 si toutes les paires sont concordantes, et traduit alors une relation forte et de même sens entre les deux variables. Il est égal à -1 si toutes les paires sont discordantes et traduit alors une relation forte et négative entre variables.

Le test de significativité s'appuie sur la table du τ de Kendall[118]. L'hypothèse nulle postule l'absence de relation entre variables. L'hypothèse nulle peut être rejetée si la valeur calculée du τ de Kendall est supérieure à la valeur critique indiquée dans la table pour le nombre d'observations et au taux de

[118] Consultable par exemple sur le site Real Statistics à l'adresse :
http://www.real-statistics.com/statistics-tables/kendalls-tau-table/

significativité considérés. La probabilité p indique le risque de se tromper en rejetant l'hypothèse nulle.

EXEMPLE 13.4

On dispose des observations décrites dans les colonnes 1 et 2 pour les cinq individus A, B, C, D et E :

	X	Y
	(1)	(2)
A	200	10
B	250	7
C	300	9
D	350	8
E	400	5

On cherche à savoir si le coefficient de Kendall révèle une relation entre les variables X et Y.

La première étape consiste à calculer les rangs de chaque individu dans chaque variable :

	X	Y	R_{x_i}	R_{y_i}
	(1)	(2)	(3)	(4)
A	200	10	1	5
B	250	7	2	2
C	300	9	3	4
D	350	8	4	3
E	400	5	5	1

On identifie ensuite la nature concordante (c) ou discordante (d) des paires d'observations :

	A	B	C	D	E
A					
B	d				
C	d	c			
D	d	c	d		
E	d	d	d	d	

Enfin on calcule le coefficient :

$$\tau = \frac{2 - 8}{10} = -0{,}6$$

Il y a donc une corrélation négative assez forte. Mais la probabilité p de se tromper en rejetant l'hypothèse nulle s'établit à 0,22. Donc cette corrélation n'est pas significative au seuil de 5%.

13.5. CORRÉLATION BISÉRIELLE PONCTUELLE

Soit un échantillon *k* de taille *n* caractérisé en termes de deux variables X et Y, l'une quantitative (X), et l'autre (Y) qualitative binaire nominale.

La corrélation entre ces deux variables peut être mesurée par le coefficient de « corrélation bisérielle

ponctuelle »[119] ou « corrélation bisérielle de point » (*point biserial correlation*).

Le coefficient de corrélation bisérielle ponctuelle est défini par :

---

[119] Si *Y* est ordinale ou comporte de façon sous-jacente une continuité entre valeurs de la variable, l'usage du coefficient de corrélation bisérielle ponctuelle peut conduire à des résultats erronés. Il y a continuité sous-jacente lorsque la variable *Y* a été artificiellement rendue binaire. C'est le cas par exemple lorsque les valeurs de la variable sont continues mais qu'on a établi un seuil de part et d'autre duquel les valeurs sont transformées en l'une ou l'autre de deux modalités (par exemple les notes à une épreuve sont transformées en l'une ou l'autre des deux modalités « inférieur à la moyenne », « supérieur à la moyenne »). Lorsque *Y* est ordinale ou comporte une continuité sous-jacente, et à condition que la variable *X* soit normalement distribuée, la procédure appropriée consiste à calculer plutôt le *coefficient de corrélation bisérielle*, défini par :

$$r_b = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_X} \times \frac{n_1 \times n_0}{n^2 \times h}$$

où *h* est l'ordonnée de la courbe de la distribution normale centrée réduite au point qui divise la distribution en deux proportions $\frac{n_0}{n}$ et $\frac{n_1}{n}$. Mais il faut alors pouvoir calculer la valeur de *h*, ce qui rend le recours au coefficient de corrélation bisérielle malaisé, d'autant que les logiciels statistiques ne le proposent pas (lorsque les logiciels mentionnent le coefficient de corrélation « bisérielle », c'est en général le coefficient de corrélation « bisérielle ponctuelle » qui est en réalité calculé).

$$r_{bp} = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_X} \times \sqrt{\frac{n_1 \times n_0}{n^2}}$$

où

- L'indice 0 désigne le groupe des individus pour lesquels Y prend la modalité 0 ("groupe 0") ;
- L'indice 1 désigne le groupe des individus pour lesquels Y prend la modalité 1 ("groupe 1") ;
- n_0 et n_1 représentent les effectifs respectifs des groupes 0 et 1 : $n_1 + n_0 = n$
- $\bar{x}_0$ désigne la moyenne de la variable X dans le groupe 0 ;
- $\bar{x}_1$ désigne la moyenne de la variable X dans le groupe 1 ;
- σ_X représente l'écart-type de la variable X pour l'ensemble des individus (groupe 0 et groupe 1 pris ensemble).

Le coefficient de corrélation bisérielle ponctuelle varie entre -1 et 1 et s'interprète comme les coefficients r, ρ et τ. La significativité du coefficient de corrélation bisérielle ponctuelle s'évalue au moyen de la statistique définie par :

$$t = r_{bp} \times \sqrt{\frac{n-2}{1-r^2}}$$

L'évaluation consiste à comparer la valeur calculée de t avec la valeur critique correspondant au seuil de significativité souhaité pour $DDL = n - 2$ sur une table de Student (voir Annexe 3). L'hypothèse nulle du test de significativité du coefficient de corrélation bisérielle ponctuelle postule l'absence de relation entre variables. Le test de significativité est concluant si la valeur calculée de t est supérieure à la valeur critique pertinente. Les logiciels statistiques indiquent la probabilité p correspondante.

EXEMPLE 13.5

L'enquête internationale TALIS 2013 de l'OCDE sur les enseignants, l'enseignement et l'apprentissage fournit notamment des données sur les opinions des enseignants au sujet de la démarche constructiviste en apprentissage. Le Tableau 13.4 montre les 50 premières réponses des enseignants de lycée en Norvège à la question de savoir s'ils pensent que l'apprenant apprend mieux au travers de la résolution autonome de problèmes. On cherche à savoir s'il y a un lien entre l'âge de l'enseignant et sa position vis-à-vis de l'apprentissage par résolution de problèmes.

Tableau 13.4.
Age et position vis-à-vis de l'apprentissage par résolution de problèmes chez 50 enseignants de lycée en Norvège : réponses à la question "*Pensez-vous que l'apprenant apprend mieux au travers de la résolution autonome de problèmes ?*"

Source : OCDE, Enquête TALIS 2013, série CTGINTT2, indicateurs TT2G02 et TT2G32B

Identifiant de l'enseignant	Age	Position 0 = Non ; 1 = Oui
500101	32	1
500103	41	1
500104	34	1
500105	39	0
500107	31	1
500108	43	0

Identifiant de l'enseignant	Age	Position 0 = Non ; 1 = Oui
500109	53	1
500110	63	0
500111	48	1
500112	40	0
500114	56	1
500115	37	0
500116	53	0
500117	55	0
500118	60	1
500119	42	1
500120	63	1
500121	42	1
500122	59	0
500123	48	1
500126	35	0
500127	31	0
500128	38	1
500129	37	1
500130	43	1
500131	32	0
500132	57	1
500133	60	1
500134	32	0
500201	49	0
500202	54	1
500203	49	1

Identifiant de l'enseignant	Age	Position 0 = Non ; 1 = Oui
500204	62	1
500205	35	1
500207	43	1
500208	49	0
500209	38	1
500210	63	1
500211	46	1
500212	41	1
500214	24	0
500216	39	1
500217	43	1
500218	36	0
500219	24	0
500220	49	0
500221	45	0
500223	36	1
500224	43	0
500225	51	1

Les calculs indiquent un coefficient de corrélation $r_{bp} = -0{,}20$ et une valeur calculée de la statistique t égale à -1,44 contre une valeur critique de l'ordre de 2,01 pour 48 DDL au seuil de significativité de 5%. $p = 0{,}15$. Il n'y a donc pas, au sein de ce groupe d'enseignants, de relation significative entre âge et

position vis-à-vis de l'apprentissage par résolution de problèmes.

13.6. MESURE DE L'INTENSITÉ DE L'ASSOCIATION ENTRE VARIABLES QUALITATIVES

Soit deux variables qualitatives entre lesquelles existe une relation (par exemple une relation détectée à partir d'un test de Khi-carré d'indépendance). Plusieurs méthodes permettent de mesurer l'intensité d'une telle relation. Le choix de la méthode dépend principalement de la nature nominale ou ordinale des variables, et de la taille du tableau de contingence de la relation (on distingue ici les tableaux à deux lignes et deux colonnes – tableaux 2X2 – des tableaux à au moins deux lignes et au moins deux colonnes).

13.6.1. Outils applicables uniquement aux relations entre variables qualitatives ordinales

Entrent dans cette catégorie deux outils de mesure des relations entre variables quantitatives, à savoir les coefficients de corrélation de Spearman et de Kendall.

EXEMPLE 13.6.
Coefficients de corrélation des rangs de Spearman et Kendall appliqués à la mesure de l'intensité de l'association entre variables qualitatives ordinales

L'enquête internationale TALIS 2013 de l'OCDE sur les enseignants, l'enseignement et l'apprentissage fournit notamment des données sur la satisfaction des enseignants au travail. Le Tableau 13.5 montre les 50 premières réponses des enseignants de collège en France aux questions de savoir s'ils considèrent la profession d'enseignant comme socialement valorisée, et s'ils regrettent d'avoir choisi cette profession. On cherche à savoir s'il y a un lien entre la considération sociale perçue et le regret.

Tableau 13.5.
Considération sociale perçue et regret du choix de profession chez 50 enseignants de collège en France

Lecture du tableau

À la question "Pensez-vous que la profession d'enseignant est socialement valorisée", les modalités de réponse sont :

1 – Absolument pas
2 – Plutôt non
3 – Plutôt oui
4 – Oui, absolument

A la question "Regrettez-vous d'avoir choisi la carrière d'enseignant", les modalités de réponse sont :

1 – Absolument pas
2 – Plutôt non
3 – Plutôt oui
4 – Oui, absolument

Source : OCDE, Enquête TALIS 2013, série BTGINTT2, indicateurs TT2G46H et TT2G46D

Identifiant Enseignant	Considération sociale	Regret
300102	2	2
300105	3	2
300106	1	1
300108	2	1
300110	1	1
300112	1	2
300113	2	1
300114	2	2
300115	1	2
300116	2	2
300118	1	2
300201	2	1
300202	2	1

Identifiant Enseignant	Considération sociale	Regret
300203	2	2
300204	2	1
300205	2	1
300206	1	1
300210	3	1
300211	1	1
300212	1	2
300213	1	2
300214	2	1
300215	3	1
300216	2	2
300219	2	1
300220	1	1
300301	1	2
300302	1	2
300303	2	1
300304	2	1
300305	2	1
300306	3	2
300307	1	2
300308	1	2
300309	2	2
300310	1	2
300311	1	2
300312	1	2
300313	2	2

Identifiant Enseignant	Considération sociale	Regret
300314	1	2
300315	1	1
300316	3	1
300317	1	1
300318	1	2
300319	2	3
300320	2	2
300401	2	2
300402	2	1
300403	2	1
300404	1	1

Les résultats indiquent que :

$$\rho = -0{,}18 \text{ avec } p = 0{,}20 \text{ ;}$$
$$\tau = -0{,}17 \text{ avec } p = 0{,}20 \text{ .}$$

Les coefficients de corrélation de Spearman et Kendall sont faibles et non significatifs. Il n'apparaît donc pas, chez ces 50 enseignants, de lien entre l'éventuel regret d'avoir choisi cette profession et leur perception de la considération pour cette profession dans la société.

13.6.2. Outils applicables aux relations entre variables qualitatives, qu'elles soient nominales ou ordinales

On distingue ici suivant la taille du tableau de contingence, c'est-à-dire entre tableaux 2X2 (tableaux de contingence à deux lignes et deux colonnes) et généralité des tableaux de relations entre deux variables (tableaux de contingence à au moins deux lignes et au moins deux colonnes). Les outils applicables à la généralité des tableaux s'appliquent aussi aux tableaux 2x2, la réciproque n'est pas vraie.

13.6.2.1. Outils spécifiques aux tableaux 2X2

Soit une relation entre deux variables qualitatives binaires décrite par un tableau de contingence à deux lignes et deux colonnes où O_{lc} désigne l'effectif observé pour la l-ième ligne de la c-ième colonne. Deux coefficients sont spécifiques à ce cas de figure : le Q et le Y de Yule.

13.6.2.1.1. Q de Yule

Le coefficient Q de Yule est défini par :

$$\Phi = \frac{(O_{11} \times O_{22}) - (O_{12} \times O_{21})}{(O_{11} \times O_{22}) + (O_{12} \times O_{21})}$$

Q varie entre -1 et 1. L'intensité de la relation est d'autant plus faible que Q est proche de zéro. L'intensité est dite :

- négligeable quand $0{,}01 \leq |Q| \leq 0{,}09$;
- faible quand $0{,}10 \leq |Q| \leq 0{,}29$;
- modérée quand $0{,}30 \leq |Q| \leq 0{,}49$;
- forte quand $0{,}50 \leq |Q| \leq 0{,}69$;
- très forte quand $0{,}70 \leq |Q| \leq 1$.

13.6.2.1.2. Y de Yule

Le coefficient Y de Yule est défini par :

$$\Phi = \frac{\sqrt{(O_{11} \times O_{22})} - \sqrt{(O_{12} \times O_{21})}}{\sqrt{(O_{11} \times O_{22})} + \sqrt{(O_{12} \times O_{21})}}$$

Y varie entre -1 et 1. L'intensité de la relation est d'autant plus faible que Y est proche de zéro.

EXEMPLE 13.7.
Coefficients Q de Yule et Y de Yule appliqués à la mesure de l'intensité de l'association entre variables qualitatives binaires

On dispose de données sur le nombre d'admis par genre et filière (série générale et série professionnelle) au diplôme national du brevet (Tableau 13.6). On cherche

à savoir s'il y a une relation entre le genre et la filière de réussite, et si oui son intensité.

Tableau 13.6.
Nombre d'admis au diplôme national du brevet selon le sexe et la filière – 2016 – France métropolitaine et Départements d'outremer

Source : Ministère de l'éducation nationale, *Repères et références statistiques 2017*, chapitre 8.9, page 223, tableau 2,
http://cache.media.education.gouv.fr/file/2017/83/0/depp-rers-2017-maj-dec-2017_861830.pdf

	Filière	
	Série générale	Série professionnelle
Filles	341553	22781
Garçons	311878	36327

On applique d'abord un test de Khi-carré d'indépendance. Les calculs montrent que tous les effectifs théoriques sont supérieurs à 5. La valeur calculée de la statistique de test s'élève à 4089,05 contre une valeur critique de 3,84 pour 1 DDL au seuil de significativité de 5%. La probabilité *p* est inférieure à 0,0001. Il y a donc, parmi les admis, une relation significative entre le genre et la filière de réussite.

On calcule ensuite :

- le Q de Yule, égal à 0,27 sur une échelle de -1 (intensité négative maximale) à +1 (intensité positive maximale) ;
- le Y de Yule, égal à 0,13 sur la même échelle de -1 à +1.

Les deux coefficients sont faibles. On peut donc considérer que la relation observée, bien que significative, est de faible intensité.

13.6.2.2. Outils applicables aux tableaux de contingence ayant au moins deux lignes et au moins deux colonnes en général

Différents outils permettent de mesurer l'intensité de l'association à partir de tableaux de contingence ayant au moins deux lignes et au moins deux colonnes. On distingue suivant que la mesure porte sur l'intensité de la relation entre variables toutes modalités confondues (c'est-à-dire l'intensité de la relation toutes catégories de l'échantillon confondues), ou au contraire sur l'intensité de la relation entre certaines modalités des variables (c'est-à-dire l'intensité de la relation pour telle ou telle case du tableau de contingence).

13.6.2.2.1. Mesure de l'intensité globale

Il s'agit ici de mesurer l'intensité de la relation entre variables, sans distinguer entre les différentes catégories. Trois coefficients parmi les plus fréquemment utilisés présentent le double avantage d'être à la fois communément proposés par les logiciels et aisément interprétables : le Φ de Pearson, le V de Cramer et le T de Tschuprow.

13.6.2.2.1.1. Le Φ de Pearson

Le coefficient Φ de Pearson est défini par :

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

où
χ^2 est le Khi-carré calculé à partir du tableau de contingence ; et
n, l'effectif de l'échantillon.

Dans le cas général, la valeur du Φ varie dans l'intervalle :

$$\left[0\,;\, min\left\{\sqrt{nombre\ de\ lignes}\,;\,\sqrt{nombre\ de\ colonnes}\right\}\right]$$

Lorsque le tableau de contingence ne contient que deux lignes et deux colonnes, la valeur du Φ est comprise entre -1 et +1. L'intensité de la relation est d'autant plus faible que Φ est proche de zéro.

EXEMPLE 13.8

On dispose de données sur le nombre d'admis par origine sociale et filière (série générale et série professionnelle) au diplôme national du brevet (Tableau 13.7). On cherche à savoir s'il y a une relation entre l'origine sociale et la filière de réussite, et si oui, son intensité.

Tableau 13.7.
Nombre d'admis au diplôme national du brevet selon l'origine sociale et la filière – 2016 – France métropolitaine et Départements d'outre-mer

Source : Ministère de l'éducation nationale, *Repères et références statistiques 2017*, chapitre 8.9, page 223, tableau 4,
http://cache.media.education.gouv.fr/file/2017/83/0/depp-rers-2017-maj-dec-2017_861830.pdf

	Filière	
Catégorie socioprofessionnelle du chef de famille	Série générale	Série professionnelle
Agriculteurs exploitants	11645	1438
Artisans, commerçants, chefs d'entreprise	60305	4598
Cadres, professions intellectuelles supérieures	151702	2931

	Filière	
Catégorie socioprofessionnelle du chef de famille	Série générale	Série professionnelle
Professions intermédiaires	98820	5699
Employés	134343	12976
Ouvriers	122494	16559
Retraités	7059	582
Inactifs	49829	8974
Indéterminé	17234	5351

Les calculs montrent que tous les effectifs théoriques sont supérieurs à 5. La valeur calculée du $\chi^2 = 22917{,}04$ contre une valeur critique de 15,51 pour 8 DDL au seuil de significativité de 5%. La probabilité *p* est inférieure à 0,0001. Il y a donc association entre origine sociale et filière de réussite.

Dès lors, il y a du sens à calculer le coefficient Φ. Le tableau de contingence permet de calculer $n = 712539$. Donc $\Phi = \sqrt{\frac{22917{,}04}{712539}} = 0{,}179$. Il s'agit ensuite d'interpréter le Φ calculé. Le tableau de contingence comporte deux colonnes et neuf lignes, donc le plus petit des nombres $\sqrt{nombre\ de\ lignes}$ et

$\sqrt{nombre\ de\ colonnes}$ est $\sqrt{2} = 1{,}41$. Donc la valeur de Φ calculée (0,179) apparaît plutôt faible sur une échelle de 0 à 1,41. On peut donc conclure que bien qu'il y ait un lien significatif entre origine sociale et filière de réussite, ce lien est de faible intensité.

13.6.2.2.1.2. Le V de Cramer

Le V de Cramer est défini par :

$$V = \sqrt{\frac{\chi^2}{n \times min\{(C-1)\ ;\ (L-1)\}}}$$

où

- χ^2 est le Khi-carré calculé à partir du tableau de contingence ;
- n, l'effectif de l'échantillon ;
- $min\{(C-1)\ ;\ (L-1)\}$, celui des deux nombres $(nombre\ de\ colonnes - 1)$ et $(nombre\ de\ lignes - 1)$ qui est le plus petit.

La valeur du V de Cramer varie entre 0 et 1 (intensité maximale).

EXEMPLE 13.9

En reprenant les valeurs de l'exemple 13.8, soit

$$\begin{cases} \chi^2 = 22917{,}04 \\ n = 712539 \\ L = 9 \\ C = 2 \end{cases}$$

on voit que le plus petit des nombres $(nombre\ de\ colonnes - 1)$ et $(nombre\ de\ lignes - 1)$ est $(nombre\ de\ colonnes - 1) = 2 - 1 = 1$. Donc $V = \sqrt{\frac{22917{,}04}{7125397 \times 1}} = 0{,}179$. Cette valeur de coefficient est faible également sur une échelle de 0 à 1, ce qui confirme que la relation observée est de faible intensité.

13.6.2.2.1.3. Le T de Tschuprow

Le coefficient d'association de Tschuprow est défini par :

$$T = \sqrt{\frac{\chi^2}{n \times \sqrt{(L-1) \times (C-1)}}}$$

La valeur du T de Tschuprow varie entre 0 et 1.

EXEMPLE 13.10

En reprenant les valeurs de l'exemple 13.8, soit

$$\begin{cases} \chi^2 = 22917{,}04 \\ n = 712539 \\ L = 9 \\ C = 2 \end{cases}$$

La valeur de T s'élève à :

$$T = \sqrt{\frac{22917{,}04}{712539 \times \sqrt{(9-1)\times(2-1)}}} = 0{,}106 \; sur\ une\ échelle\ de\ 0 \text{ à } 1$$

Ce qui confirme encore la faible intensité du lien entre les deux variables.

13.6.2.2.2. Mesure de l'intensité locale : le pourcentage d'écart maximum (PEM) local

La méthode du pourcentage d'écart maximum local (PEML) a été proposée par Cibois (1993). Le PEML permet de mesurer le degré d'attraction ou de répulsion entre deux modalités de variables définissant une case du tableau de contingence[120].

[120] Il ne s'agit donc pas ici de mesurer la *significativité* par case (voir par exemple page 288), mais de mesurer l'*intensité* de la relation par case.

Soit un tableau de contingence décrivant une population / un échantillon en termes de modalités de deux variables, la variable 1 (de modalités $1.1, 1.2, \dots, 1.\xi$) et la variable 2 (de modalités $2.1, 2.2, \dots, 2\psi$) :

Tableau de contingence observé

		Variable 2				
		Modalité 2.1	Modalité 2.2	..	Modalité 2.ψ	Total
Variable 1	Modalité 1.1	O_{11}	O_{12}	..	$O_{1\psi}$	$\mu_{1.1}$
	Modalité 1.2	O_{21}	O_{22}	..	$O_{2\psi}$	$\mu_{1.2}$
	⋮	⋮	⋮	⋮	⋮	⋮
	Modalité 1.ξ	$O_{\xi 1}$	$O_{\xi 2}$	..	$O_{\xi\psi}$	$\mu_{1.\xi}$
	Total	$\mu_{2.1}$	$\mu_{2.2}$	..	$\mu_{2.\psi}$	n

On appelle *pourcentage d'écart maximum local* pour une case lc, définie par l'intersection entre la ligne l et la colonne c, le rapport :

$$PEML_{lc} = \frac{O_{lc} - T_{lc}}{max_{lc} - T_{lc}}$$

$$avec \begin{cases} max_{lc} = min\{\mu_{1.l}\ ;\ \mu_{2.c}\}\ si\ O_{lc} > T_{lc} \\ max_{lc} = 0\ si\ O_{lc} < T_{lc} \end{cases}$$

où

- O_{lc} est l'effectif observé de la case lc ;
- T_{lc} , l'effectif théorique de la case lc. L'effectif théorique est défini à partir des sommes marginales[121] :

$$T_{lc} = \frac{\mu_{1.l} \times \mu_{2.c}}{n}$$

- max_{lc} est la valeur maximale que pourrait atteindre l'effectif observé de la case lc. Cette valeur maximale est nulle en cas de répulsion, or il y a répulsion si l'effectif observé est inférieur à l'effectif théorique.

Le PEML n'est généralement pas proposé par les logiciels. Il faut donc le calculer « à la main » sur un tableur.

EXEMPLE 13.11

Le tableau 13.8 ci-après présente la répartition d'une population en 9 cases en fonction de deux variables qualitatives comprenant chacune trois modalités.

[121] Voir chapitre 11.

Tableau 13.8.
Exemple pour le calcul du PEML

		Variable 2			
		Modalité 2.1	Modalité 2.2	Modalité 2.3	Total
Variable 1	Modalité 1.1	84	11	66	161
	Modalité 1.2	2788	147	56	2991
	Modalité 1.3	7	62	64	133
	Total	2879	220	186	3285

Les calculs indiquent que tous les effectifs théoriques sont supérieurs à 5. $\chi^2 = 1321{,}62$ contre une valeur critique de 9,48 pour 4 DDL. $p < 0{,}0001$. Il y a donc entre les deux variables une relation significative au seuil de significativité de 5%.

On cherche à en savoir davantage sur l'intensité de la relation. On mesure d'abord l'intensité globale. Le calcul des coefficients de Pearson, Cramer et Tschuprow indique :

- $\Phi = 0{,}63$ sur une échelle de 0 à 1,73 ;
- $\mathrm{V} = 0{,}45$ sur une échelle de 0 à 1 ;
- $\mathrm{T} = 0{,}45$ sur une échelle de 0 à 1.

La relation d'ensemble peut donc être considérée comme d'intensité modérée. On cherche à présent à mesurer l'intensité de la relation case par case.

On peut commencer par analyser la significativité par case. Le tableau de significativité par case indique que l'écart entre effectif observé et effectif théorique est significatif pour toutes les cases à l'exclusion de la deuxième sur la première ligne :

	Modalité 2.1	Modalité 2.2	Modalité 2.3
Modalité 1.1	<	(>)	>
Modalité 1.2	>	<	<
Modalité 1.3	<	>	>

On mesure ensuite l'intensité de la relation entre modalités pour chaque case.

Les étapes du calcul des PEM locaux s'établissent comme suit :

Effectifs observés				
	Modalité 2.1	Modalité 2.2	Modalité 2.3	Total
Modalité 1.1	84	11	66	161
Modalité 1.2	2788	147	56	2991
Modalité 1.3	7	62	64	133
Total	2879	220	186	3285

Effectifs théoriques			
	Modalité 2.1	Modalité 2.2	Modalité 2.3
Modalité 1.1	141,10	10,78	9,12
Modalité 1.2	2621,34	200,31	169,35
Modalité 1.3	116,56	8,91	7,53

Effectifs observés – Effectifs théoriques			
	Modalité 2.1	Modalité 2.2	Modalité 2.3
Modalité 1.1	-57,10	0,22	56,88
Modalité 1.2	166,66	-53,31	-113,35
Modalité 1.3	-109,56	53,09	56,47

Maximum de la valeur que pourrait atteindre l'effectif observé (max_{lc})			
	Modalité 2.1	Modalité 2.2	Modalité 2.3
Modalité 1.1	0	161	161
Modalité 1.2	2879	0	0
Modalité 1.3	0	133	133

Calcul des PEM locaux			
	Modalité 2.1	Modalité 2.2	Modalité 2.3
Modalité 1.1	-57,10/-141,10	0,22/(161-10,78)	56,88/(161-9,12)
Modalité 1.2	166,66/(2879-2621,34)	-53,31/-200,31	-113,35/-169,35
Modalité 1.3	-109,56/-116,56	53,09/(133-8,91)	56,47/(133-7,53)

Tableau des PEM locaux			
	Modalité 2.1	Modalité 2.2	Modalité 2.3
Modalité 1.1	0,40	0,00	0,37
Modalité 1.2	0,65	0,27	0,67
Modalité 1.3	0,94	0,43	0,45

Tous les PEML sont positifs, mais il faut se souvenir qu'ils traduisent une répulsion lorsque la valeur observée était inférieure à la valeur théorique. On constate donc :

- une très forte répulsion entre modalités 1.3 et 2.1. ;

- une forte répulsion entre modalités 1.2 et 2.3 ;

- une répulsion modérée entre modalités 1.1 et 2.1, et entre modalités 1.2 et 2.2 ;
- une forte attraction entre modalités 1.2 et 2.1 ;
- une attraction modérée entre modalités 1.1 et 2.3 ; 1.3 et 2.2 ; 1.3 et 2.3 ;
- et une indépendance entre modalités 1.1 et 2.2 (mais l'absence de relation significative entre ces deux modalités avait déjà été mise en évidence par le tableau de significativité par case).

Chapitre 14. Repérer des facteurs sous-jacents : l'analyse factorielle exploratoire

14.1. PROBLÉMATIQUE

L'analyse factorielle (*factor analysis*) ou analyse factorielle exploratoire (*exploratory factor analysis – EFA*) ne doit pas être confondue avec l'analyse factorielle des correspondances (AFC). L'AFC vise à analyser la dépendance entre deux variables qualitatives à partir de leur tableau de contingence en hiérarchisant les relations de dépendance. Elle n'est pas présentée dans cet ouvrage. Dans l'analyse factorielle exploratoire[122], on dispose de données quantitatives sur un grand nombre de variables (en pratique une demi-douzaine au moins) pour les individus d'un échantillon, et on cherche à identifier un nombre limité (deux ou trois en général) de facteurs sous-jacents qui permettraient de synthétiser ces variables. C'est le cas par exemple lorsqu'on cherche à (ré)organiser les items d'un questionnaire en domaines cohérents distincts, chaque facteur sous-jacent constituant ici un domaine dont les items sont les variables relevant de ce facteur. L'analyse factorielle exploratoire permet aussi de caractériser les individus de l'échantillon en fonction du facteur dont ils relèvent. C'est à l'analyse factorielle exploratoire qu'est consacré le présent chapitre.

---

[122] Accessible dans XLSTAT via la commande *Analyse des données / Analyse factorielle*.

L'analyse factorielle présente une certaine proximité avec l'analyse classificatoire dans la mesure où les deux méthodes peuvent conduire à constituer des groupes d'individus. Ces méthodes n'ont pas le même objectif cependant : identifier les facteurs sous-jacents qui permettent de regrouper les variables dans le cas de l'analyse factorielle ; regrouper les individus entre eux sur la base de leur proximité au regard des variables, et sans recherche de liens sous-jacents, dans le cas de l'analyse classificatoire. La distinction est cependant moins nette avec la classification par classes latentes[123].

14.2. DÉMARCHE GÉNÉRALE DE L'ANALYSE FACTORIELLE EXPLORATOIRE

On dispose par exemple des données suivantes, relatives aux notes d'un groupe d'élèves ainsi que de caractéristiques démographiques, sociales et économiques de ces élèves (Tableau 14.1) :

Notes
- Français (F)
- Anglais (A)
- Maths (M)
- Histoire (H)

[123] Voir chapitre 5.

- Éducation physique et sportive (EPS)

Caractéristiques démographiques, sociales et économiques
- Âge
- Diplôme le plus élevé du chef de famille (DP)[124]
- Revenu mensuel du chef de famille, en Euro (RM)
- Nombre de frères et sœurs (FS)
- Distance domicile-établissement, en kms (DDE).

On cherche à identifier les facteurs sous-jacents qui permettraient de catégoriser ces élèves de façon synthétique.

Tableau 14.1
Notes et caractéristiques démographiques, sociales et économiques d'un groupe de 25 élèves

	Notes					Caractéristiques				
	F	A	M	H	EPS	Âge	DP	RM	FS	DDE
1	11	17	10	2	1	14	0	8390	0	0
2	9	12	7	14	17	19	3	5060	0	0
3	19	14	18	9	9	14	5	2695	3	1
4	7	17	8	2	4	17	1	5166	5	1
5	18	1	5	14	8	14	1	10894	2	1
6	18	9	17	19	2	18	6	3409	3	3
7	16	19	18	10	13	19	4	3463	0	4

[124] 0 = Aucun diplôme ; 1 = Fin d'études primaires ; 2 = Collège ; 3 = Lycée ; 4 = Bac+2 ; 5 = Licence ; 6 = Master ; 7 = Doctorat

	Notes					Caractéristiques				
	F	A	M	H	EPS	Âge	DP	RM	FS	DDE
8	12	17	14	1	13	16	4	3675	2	4
9	4	9	8	0	2	17	4	5969	1	4
10	1	16	13	17	0	16	5	7515	0	4
11	11	5	8	1	16	16	2	7324	0	5
12	6	4	19	19	10	16	5	11734	4	5
13	8	12	16	17	10	17	7	3269	3	6
14	3	17	3	8	5	14	0	2605	0	6
15	7	3	6	6	12	14	2	10735	5	6
16	15	20	14	10	1	16	7	3080	3	7
17	15	13	9	10	3	15	5	3697	1	7
18	5	12	17	18	17	14	5	7408	4	7
19	11	16	17	12	4	18	7	2622	0	8
20	2	17	9	15	9	15	3	10550	0	8
21	7	4	12	11	8	14	0	10525	0	9
22	5	14	1	8	7	18	2	11360	4	9
23	13	10	20	14	0	16	5	3740	1	10
24	9	18	14	18	7	14	6	3269	1	10
25	9	15	13	10	7	18	5	3568	2	10

Une fois constitué le fichier de données à analyser, la démarche de l'analyse factorielle comporte trois principales étapes.

14.2.1. Tableau des valeurs propres

La première étape consiste à dresser[125] et à analyser le « tableau des valeurs propres », qui indique le nombre total de facteurs possibles, ainsi que la « valeur propre » et la contribution de chaque facteur à la variabilité totale des données.

Dans l'exemple étudié ici, le tableau des valeurs propres se présente comme suit[126] :

	F1	F2	F3	F4	F5	F6
Valeur propre	2,673	2,139	1,141	0,549	0,213	0,092
Variabilité (%)	24,300	19,449	10,373	4,988	1,938	0,836
% cumulé	24,300	43,749	54,122	59,110	61,047	61,883

[125] En pratique, les logiciels fournissent directement ces tableaux en même temps que les autres résultats des calculs.

[126] Différentes méthodes d'extraction des facteurs latents peuvent être utilisées, notamment la méthode du maximum de vraisemblance (qui suppose que les variables sont normalement distribuées) et la méthode des facteurs principaux. La valeur propre d'un facteur peut varier suivant la méthode d'extraction utilisée.

Dans cet exemple, le tableau indique qu'il pourrait y avoir jusqu'à 6 facteurs (F1 à F6).

En application du *critère de Kaiser*, on ne retient dans l'analyse que les facteurs ayant une valeur propre supérieure à 1 (c'est-à-dire des facteurs qui sont chacun au moins l'équivalent d'une variable).

La deuxième ligne indique la contribution de chaque facteur à la variabilité totale des données, soit 24,30% de la variabilité totale pour le premier facteur, 19,44% pour le 2ème, et ainsi de suite. On voit que chaque facteur supplémentaire n'apporte qu'une contribution de plus en plus faible. Étant donné que l'exercice consiste à repérer des facteurs qui synthétisent et permettent de réduire le nombre de variables à manipuler, il n'y a pas lieu de prendre en compte un trop grand nombre de facteurs, ce qui amène à laisser de côté les facteurs dont la contribution est faible.

Dans cet exemple, on peut se limiter aux trois premiers facteurs (F1 à F3), seuls à vérifier le critère de Kaiser, et au-delà desquels la contribution additionnelle devient trop faible.

14.2.2. Tableau des coordonnées factorielles

La deuxième étape consiste à dresser et analyser le « *tableau des coordonnées factorielles* ». Les

coordonnées factorielles (ou poids factoriels, « *factor loadings* ») sont les coefficients de corrélation entre les variables et les facteurs[127]. Le tableau des coordonnées factorielles indique à quel facteur chaque variable est le plus étroitement liée[128] ainsi que la « communalité » de chaque variable, c'est-à-dire la variance de chaque variable qui peut être expliquée par les autres variables. Les variables ayant une communalité inférieure à 40% peuvent être exclues de l'analyse car on peut considérer qu'elles ont peu de liens avec les autres variables[129].

---

[127] Ce qui peut du reste permettre d'écrire chaque facteur comme une combinaison linéaire des variables qui le composent, par exemple ici $F1 = 0{,}89 \times Diplôme\ du\ chef\ de\ famille + 0{,}62 \times note\ en\ maths + \cdots$

[128] Les variables associées à un facteur peuvent différer suivant la méthode d'extraction de facteurs latents utilisée. Par ailleurs, afin de mieux discerner le positionnement des variables par rapport aux facteurs, et faciliter l'interprétation des résultats, des techniques de « rotation » sont usuellement appliquées. Le principe des rotations est d'accentuer la nature des corrélations entre variables et facteurs en minimisant les faibles corrélations et en maximisant les fortes corrélations, de façon à faire plus clairement apparaître les variables dominantes pour chaque facteur. Les rotations « orthogonales » (par exemple « Varimax ») s'appliquent lorsque les facteurs ne sont pas corrélés, les rotations « obliques » (par exemple « Quartimin », « Oblimin ») lorsqu'ils le sont.

[129] Voir en liste des références Costello & Osborne (2005).

Dans cet exemple, le tableau des coordonnées factorielles se présente comme suit :

	F1	F2	F3	F4	Communalité		Variance
					Initiale	Finale	spécifique
Âge	0,22	0,24	-0,09	0,17	0,259	0,146	0,854
F	0,15	0,61	0,26	-0,55	0,401	0,775	0,225
A	0,38	0,23	-0,62	0,17	0,461	0,621	0,379
M	0,62	0,27	0,32	0,03	0,639	0,577	0,423
H	0,44	-0,13	0,41	0,17	0,505	0,413	0,587
EPS	-0,17	-0,03	0,23	0,07	0,103	0,091	0,909
DP	0,89	0,23	0,26	0,23	0,735	0,987	0,013
RM	-0,58	-0,54	0,40	0,14	0,651	0,830	0,170
FS	-0,03	-0,02	0,29	0,05	0,211	0,092	0,908
DDE	0,60	-0,74	-0,10	-0,19	0,986	0,963	0,037

Le tableau indique que :

- le facteur F1 est, de façon principale, positivement lié au diplôme du chef de famille, et aux notes en maths et histoire, et négativement lié au revenu du chef de famille. La distance, dont le coefficient de corrélation est élevé (0,60), est cependant encore plus corrélée avec le facteur F2 ;
- le facteur F2 est, de façon principale, positivement lié à la note en français, et négativement à la distance domicile – établissement. Le revenu, fortement corrélé, l'est cependant encore plus avec le facteur F1 ;

- le facteur F3 se caractérise essentiellement par son lien négatif avec la note en anglais. La note en EPS et la taille de la fratrie ont une communalité trop faible.

Il apparaît donc *a priori* que trois profils se détachent dans l'échantillon. Le premier est celui d'individus issus de familles dont le chef a un niveau d'éducation élevé mais un revenu faible, et qui excellent en maths et histoire. Le deuxième est celui d'individus caractérisés par une proximité géographique entre domicile et établissement, et qui excellent en français. Le troisième caractérise un groupe d'individus réfractaires à l'anglais. Les trois profils constituent les facteurs sous-jacents.

Cependant, il importe de vérifier la solidité des liens entre variables ainsi établis. On calcule à cette fin le coefficient alpha de Cronbach[130]. Dans cet exemple, les valeurs du coefficient alpha de Cronbach s'établissent comme suit :

	Alpha de Cronbach
F1	0,364
F2	0,177
F3	-0,442

[130] Voir section 2.1.

Ce qui suggère que les facteurs ne sont pas homogènes/cohérents (chaque facteur mesure vraisemblablement plus qu'un seul construit).

14.2.3. Coordonnées des observations

La troisième étape consiste à dresser et analyser le « *tableau des coordonnées des observations* », qui indique à quel facteur chaque individu est le plus étroitement lié.

Dans cet exemple, le tableau des coordonnées des observations se présente comme suit :

Observation	F1	F2	F3	F4
1	-1,747	0,769	-0,611	-0,078
2	-0,856	0,955	-0,038	0,915
3	0,182	1,814	0,604	-0,494
4	-1,468	0,488	-1,364	0,110
5	-1,661	0,294	1,691	-1,220
6	0,383	1,665	0,877	-0,409
7	0,126	1,406	-0,468	-0,462
8	0,067	1,048	-0,517	0,065
9	-0,277	0,234	-0,249	1,225
10	0,142	0,068	0,006	2,274
11	-0,721	-0,020	0,308	-0,524
12	-0,182	-0,511	2,137	1,230
13	1,064	0,678	0,357	1,042
14	-0,882	-0,211	-2,610	-0,394

Observation	F1	F2	F3	F4
15	-0,963	-0,961	0,847	-0,147
16	1,447	0,704	-0,156	-0,143
17	0,819	0,173	-0,078	-0,882
18	0,533	-0,865	0,776	0,772
19	1,638	-0,091	-0,070	0,067
20	-0,021	-1,452	-0,150	1,003
21	-0,957	-1,819	0,161	-1,384
22	-0,581	-1,695	-0,481	-0,083
23	1,147	-0,751	0,071	-1,493
24	1,610	-0,624	-0,415	-0,169
25	1,156	-1,297	-0,630	-0,821

Le tableau indique le positionnement des individus par rapport aux facteurs précédemment identifiés :

- Les individus 19, 24 et 16 représentent le mieux le profil décrit par le facteur F1. Les individus 1 et 4 le représentent le moins bien ;
- Les individus 3 et 6 représentent le mieux le profil décrit par le profil F2 ;
- Les individus 12 et 5 représentent le mieux le profil décrit par le facteur F3.

Chapitre 15. Analyse de variance : mesurer les effets de l'appartenance à une catégorie spécifique

Soit un ensemble d'individus (par exemple des élèves, des établissements, etc…) caractérisables par une ou plusieurs variables qualitatives (par exemple l'origine sociale des élèves, le statut public ou privé des établissements…) et quantitatives (par exemple les notes des élèves, les effectifs des établissements…). L'analyse de variance vise à tester si, pour une variable qualitative donnée, le fait que des individus appartiennent à des catégories différentes de cette variable exerce ou non un effet statistiquement significatif sur les caractéristiques de ces individus en termes de telle ou telle variable quantitative.

On appelle *variable indépendante* ou "*facteurs de variabilité*" (ou simplement "*facteurs*") les caractéristiques qualitatives (et leurs catégories) dont on étudie l'impact potentiel sur les caractéristiques quantitatives. Ces dernières sont les variables *dépendantes* (ou « *variables-réponses* »).

Il existe un grand nombre de techniques d'analyse de variance. La technique de base est l'analyse de variance au sens strict ou *ANOVA* (*analysis of variance*). Dans une anova, on considère une unique variable dépendante. L'anova elle-même recouvre deux grands cas de figure, l'anova à un seul facteur (section 1 de ce

chapitre) et l'anova multifactorielle (section 15.2), cette dernière comportant elle-même plusieurs possibilités : l'anova à deux facteurs, l'anova à trois facteurs, etc. Dans la pratique cependant, l'anova multifactorielle se limite à 2 ou 3 facteurs étant donné le degré de complexité des calculs au-delà.

D'autres développements de l'analyse de variance sont, notamment, l'analyse de variance multivariée (*MANOVA*), l'analyse de covariance (*ANCOVA*), et l'analyse de covariance multivariée (*MANCOVA*). L'analyse de variance multivariée (section 15.3) prend en compte deux (ou plus) variables dépendantes. Dans l'analyse de covariance[131], il n'y a qu'une unique variable dépendante, mais on peut prendre en compte des variables quantitatives parmi les variables indépendantes. L'analyse de covariance multivariée admet plusieurs variables dépendantes[132].

15.1. ANOVA À UN FACTEUR

Soit un échantillon de n individus i caractérisables par une variable indépendante qualitative X comportant ξ modalités (catégories), et une variable dépendante quantitative Y :

[131] L'analyse de covariance est présentée au chapitre 19.

[132] L'analyse de covariance multivariée est citée ici seulement pour mémoire, elle n'est pas présentée dans cet ouvrage.

Individus	Catégorie d'appartenance	Valeur de la variable dépendante
i_1	$Cat.1\ ou\ Cat.2\ ou\ \cdots ou\ Cat.\xi$	y_1
i_2	$Cat.1\ ou\ Cat.2\ ou\ \cdots ou\ Cat.\xi$	y_2
⋮	⋮	⋮
i_n	$Cat.1\ ou\ Cat.2\ ou\ \cdots ou\ Cat.\xi$	y_n

La question est de savoir si le fait pour un individu d'appartenir à telle catégorie plutôt qu'à telle autre exerce un effet sur la valeur de la variable dépendante pour cet individu. En d'autres termes, la valeur de la variable dépendante est-elle significativement différente si l'individu appartient à telle catégorie plutôt qu'à telle autre ?

ENCADRÉ 15.1 – ANOVA : LE CADRE THÉORIQUE

L'analyse de variance est fondée sur l'idée que la relation entre la variable indépendante et la variable dépendante est de forme linéaire. En termes généraux, la relation s'écrit :

$$Y = \alpha + Y(1, 2, \cdots, \boldsymbol{\xi})$$

où

- Y est la variable dépendante, quantitative ;
- α est une constante, et

- $Y(1, 2, \cdots, \xi)$ est la variable indépendante, qualitative, elle-même fonction de ses ξ catégories / modalités.

La relation indique que la valeur de la variable dépendante pour un individu est fonction des catégories de la variable indépendante.

L'analyse de variance consiste à comparer les moyennes de la variable dépendante pour les différentes catégories : plus ces moyennes sont différentes, plus le fait pour un individu d'appartenir à telle ou telle catégorie est crucial quant à la valeur de la variable dépendante pour cet individu.

On peut mesurer l'ampleur des différences entre moyennes catégorielles grâce à la dispersion de ces moyennes puisque plus les moyennes catégorielles sont différentes, plus elles s'écartent de la moyenne de l'ensemble des catégories. La dispersion des moyennes catégorielles peut être estimée à partir de la dispersion de *Y* pour l'ensemble de l'échantillon. La dispersion de *Y* pour l'ensemble des individus de l'échantillon (dispersion totale) est en effet la somme de la dispersion entre catégories et de la dispersion

entre individus au sein même de chaque catégorie. En mesurant la dispersion par la somme des carrés des écarts à la moyenne, c'est-à-dire le numérateur de la variance, la relation s'écrit :

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \left[\sum_{m=1}^{\xi} n_m(\overline{y}_m - \overline{y})^2\right] + \left[\sum_{m=1}^{\xi}\sum_{i=1}^{n}(y_{im} - \overline{y}_m)^2\right]$$

où

- $i = 1, 2, \cdots, n$ désigne les n individus composant l'échantillon ;
- m désigne les catégories 1 à ξ composant la variable indépendante : $m = 1, 2, \cdots, \boldsymbol{\xi}$;
- $\overline{y}$, la moyenne de la variable dépendante pour l'ensemble de l'échantillon ;
- $\overline{y}_m$, la moyenne de la variable dépendante pour les individus de la catégorie m ;
- n_m, le nombre d'individus dans la catégorie m ;
- y_{im}, la valeur de la variable dépendante pour l'individu i de la catégorie m.

Le premier bloc entre crochets représente la dispersion des moyennes catégorielles, qu'on appelle *variabilité interclasses* ou *variabilité intergroupes*.

Le second bloc entre crochets représente la dispersion interne aux catégories, qu'on appelle *variabilité intraclasses* ou *variabilité intra-groupes* ou *variabilité résiduelle*.

On peut considérer que l'appartenance à telle ou telle catégorie a d'autant plus d'effet sur la variable dépendante que la variabilité interclasses est supérieure à la variabilité résiduelle.

L'analyse de variance consiste finalement à vérifier si la variabilité interclasses est significativement supérieure à la variabilité résiduelle. Lorsque c'est le cas, on peut conclure que la catégorie d'appartenance exerce une influence déterminante sur la valeur de la variable dépendante.

On effectue l'analyse de variance en calculant la variabilité interclasses et la variabilité intraclasses, puis, en cas de différence, en appliquant un test F[133] de façon à déterminer si la différence est significative. La statistique du test F d'anova est le rapport :

$$F = \frac{Variabilité\ interclasses}{Variabilité\ intraclasses}$$

L'hypothèse nulle du test est que numérateur et dénominateur sont égaux. Si au contraire la variabilité interclasses est significativement supérieure à la variabilité intraclasses, on conclut que la catégorie d'appartenance importe.

Les logiciels statistiques[134] fournissent les valeurs de la variabilité interclasses et de la variabilité intraclasses, ainsi que les résultats du test F et la probabilité *p* de se tromper en rejetant l'hypothèse nulle d'égalité des variabilités.

La validité de l'anova suppose que deux conditions préalables[135] soient remplies :

---

[133] Dont la statistique diffère de celle du test F de Fisher pour l'homogénéité des variances présenté page 180, mais qui utilise la même table de Fisher-Snedecor figurant en Annexe 2.

[134] L'anova (mono- ou multifactorielle) est accessible dans XLSTAT via la commande *Modélisation des données / Anova*.

[135] D'autres conditions de validité peuvent être vérifiées *a posteriori*, à partir des résidus : voir chapitre 16.

- la première est la condition dite d'« indépendance des échantillons » : lors de la composition de l'échantillon, il doit être fait en sorte que les catégories soient indépendantes, ce qui signifie qu'un même individu (une même unité statistique) ne doit faire partie que d'une seule catégorie à la fois ;
- la seconde condition est qu'aucune catégorie ne doit comporter d'individu caractérisé par une valeur aberrante de la variable dépendante. L'anova repose sur la mesure des variances, donc la fiabilité de ses résultats peut être sensiblement affectée par la présence de valeurs aberrantes. La détection de valeurs aberrantes se conduit au sein de chaque catégorie. Elle s'effectue soit visuellement par l'examen d'un nuage de points, soit au moyen du test de Grubbs.

EXEMPLE 15.1

Afin de mesurer l'efficacité d'un enseignement, un test d'évaluation des connaissances est administré à des étudiants avant le début puis après la fin de l'enseignement. Les résultats du test initial avaient permis de distinguer trois catégories : faibles, moyens, forts. Le Tableau 15.1 montre la différence entre la note au test initial et la note au test final pour chaque étudiant. On constate une progression de l'ensemble des notes en moyenne, mais la progression moyenne par catégorie est variable, la progression de la moyenne étant d'autant plus forte que la catégorie était

initialement faible. On souhaite vérifier par une anova si la différence de progression suivant la catégorie est significative.

Tableau 15.1.
Effets de l'enseignement sur la performance aux tests d'évaluation

Étudiants	Catégorie d'appartenance (X)	Écart entre test initial et test final (Y)
Étudiant-e 1	Faible	4
Étudiant-e 2	Faible	5
Étudiant-e 3	Faible	9
Étudiant-e 4	Faible	6
Étudiant-e 5	Faible	6
Étudiant-e 6	Faible	8
Étudiant-e 7	Faible	3
Étudiant-e 8	Faible	-2
Étudiant-e 9	Faible	-3
Étudiant-e 10	Faible	11
Étudiant-e 11	Faible	8
Étudiant-e 12	Faible	7
Étudiant-e 13	Faible	3
Étudiant-e 14	Faible	10
Étudiant-e 15	Faible	4

Étudiants	Catégorie d'appartenance (X)	Écart entre test initial et test final (Y)
Étudiant-e 16	Moyen	3
Étudiant-e 17	Moyen	2
Étudiant-e 18	Moyen	2
Étudiant-e 19	Moyen	2
Étudiant-e 20	Moyen	-3
Étudiant-e 21	Moyen	4
Étudiant-e 22	Moyen	8
Étudiant-e 23	Moyen	3
Étudiant-e 24	Moyen	2
Étudiant-e 25	Moyen	6
Étudiant-e 26	Moyen	7
Étudiant-e 27	Moyen	8
Étudiant-e 28	Moyen	6
Étudiant-e 29	Moyen	8
Étudiant-e 30	Moyen	7
Étudiant-e 31	Fort	4
Étudiant-e 32	Fort	9
Étudiant-e 33	Fort	-5
Étudiant-e 34	Fort	10
Étudiant-e 35	Fort	-4

Étudiants	Catégorie d'appartenance (X)	Écart entre test initial et test final (Y)
Étudiant-e 36	Fort	3
Étudiant-e 37	Fort	-5
Étudiant-e 38	Fort	-2
Étudiant-e 39	Fort	-3
Étudiant-e 40	Fort	9
Étudiant-e 41	Fort	-4
Étudiant-e 42	Fort	5
Étudiant-e 43	Fort	3
Étudiant-e 44	Fort	6
Étudiant-e 45	Fort	-3
Progression moyenne de la catégorie Faible		*5,26*
Progression moyenne de la catégorie Moyen		*4,33*
Progression moyenne de la catégorie Fort		*1,53*
Progression moyenne de l'ensemble des 45 étudiants		*3,71*

Vérification des conditions de validité

La condition d'indépendance des catégories est remplie puisqu'un même individu ne peut appartenir qu'à une seule des trois catégories.

Avant d'appliquer le test de Grubbs, on vérifie la normalité de la variable dépendante pour chaque catégorie de la variable indépendante :

	Shapiro -Wilk	Anderson -Darling	Lilliefors	Jarque-Bera
Faible	0,36	0,42	0,47	0,56
Moyen	0,07	0,10	0,31	0,63
Fort	0,05	0,07	0,09	0,48

La variable dépendante suit donc une distribution gaussienne dans chacune des catégories de la variable indépendante. Le test de Grubbs, ensuite, indique :

	Fort	Moyen	Faible
Valeur observée de la statistique de Grubbs	2,07	2,32	1,54
Valeur critique	2,55	2,55	2,55
p-value (bilatérale)	0,38	0,15	0,28

Aucune catégorie de la variable indépendante ne comporte donc de valeur aberrante au seuil de significativité de 5%.

Analyse des résultats

Les résultats de l'analyse de variance s'établissent comme suit :

Source	DDL	Somme des carrés	Moyenne des carrés	F	p
Interclasses	2	113,24	56,62	3,018	0,06
Intraclasses	42	788,00	18,76		
Total	44	901,24			

Le nombre de DDL pour la variabilité interclasses est égal a $nombre\ de\ catégories - 1$. Pour la variabilité intraclasses, $nombre\ de\ DDL = \sum_{m=1}^{\xi}(n_m - 1)$ où n_m est l'effectif de la catégorie *m*, soit ici $(15 - 1) + (15 - 1) + (15 - 1) = 42$. La variabilité interclasses est inférieure à la variabilité résiduelle. La moyenne des carrés est obtenue en divisant la somme des carrés par le nombre de DDL. La valeur calculée de la statistique F s'élève à 3,018 contre une valeur critique de 19,47 pour les DDL considérés (la plus forte variance au

numérateur) au seuil de significativité de 5%. La probabilité p de se tromper en rejetant l'hypothèse nulle d'égalité des variances s'établit à 0,06 et est donc trop élevée pour que l'on puisse rejeter l'hypothèse nulle. On conclut donc que la différence de progression entre catégories n'est pas statistiquement significative.

EXEMPLE 15.2

On dispose de données sur la répartition des effectifs d'enseignants dans le premier degré public par Académie et Département (Tableau 15.2). On cherche à savoir si l'effectif d'enseignants par Département diffère significativement d'une Académie à l'autre.

Tableau 15.2.
Effectifs d'enseignants dans le premier degré public par Académie et Département pour vingt Académies métropolitaines – 2013-2014

Source : MENESR (2014), Repères et références statistiques sur les enseignements, la formation et la recherche, page 299
http://www.enseignementsup-recherche.gouv.fr/cid81900/reperes-et-references-statistiques-edition-septembre-2014.html

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Alpes-Haute-Provence	Aix-Marseille	917
Hautes-Alpes	Aix-Marseille	770
Bouches-du-Rhône	Aix-Marseille	9 618
Vaucluse	Aix-Marseille	2 839
Aisne	Amiens	2 918
Oise	Amiens	4 648
Somme	Amiens	2 817
Doubs	Besançon	2 932
Jura	Besançon	1 316
Haute-Saône	Besançon	1 321
Territoire de Belfort	Besançon	734
Dordogne	Bordeaux	1 776
Gironde	Bordeaux	6 930
Landes	Bordeaux	1 818
Lot-et-Garonne	Bordeaux	1 514
Pyrénées-Atlantiques	Bordeaux	2 631
Calvados	Caen	3 308
Manche	Caen	2 175
Orne	Caen	1 276
Allier	Clermont-Ferrand	1 624
Cantal	Clermont-Ferrand	761
Haute-Loire	Clermont-Ferrand	959

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Puy-de-Dôme	Clermont-Ferrand	2 864
Corse-du-Sud	Corse	656
Haute-Corse	Corse	782
Seine-et-Marne	Créteil	8 009
Seine-Saint-Denis	Créteil	9 838
Val-de-Marne	Créteil	6 933
Côte-d'Or	Dijon	2 624
Nièvre	Dijon	1 033
Saône-et-Loire	Dijon	2 742
Yonne	Dijon	1 815
Ardèche	Grenoble	1 379
Drôme	Grenoble	2 534
Isère	Grenoble	6 425
Savoie	Grenoble	2 086
Haute-Savoie	Grenoble	3 795
Nord	Lille	12 946
Pas-de-Calais	Lille	7 926
Ain	Lyon	3 272
Loire	Lyon	3 383
Rhône	Lyon	8 436
Aude	Montpellier	1 703
Gard	Montpellier	3 538
Hérault	Montpellier	4 999
Lozère	Montpellier	426

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Pyrénées-Orientales	Montpellier	2 272
Meurthe-et-Moselle	Nancy-Metz	3 640
Meuse	Nancy-Metz	1 064
Moselle	Nancy-Metz	5 381
Vosges	Nancy-Metz	2 029
Loire-Atlantique	Nantes	5 087
Maine-et-Loire	Nantes	2 962
Mayenne	Nantes	1 230
Sarthe	Nantes	2 723
Vendée	Nantes	1 815
Cher	Orléans-Tours	1 460
Eure-et-Loir	Orléans-Tours	2 384
Indre	Orléans-Tours	1 035
Indre-et-Loire	Orléans-Tours	2 722
Loir-et-Cher	Orléans-Tours	1 616
Loiret	Orléans-Tours	3 419
Charente	Poitiers	1 590
Charente-Maritime	Poitiers	2 832
Deux-Sèvres	Poitiers	1 664
Vienne	Poitiers	2 017
Ardennes	Reims	1 641
Aube	Reims	1 614
Marne	Reims	2 875
Haute-Marne	Reims	1 056
Côtes-d'Armor	Rennes	2 267

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Finistère	Rennes	3 092
Ille-et-Vilaine	Rennes	3 743
Morbihan	Rennes	2 134
Ariège	Toulouse	735
Aveyron	Toulouse	1 137
Haute-Garonne	Toulouse	5 999
Gers	Toulouse	874
Lot	Toulouse	775
Hautes-Pyrénées	Toulouse	1 044
Tarn	Toulouse	1 653
Tarn-et-Garonne	Toulouse	1 304

Vérification des conditions de validité

La condition d'indépendance des catégories est manifestement remplie. La condition de normalité peut être vérifiée pour les Académies couvrant au moins quatre Départements :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Aix-Marseille	0,08	0,09	0,17	0,68
Besançon	0,23	0,14	0,08	0,74
Bordeaux	0,01	0,01	0,04	0,45
Clermont-Ferrand	0,41	0,35	0,61	0,77
Dijon	0,45	0,40	0,42	0,79

Grenoble	0,41	0,36	0,45	0,70
Montpellier	0,97	0,90	0,90	0,86
Nancy-Metz	0,85	0,75	0,82	0,83
Nantes	0,58	0,49	0,41	0,78
Orléans-Tours	0,81	0,76	0,57	0,79
Poitiers	0,25	0,22	0,47	0,74
Reims	0,38	0,23	0,13	0,79
Rennes	0,47	0,42	0,42	0,79
Toulouse	<0,0001	0,00	0,00	0,01

La condition de normalité est remplie dans toutes les Académies couvrant au moins quatre Départements, à l'exception de Bordeaux et Toulouse. Le test de Grubbs est appliqué aux académies dans lesquelles la condition de normalité est remplie :

	G	Valeur critique	p-value
Aix-Marseille	1,46	1,48	0,10
Besançon	1,43	1,48	0,17
Clermont-Ferrand	1,38	1,48	0,31
Dijon	1,28	1,48	0,58
Grenoble	1,60	1,48	0,20
Montpellier	1,38	1,48	0,64
Nancy-Metz	1,24	1,48	0,69
Nantes	1,58	1,48	0,24
Orléans-Tours	1,47	1,48	0,64
Poitiers	1,42	1,48	0,22
Reims	1,40	1,48	0,25
Rennes	1,24	1,48	0,69

Donc aucune valeur aberrante n'apparait dans aucune des 12 (sur 21) Académies dans lesquelles il a pu être vérifié que la condition de normalité est remplie. L'examen visuel des effectifs d'enseignants dans les neuf autres Académies ne suggère aucune valeur exceptionnelle manifestement aberrante.

Analyse des résultats

Les résultats de l'analyse de variance s'établissent comme suit :

	DDL	Somme des carrés	Moyenne des carrés	F	p
Interclasses	19	268548641,58	14134139,031	4,46	< 0,0001
Intraclasses	62	196208172,04	3164647,936		
Total corrigé	81	464756813,62			

La variabilité interclasses est supérieure à la variabilité résiduelle. La valeur calculée de la statistique F s'élève à 4,46 contre une valeur critique de 1,76 pour les DDL

considérés au seuil de significativité de 5%. La probabilité p est inférieure à 0,0001. Donc l'hypothèse nulle d'égalité des variances peut être rejetée, et on peut conclure que l'Académie d'appartenance exerce un effet significatif sur le nombre d'enseignants par Département.

On peut alors chercher à en savoir plus sur les Académies qui ont entre elles des différences significatives en matière d'effectifs d'enseignement du premier degré public. On effectue à cette fin un test de comparaison par paires. Le test consiste à mesurer la significativité des différences deux-à-deux. Les logiciels proposent plusieurs de ces tests. Par exemple, le test de Tukey indique les résultats suivants :

Contraste	Différence standardisée	Valeur critique	Pr > Diff	Significatif
Corse vs Lille	-5462,00	3701,00	0,000	Oui
Corse vs Créteil	-4644,00	3701,00	0,003	Oui
Clermont-Ferrand vs Lille	-5767,00	3701,00	< 0,0001	Oui
Clermont-Ferrand vs Créteil	-4937,00	3701,00	0,001	Oui

Contraste	Différence standardisée	Valeur critique	Pr > Diff	Significatif
Besançon vs Lille	-5751,00	3701,00	< 0,0001	Oui
Besançon vs Créteil	-4920,00	3701,00	0,001	Oui
⋮	⋮	⋮	⋮	⋮
Grenoble vs Lille	-4832,00	3701,00	0,001	Oui
Grenoble vs Créteil	-3861,00	3701,00	0,032	Oui
Amiens vs Lille	-4295,00	3701,00	0,009	Oui
Aix-Marseille vs Lille	-4479,00	3701,00	0,005	Oui
Corse vs Lyon	-2655,00	3701,00	0,478	Non
Corse vs Aix-Marseille	-1828,00	3701,00	0,947	Non

Contraste	Différence standardisée	Valeur critique	Pr > Diff	Significatif
⋮	⋮	⋮	⋮	⋮
Amiens vs Créteil	-3304,00	3701,00	0,137	Non
Amiens vs Lyon	-1080,00	3701,00	1,000	Non
Amiens vs Aix-Marseille	-0,055	3701,00	1,000	Non
Aix-Marseille vs Créteil	-3477,00	3701,00	0,090	Non
Aix-Marseille vs Lyon	-1100,00	3701,00	1,000	Non
Lyon vs Lille	-3329,00	3701,00	0,130	Non
Lyon vs Créteil	-2224,00	3701,00	0,775	Non
Créteil vs Lille	-1340,00	3701,00	0,998	Non

L'information importante est celle fournie dans la dernière colonne (significativité des différences deux à deux). On constate qu'il n'y a pas de différence significative entre Créteil et Lille, mais qu'en revanche ces deux académies sont significativement différentes de toutes les autres, à l'exception de Lyon (pas de différence entre Lyon et Lille ni entre Lyon et Créteil) ; et à l'exception d'Amiens et Aix-Marseille par rapport à Créteil (pas de différence entre Amiens et Créteil ni entre Aix-Marseille et Créteil). Le tableau de synthèse des comparaisons multiples résume ce résultat :

Synthèse des comparaisons multiples par paires			
Modalités	Classes		
Corse	A		
Clermont-Ferrand	A		
Besançon	A		
Toulouse	A		
Reims	A		
Poitiers	A		
Dijon	A		
Orléans-Tours	A		
Caen	A		
Montpellier	A		

Synthèse des comparaisons multiples par paires			
Modalités	Classes		
Nantes	A		
Rennes	A		
Bordeaux	A		
Nancy-Metz	A		
Grenoble	A		
Amiens	A	B	
Aix-Marseille	A	B	
Lyon	A	B	C
Créteil		B	C
Lille			C

15.2. ANOVA MULTIFACTORIELLE

Dans l'anova multifactorielle, les variables indépendantes sont au moins au nombre de deux[136]

[136] La relation entre variables indépendantes et variable dépendante s'écrit :

$$Y = \alpha + Y(X_1, X_2, \ldots, X_\omega)$$

où

- *Y* est la variable quantitative dépendante ;

(chacune ayant au moins deux modalités). La conséquence en est que la variable dépendante peut être affectée non seulement par chaque facteur, mais aussi par les combinaisons de facteurs, qu'on appelle aussi *interactions*. Les résultats à interpréter peuvent donc porter (si le chercheur s'y intéresse) non seulement sur l'effet potentiel de l'appartenance à telle ou telle catégorie de chaque facteur, mais également sur l'effet potentiel de l'appartenance à telle ou telle combinaison de catégories (ou « groupe »).

L'utilisation de l'anova multifactorielle suppose que soit remplie la condition d'indépendance des groupes (un même individu ne peut appartenir qu'à une unique combinaison de modalités). Aucun groupe ne doit non plus contenir de valeur aberrante. En outre, il ne doit pas y avoir *multicolinéarité des facteurs*. En règle générale, la condition de non-multicolinéarité concerne les variables *indépendantes* quantitatives ou qualitatives ordinales. Dans une anova multifactorielle, l'éventuelle multicolinéarité des facteurs ne peut donc provenir que de variables qualitatives ordinales. La multicolinéarité des facteurs signifie que les variables

---

- α la constante ;
- $X_1, X_2, \ldots, X_\omega$ sont les ω facteurs, variables qualitatives indépendantes.

indépendantes ne doivent pas être liées par une relation de dépendance linéaire entre elles. En d'autres termes, elles ne doivent pas être excessivement intercorrélées. La présence de multicolinéarité indique que les variables sont redondantes. Elle fausse les calculs de significativité[137]. Elle se mesure par le « facteur d'inflation de variance » (*variance inflation factor* – VIF)[138] ou par la « *tolérance* », qui en est l'inverse[139] : il n'y pas de multicolinéarité si le VIF est égal à 1, il y en a à partir de 5 ; au-delà de 10 (c'est-à-dire si la tolérance n'est pas au moins égale à 0,10), elle devient préoccupante car la fiabilité des calculs de significativité devient douteuse. Il faut alors retirer au moins l'une des variables redondantes[140].

EXEMPLE 15.3

On dispose de données sur le taux de réussite aux concours de recrutement d'enseignants du second degré

---

[137] La multicolinéarité augmente l'erreur standard, ce qui réduit la valeur du rapport entre coefficient de régression et erreur standard qui détermine la statistique de significativité, et augmente la p-value.

[138] C'est-à-dire techniquement, le facteur par lequel est multiplié l'erreur standard, par rapport à une situation sans intercorrélation.

[139] $Tolérance = 1/VIF$

[140] XLSTAT permet de mesurer la multicolinéarité à partir de la commande *Statistiques de multicolinéarité* sous l'onglet *Description des données*.

pour 40 centres de préparation aux concours. Les données disponibles couvrent 4 disciplines et 5 sites de formation (Tableau 15.3). On cherche à savoir si le taux de réussite diffère significativement suivant la discipline et/ou le site de formation.

Tableau 15.3.
Taux de réussite aux concours de recrutement d'enseignants du second degré pour 40 centres de préparation couvrant 4 disciplines (D1 à D4) et 5 sites (S1 à S5)

Centre	Discipline	Site	Taux de réussite
1D1S1	D1	S1	0,41
2D1S1	D1	S1	0,57
3D1S2	D1	S2	0,88
4D1S4	D1	S4	0,16
5D1S3	D1	S3	0,09
6D1S4	D1	S4	0,17
7D1S1	D1	S1	0,43
8D1S1	D1	S1	0,57
9D1S2	D1	S2	0,97
10D1S4	D1	S4	0,18
11D1S3	D1	S3	0,16
12D1S4	D1	S4	0,19
13D1S5	D1	S5	0,44
14D2S1	D2	S1	0,44

Centre	Discipline	Site	Taux de réussite
15D2S2	D2	S2	0,73
16D2S2	D2	S2	0,98
17D2S5	D2	S5	0,46
18D2S4	D2	S4	0,08
19D2S2	D2	S2	0,99
20D2S5	D2	S5	0,51
21D2S5	D2	S5	0,54
22D3S1	D3	S1	0,46
23D3S2	D3	S2	0,74
24D3S3	D3	S3	0,01
25D3S4	D3	S4	0,09
26D3S3	D3	S3	0,02
27D3S1	D3	S1	0,51
28D3S2	D3	S2	0,75
29D3S3	D3	S3	0,06
30D3S4	D3	S4	0,09
31D3S3	D3	S3	0,08
32D3S5	D3	S5	0,57
33D3S5	D3	S5	0,57
34D4S1	D4	S1	0,54
35D4S2	D4	S2	0,79
36D4S3	D4	S3	0,08
37D4S5	D4	S5	0,59
38D4S4	D4	S4	0,16

Centre	Discipline	Site	Taux de réussite
39D4S3	D4	S3	0,09
40D4S5	D4	S5	0,60

Vérification des conditions de validité

La condition d'indépendance des groupes est manifestement remplie. On peut vérifier la normalité de la variable dépendante pour les trois groupes (D1S1, D1S4, D3S3) disposant d'au moins quatre observations :

	Shapiro -Wilk	Anderson -Darling	Lilliefors	Jarque-Bera
D1S1	0,07	0,09	0,21	0,72
D1S4	0,97	0,86	0,99	0,86
D3S3	0,51	0,45	0,49	0,78

La condition de normalité est remplie pour ces trois groupes. Le test de Grubbs montre qu'aucun de ces groupes ne comporte de valeur aberrante :

	G	Valeur critique	p-value
D1S1	0,98	1,48	0,61
D1S4	1,16	1,48	0,90
D3S3	1,13	1,48	0,97

L'examen visuel des taux de réussite dans les autres groupes ne suggère dans aucun groupe la présence d'une valeur exceptionnellement différente des autres valeurs du groupe.

Enfin, la question de la multicolinéarité ne se pose pas puisque tous les facteurs sont qualitatifs nominaux[141].

Analyse des résultats

Les résultats de l'anova sont présentés dans la table d'anova ci-après. La variabilité interclasses est supérieure à la variabilité résiduelle. La probabilité p de se tromper en rejetant l'hypothèse nulle d'égalité des variances est inférieure à 0,0001. Donc au seuil de significativité de 5%, on peut considérer que la discipline d'appartenance et/ou le site d'implantation de la formation ont un effet significatif sur le taux de réussite au concours. Il importe donc d'identifier exactement où sont les différences qui comptent.

[141] Le logiciel peut fournir des statistiques de multicolinéarité car lors de la saisie des données, les facteurs qualitatifs ne sont pas différenciés comme nominaux ou ordinaux. Il revient donc à l'utilisateur de ne pas tenir compte des statistiques de multicolinéarité fournies lorsqu'elles n'ont pas de sens étant donné le caractère nominal de tous les facteurs.

Table d'anova

	DDL	Somme des carrés	Moyenne des carrés	F	p
Interclasses	18*	3,33	0,185	47,97	<0,0001
Intraclasses	21*	0,08	0,004		
Total corrigé	39*	3,41			

* Le nombre de DDL est calculé ici en tenant compte non seulement du nombre de catégories et de l'effectif des catégories, mais aussi du nombre d'interactions entre modalités des catégories

Tests de comparaison

Le test de comparaison de Tukey effectué par discipline, par site, et par combinaison discipline-site permet de repérer les points de différence.

- Comparaison entre disciplines

Contraste	Différence standardisée	Valeur critique	Pr > Diff	Significatif
D3 vs D2	-9,318	2,787	<0,0001	Oui
D3 vs D4	-2,694	2,787	0,061	Non
D3 vs D1	-2,942	2,787	0,036	Oui
D1 vs D2	-6,843	2,787	<0,0001	Oui
D1 vs D4	-0,221	2,787	0,996	Non
D4 vs D2	-5,742	2,787	<0,0001	Oui

Toutes les paires montrent une différence significative, sauf deux, à savoir D3-D4 et D1-D4. Pour autant, D1, D3 et D4 n'appartiennent pas à la même classe puisque D3 et D1 sont significativement différentes. Donc la classe se limite à D1-D4. En définitive, on observe trois classes de disciplines entre lesquelles les différences comptent :

- D1-D4 ;
- D2, distincte des deux autres classes ;
- D3, distincte de D2 et en partie de D1-D4 (D3 différente de D1 mais pas de D4).

Le tableau de synthèse des comparaisons multiples résume ce résultat :

	Moyenne	Erreur standard	Borne inférieure (95%)	Borne supérieure (95%)	Classes		
D3	0,328	0,018	0,291	0,365	A		
D1	0,401	0,017	0,365	0,437		B	
D4	0,407	0,023	0,359	0,456		B	
D2	0,591	0,022	0,546	0,637			C

La classe A réunit tous les centres de préparation relevant de la discipline D3. Le taux de réussite de ces centres s'établit à 32,8% en moyenne. La classe B réunit les centres relevant des disciplines D1 et D4, avec un taux de réussite moyen de 40%. La classe C réunit les centres relevant de la discipline D2, dont le taux de réussite au concours est en moyenne de 59,1%. Les moyennes sont statistiquement significatives car elles se situent entre les bornes des intervalles de confiance avec un degré de confiance de 95%.

- Comparaison entre sites de formation

Contraste	Différence standardisée	Valeur critique	Pr > Diff	Significatif
S3 vs S2	-25,168	2,979	< 0,0001	Oui
S3 vs S5	-14,897	2,979	< 0,0001	Oui
S3 vs S1	-13,462	2,979	< 0,0001	Oui
S3 vs S4	-2,145	2,979	0,239	Non
S4 vs S2	-23,023	2,979	< 0,0001	Oui
S4 vs S5	-12,752	2,979	< 0,0001	Oui

Contraste	Différence standardisée	Valeur critique	Pr > Diff	Significatif
S4 vs S1	-11,317	2,979	< 0,0001	Oui
S1 vs S2	-11,706	2,979	< 0,0001	Oui
S1 vs S5	-1,435	2,979	0,613	Non
S5 vs S2	-10,271	2,979	< 0,0001	Oui

Toutes les paires montrent une différence significative, sauf S3-S4 et S1-S5. Il y a donc trois différentes classes de sites :
- S1-S5 ;
- S2 ;
- S3-S4.

Le tableau de synthèse des comparaisons multiples résume ce résultat :

	Moyenne	Erreur standard	Borne inférieure (95%)	Borne supérieure (95%)	Classes
S3	0,074	0,022	0,028	0,119	A
S4	0,140	0,022	0,095	0,186	A
S1	0,490	0,022	0,445	0,536	B
S5	0,535	0,022	0,489	0,580	B
S2	0,852	0,022	0,807	0,898	C

La classe A réunit les sites S3 et S4, dont le taux de réussite au concours est en moyenne compris entre 7 et 14%. La classe B réunit les sites S1 et S5, dont le taux

de réussite au concours est en moyenne compris entre 49 et 53%. La classe C réunit les centres de préparation implantés sur le site S2, qui ont un taux de réussite moyen de 85,2%.

- *Comparaison des combinaisons discipline-site*

Le tableau de synthèse des comparaisons multiples des combinaisons discipline-site montre :

Combinaison		Moyenne	Erreur standard	Borne inf. (95%)	Borne sup. (95%)				
1	D3*S3	0,043	0,031	-0,021	0,108	A			
2	D2*S4	0,084	0,062	-0,044	0,213	A			
3	D4*S3	0,086	0,044	-0,005	0,177	A			
4	D3*S4	0,088	0,044	-0,003	0,179	A			
5	D1*S3	0,123	0,044	0,032	0,214	A			
6	D4*S4	0,157	0,062	0,029	0,286	A	B		
7	D1*S4	0,176	0,031	0,112	0,240	A	B		
8	D2*S1	0,440	0,062	0,311	0,568		B	C	
9	D1*S5	0,440	0,062	0,311	0,568		B	C	
10	D3*S1	0,485	0,044	0,394	0,576			C	
11	D1*S1	0,495	0,031	0,430	0,559			C	
12	D2*S5	0,502	0,036	0,427	0,576			C	
13	D4*S1	0,535	0,062	0,407	0,664			C	D
14	D3*S5	0,569	0,044	0,478	0,660			C	D
15	D4*S5	0,598	0,044	0,507	0,689			C	D

Combinaison		Moyenne	Erreur standard	Borne inf. (95%)	Borne sup. (95%)			
16	D3*S2	0,741	0,044	0,650	0,832	C	D	E
17	D4*S2	0,790	0,062	0,662	0,919		D	E
18	D2*S2	0,901	0,036	0,826	0,975			E
19	D1*S2	0,923	0,044	0,832	1,014			E

Émergent finalement trois différentes classes ;

- Classe A, comprenant les combinaisons 1 à 7, dont le taux de réussite moyen au concours est compris entre 4,3% et 17,6% ;
- Classe C/CD, comprenant les combinaisons 8 à 15, dont le taux de réussite moyen est compris entre 44 et 59,8% ;
- Classe E, comprenant les combinaisons 16 à 19, dont le taux de réussite moyen est compris entre 74 et 92,3%.

En définitive, les différences entre disciplines et sites qui influencent significativement le taux de réussite apparaissent clairement :

- indépendamment du site, les taux de réussite sont plus élevés dans la discipline D2 que dans les autres, et plus faibles dans la discipline D3 ;

- indépendamment de la discipline, les taux de réussite sont plus élevés sur le site S2, et plus faibles sur les sites S3-S4 ;
- les différences entre combinaisons sont tout aussi claires. Elles permettent de conclure par exemple que pour un candidat aux concours, les stratégies les plus efficaces consistent à se préparer dans les disciplines D1 ou D2 sur le site S2 ; les moins efficaces à se préparer dans la discipline D3 sur le site S3, et plus généralement sur les sites S3 ou S4, quelle que soit la discipline.

15.3. MANOVA : L'ANOVA MULTIVARIÉE

L'anova multivariée (manova)[142] permet d'analyser l'effet d'une ou plusieurs variables qualitatives sur au moins deux variables quantitatives.

Soit un échantillon d'individus appartenant à différentes catégories d'une ou plusieurs variables qualitatives, et caractérisés en termes de deux ou plus variables quantitatives. Considérons par exemple des élèves d'un même établissement appartenant à différents groupes d'âge et d'origines socio-économique (catégories socioprofessionnelles des

[142] Accessible dans XLSTAT via la commande *Modélisation des données / Manova*.

chefs de famille) différentes, et pour lesquels on dispose de données sur leurs compétences en langues, sciences et arts.

L'anova multivariée (manova) permet d'identifier l'existence de relations significatives entre l'appartenance catégorielle et tel ou tel profil en termes de variables quantitatives. Dans l'exemple ci-dessus, la manova peut permettre d'identifier un lien significatif entre d'une part appartenance à tel groupe d'âge de telle origine socio-économique et d'autre part, tel ou tel profil de compétences en langues, sciences et arts.

L'approche et la méthodologie de la manova diffèrent suivant que la variable qualitative indépendante est unique ou non.

15.3.1. Manova monofactorielle

Dans ce cas de figure, il y au moins deux variables dépendantes, et toutes deux doivent être quantitatives ; il y a une seule variable indépendante, elle est qualitative et comporte au moins deux modalités / catégories.

L'objectif de l'analyse est de savoir si l'appartenance catégorielle exerce un effet significatif sur le profil en termes de variables-réponses.

L'analyse consiste à tester une hypothèse alternative postulant l'existence d'un effet significatif contre une hypothèse nulle postulant l'absence d'effet. Différents tests existent : tests de Wilks, Hotelling-Lawley, Pillai, Roy, notamment. L'hypothèse nulle est rejetée si la probabilité *p* est inférieure au seuil de significativité.

Lorsqu'un effet significatif est mis en évidence, la question reste posée de savoir quelle(s) catégorie exactement se caractérise par quel profil. Il faut pour le savoir procéder à une analyse complémentaire, par exemple appliquer un test de Tukey de comparaisons multiples.

La validité d'une manova nécessite que trois principales conditions préalables soient remplies :

- il ne doit y avoir de valeur aberrante dans aucune des catégories de la variable indépendante, car la présence de valeurs aberrantes fausse les résultats de la manova. On distingue valeurs aberrantes « univariées » et « multivariées ». On est dans le cas univarié lorsque la recherche de valeur(s) aberrante(s) s'effectue au sein d'une même série. Les individus d'une catégorie de la variable indépendante qui présentent une valeur aberrante dans au moins une des séries quantitatives dépendantes (« valeur aberrante

univariée ») doivent être retirés de l'échantillon[143]. Doivent être retirés également les individus d'une catégorie qui présentent des « valeurs aberrantes multivariées », c'est-à-dire des profils marginaux en termes de combinaison de variables dépendantes. Un grand nombre de méthodes permettent de détecter les valeurs aberrantes multivariées[144], mais elles ne sont pas systématiquement disponibles sur tous les logiciels statistiques. La méthode la plus utilisée est la *Distance de Mahalanobis* : pour un ensemble de variables $(X_1, X_2, , \cdots, X_\xi)$, la distance de Mahalanobis mesure l'éloignement entre le vecteur $(x_{1i}, x_{2i}, \dots, x_{\xi i})$ de l'individu *i* et le vecteur moyen $(\overline{x_1}, \overline{x_2}, \dots, \overline{x_\xi})$ de l'échantillon[145]. S'il n'y a pas plus de deux ou trois

---

[143] Rappelons que les valeurs aberrantes univariées se détectent visuellement ou alors au moyen du test de Grubbs si la série est normalement distribuée.

[144] Voir par exemple en liste des références la revue effectuée par Planchon (2005).

[145] Le « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9) dispose d'une fonction MOUTLIERS qui permet de calculer la distance de Mahalanobis. La fonction MOUTLIERS est une fonction matricielle. Supposons par exemple qu'on recherche des valeurs aberrantes multivariées pour trois séries quantitatives dépendantes Y_1, Y_2, Y_3 dans une catégorie de la variable indépendante. Il y a 15 observations dans la catégorie. On inscrit la première série, Y_1, dans les cellules A1 à A15 de la feuille Excel, puis Y_2 dans la plage B1:B15 et Y_3 dans la plage

séries quantitatives à inspecter simultanément, on peut aussi détecter visuellement les valeurs aberrantes multivariées par examen du nuage de points en deux ou trois dimensions ;

- il doit y avoir *multinormalité* (ou « normalité multivariée ») des variables dépendantes, c'est-à-dire normalité des variables dépendantes prises toutes ensemble pour la totalité de l'échantillon. La multinormalité est l'équivalent de la normalité que vérifient les tests usuels de normalité (Shapiro-Wilk, Anderson-Darling, …) dans les analyses univariées. La multinormalité se vérifie au moyen de tests spécifiques tels que les tests de Mardia, le test de Henze-Zirkler, ou le test de Rao-Ali, mais ces tests ne sont pas systématiquement disponibles sur tous les logiciels statistiques[146] ;

C1:C15. Puis on sélectionne la plage D1:E15 et on inscrit dans la fenêtre la formule

=MOUTLIERS(A1:C15)

Enfin on appuie en même temps sur les touches Ctrl Maj Entrée. Deux colonnes s'affichent alors. La première indique la distance de Mahalanobis pour chacun des 15 vecteurs y_{1i}, y_{2i}, y_{3i}. La deuxième colonne indique la p-value associée à chaque vecteur. On considère comme valeurs aberrantes les vecteurs pour lesquels la p-value est inférieure à 0,001 (seuil de significativité recommandé par Tabachnick et Fidell, 2007, p.74).

[146] Cependant, le « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9) dispose d'une fonction MKURTTEST

- il doit y avoir homogénéité des matrices de variances-covariances, ce qui est l'équivalent pour le cas multivarié de l'homogénéité des variances que vérifient les tests usuels d'homogénéité des variances (Bartlett, Levene, F de Fisher) en analyse univariée. L'homogénéité des matrices de variances-covariances se vérifie au moyen du test M de Box147. L'hypothèse

---

qui propose le test du coefficient d'aplatissement (*kurtosis*) de Mardia. Supposons par exemple qu'on veuille tester la multinormalité de trois séries quantitatives dépendantes Y_1, Y_2, Y_3. Il y a 45 observations dans l'échantillon. On inscrit la première série, Y_1, dans la plage A1:A45 de la feuille Excel, puis Y_2 dans la plage B1:B45 et Y_3 dans la plage C1:C45. Puis on sélectionne la plage D1:E3 et on inscrit dans la fenêtre la formule

=MKURTTEST(A1:C45, VRAI)

Enfin on appuie en même temps sur les touches Ctrl Maj Entrée. Les trois cellules D1 à E3 se remplissent alors. E1 indique le coefficient d'aplatissement, et E2 la statistique z associée. L'information essentielle figure en E3 : c'est la p-value du test d'aplatissement de Mardia. Il y a multinormalité si la p-value est supérieure ou égale à 0,05. Au contraire, la condition de multinormalité n'est pas remplie si la p-value est inférieure à 0,05.

147 Le « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9) dispose d'une fonction BOX qui propose le test M de Box (à condition que les valeurs des séries ne soient pas des nombres compris entre 0 et 1). Supposons par exemple qu'on veuille vérifier l'homogénéité des matrices de variances-covariances pour un échantillon caractérisé au regard d'une variable qualitative indépendante X et trois variables

nulle du test postule l'homogénéité. L'hypothèse nulle peut être rejetée si p<0,001 : on peut considérer qu'il y a homogénéité tant que $p \geq 0{,}001$[148].

EXEMPLE 15.4

On dispose de données sur les résultats à trois épreuves (orthographe, éducation civique, géographie) pour des élèves provenant de trois établissements. On cherche à déceler un éventuel lien entre établissement d'origine et profil des résultats aux différentes épreuves.

---

quantitatives dépendantes Y_1, Y_2, Y_3. Il y a 40 observations dans l'échantillon. On inscrit les valeurs de X dans la plage A1:A40 de la feuille Excel, celles de Y_1 dans la plage B1:B40, celles de Y_2 dans la plage C1:C40, et celles de Y_3 dans la plage D1:D40. Puis on sélectionne la plage E1:E5 et on inscrit dans la fenêtre la formule

=BOX(A1:D40)

Enfin on appuie en même temps sur les touches Ctrl Maj Entrée. Les cinq cellules E1 à E5 se remplissent alors. E1 indique la valeur calculée de la statistique de Box ; E2 et E3 les degrés de liberté ; et E4 la statistique F. L'information essentielle figure en E5 : c'est la p-value du test de Box. Il y a homogénéité des matrices de variances-covariances si la p-value est supérieure ou égale à 0,001. Au contraire, la condition d'homogénéité n'est pas remplie si la p-value est inférieure à 0,001.

[148] Le test est en effet très sensible à l'absence de multinormalité et à la taille de l'échantillon, d'où la nécessité d'un seuil de significativité très bas pour s'assurer de ne rejeter l'hypothèse nulle qu'à bon escient : voir par exemple Hahs-Vaughn (2016).

Tableau 15.4.
Résultats d'élèves de trois établissements à trois épreuves :
- Orthographe (O)
- Éducation civique (EC)
- Géographie (Géo)

Identifiant de l'élève	Établissement d'origine	Résultats		
		O	EC	Géo
1	1	20	11	13
2	3	12	8	8
3	1	16	10	20
4	1	17	0	13
5	2	13	17	12
6	2	11	20	18
7	1	15	9	8
8	2	20	20	9
9	3	7	12	17
10	3	9	9	17
11	1	17	11	13
12	2	10	17	6
13	2	11	20	8
14	3	5	13	7
15	2	17	18	7
16	1	19	15	20
17	1	20	4	11
18	2	12	16	20
19	3	8	8	9
20	1	15	6	7

Identifiant de l'élève	Établissement d'origine	Résultats		
		O	EC	Géo
21	2	8	16	9
22	3	8	12	3
23	3	10	12	12
24	3	13	6	17
25	2	20	16	7
26	3	10	12	16
27	1	14	6	16
28	1	20	12	20
29	3	4	14	11
30	3	11	12	8
31	2	12	20	2
32	1	15	4	17
33	2	11	17	9

Vérification des conditions de validité

- On vérifie tout d'abord si aucun des trois établissements ne présente de valeurs aberrantes univariées ou multivariées :
  - Valeurs aberrantes univariées

Les trois séries de notes sont normalement distribuées :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
O	0,18	0,29	0,48	0,58
EC	0,33	0,47	0,54	0,63
Géo	0,06	0,05	0,03	0,46

Le test G de Grubbs signale la présence de valeurs aberrantes :

		Orthographe		Éducation civique		Géographie	
		G*	p	G*	p	G*	p
Établissements	1	1,34	0,13	1,82	0,53	1,58	0,92
	2	1,69	0,78	1,18	<0,0001	1,96	0,32
	3	1,72	0,71	1,86	0,46	1,72	0,72

* Valeur critique : 2,35

Le test signale la note 20 de l'élève 6 comme valeur aberrante pour l'Éducation civique dans l'établissement 2. Cette note est donc retirée de l'échantillon pour la suite des analyses.

- Valeurs aberrantes multivariées

Les distances de Mahalanobis (D) et les p-values associées (p) se présentent comme suit :

Élève	Établissement	O	EC	Géo	D	p
1	1	20	11	13	2,22	0,53
3	1	16	10	20	2,13	0,55
4	1	17	0	13	3,98	0,26

Élève	Établissement	O	EC	Géo	D	p
7	1	15	9	8	3,44	0,33
11	1	17	11	13	0,93	0,82
16	1	19	15	20	2,96	0,40
17	1	20	4	11	4,11	0,25
20	1	15	6	7	2,96	0,40
27	1	14	6	16	2,34	0,51
28	1	20	12	20	2,50	0,48
32	1	15	4	17	2,43	0,49
5	2	13	17	12	0,50	0,92
6	2	11		18	8,41	0,04
8	2	20	20	9	3,16	0,37
12	2	10	17	6	1,40	0,70
13	2	11	20	8	1,02	0,80
15	2	17	18	7	1,05	0,79
18	2	12	16	20	6,30	0,10
21	2	8	16	9	1,80	0,62
25	2	20	16	7	3,44	0,33
31	2	12	20	2	2,53	0,47
33	2	11	17	9	0,39	0,94
2	3	12	8	8	2,57	0,46
9	3	7	12	17	2,31	0,51
10	3	9	9	17	1,87	0,60

Élève	Établissement	O	EC	Géo	D	p
14	3	5	13	7	2,28	0,52
19	3	8	8	9	3,73	0,29
22	3	8	12	3	3,01	0,39
23	3	10	12	12	1,43	0,70
24	3	13	6	17	4,03	0,26
26	3	10	12	16	2,54	0,47
29	3	4	14	11	3,13	0,37
30	3	11	12	8	3,11	0,38

Aucune p-value n'est inférieure à 0,001. On peut donc conclure qu'aucune des trois catégories ne comporte de valeur aberrante multivariée.

- On vérifie ensuite la condition de multinormalité. Les résultats du test du coefficient d'aplatissement de Mardia indiquent :

kurt	11,62
z-stat	-1,76
p-value	0,07

La p-value du test est supérieure à 5%, donc la condition de multinormalité est remplie.

- On vérifie enfin la condition d'homogénéité des matrices de variances-covariances :

Statistique du test M de Box	25,79
DDL 1	12,00
DDL 2	5277,46
F	1,86
p-value	0,03

La p-value du test M de box est supérieure à 0,001. Donc la condition d'homogénéité des matrices de variances-covariances est remplie.

Analyse des résultats

Manova : tests de l'effet de l'appartenance catégorielle

Les tests montrent que l'appartenance catégorielle influence significativement le profil en termes de variables-réponses :

p			
Wilks	Hotelling-Lawley	Pillai	Roy
<0,0001	<0,0001	<0,0001	<0,0001

On va donc chercher à identifier les différences significatives entre catégories en termes de variables-réponses.

Test de Tukey de comparaisons multiples[149]

Contrastes entre établissements selon la variable-réponse

	Contraste	Différence standardisée*	p	Signifi-catif
O	3 vs 1	-6,22	<0,0001	Oui
	3 vs 2	-3,28	0,007	Oui
	2 vs 1	-2,94	0,017	Oui
EC	1 vs 2	-7,13	<0,0001	Oui
	1 vs 3	-1,99	0,131	Non
	3 vs 2	-5,14	<0,0001	Oui
Géo	2 vs 1	-2,21	0,08	Non
	2 vs 3	-0,78	0,72	Non
	3 vs 1	-1,43	0,34	Non

* Valeur critique : 2,47

[149] Dans XLSTAT, les tests de comparaisons multiples sont proposés avec les sorties de l'ANOVA. Il faut donc soumettre les données à une procédure d'ANOVA en décochant toutes les options sauf *Sorties / Moyennes / Comparaisons multiples / Appliquer à tous les facteurs / Comparaisons par paires / Tukey*. Parmi les résultats, on ne s'intéresse ici qu'aux comparaisons multiples.

Classification des catégories par variable réponse

	Modalité	Moyenne	Classes		
O	3	8,82	A		
	2	13,18		B	
	1	17,09			C
EC	1	8,00	A		
	3	10,73	A		
	2	17,91		B	
Géo	2	9,73	A		
	3	11,36	A		
	1	14,36	A		

Le test de Tukey montre que l'établissement 1 domine significativement les deux autres à l'épreuve d'orthographe : ses élèves y obtiennent une moyenne significativement supérieure.

L'établissement 2 domine significativement les deux autres à l'épreuve d'éducation civique : ses élèves y obtiennent une moyenne significativement supérieure.

En revanche, il n'y a pas d'écart significatif entre établissements à l'épreuve de géographie.

Le test complémentaire de comparaisons multiples permet donc de préciser l'effet de profil signalé par la manova en distinguant :

- le profil des élèves de l'établissement 1, meilleurs que les autres en orthographe, moins bons que ceux de l'établissement 2 en éducation civique, et semblables aux autres en géographie ;
- le profil des élèves de l'établissement 2, meilleurs que les autres en éducation civique, moins bons que ceux de l'établissement 1 mais meilleurs que ceux de l'établissement 3 en orthographe, et semblables aux autres en géographie ;
- le profil des élèves de l'établissement 3, moins bons que les autres en orthographe, moins bons que ceux de l'établissement 2 mais semblables à ceux de l'établissement 1 en éducation civique, et semblables aux autres en géographie.

15.3.2. Manova multifactorielle

La différence avec la manova à un facteur est qu'il y a ici au moins deux variables indépendantes, ce qui a pour conséquence que les interactions entre variables indépendantes peuvent, elles aussi, exercer un effet sur la variable dépendante. Comme dans l'anova unifactorielle, l'anova multifactorielle, et la manova monofactorielle, les variables indépendantes sont, ici aussi, exclusivement qualitatives.

Les conditions préalables sont les mêmes que pour la manova monofactorielle, sauf à préciser que, étant

donné qu'il y a plusieurs variables indépendantes, c'est au niveau de chaque groupe (combinaison de modalités des catégories) des variables indépendantes, et non pas au niveau des catégories, que se vérifie la condition relative à l'absence de valeurs aberrantes.

EXEMPLE 15.5

Une institution de formation continue souhaite adapter ses approches pédagogiques aux caractéristiques de son public. Elle effectue une enquête auprès de 100 usagers (hommes et femmes âgés de 23 à 48 ans) afin d'identifier leurs caractéristiques au regard de cinq critères :

(a) Ouverture multiculturelle (OM) ;

(b) Esprit d'entreprise, être autonome, jouer un rôle actif (EE) ;

(c) Attirance pour les médias sociaux (MS) ;

(d) Attirance pour les technologies mobiles (TM) ;

(e) Préférence pour le travail individuel (TI).

Le Tableau 15.5 récapitule les données collectées. On cherche à discerner d'éventuels profils d'apprenants en termes de ces cinq critères.

Tableau 15.5.
Genre, tranche d'âge et caractérisation au regard de cinq critères dans un échantillon de 100 stagiaires de formation continue

Note
- Genre (H/F) : 1=Homme ; 2=Femme
- Tranche d'âge : 1=18-21ans ; 2=22-32 ans ; 3=33-48 ans
- Caractérisation au regard des critères : pour chaque critère, score obtenu par le répondant sur une échelle de 1 à 100

Id.	H/F	Age	OM	EE	MS	TM	TI
1	2	3	45	53	19	44	68
2	1	3	52	17	57	23	78
3	2	1	83	35	40	81	47
4	2	3	50	28	78	2	98
5	2	1	75	46	43	70	55
6	2	1	76	50	30	86	34
7	1	3	24	60	54	52	67
8	2	1	67	38	55	100	44
9	1	3	21	12	41	17	77
10	1	2	47	73	64	89	40
11	2	2	29	100	73	14	26
12	1	3	48	14	59	30	82
13	2	3	18	49	58	29	77
14	1	1	78	62	35	73	31
15	2	3	38	55	21	42	73
16	1	2	1	83	100	38	36

Id.	H/F	Age	OM	EE	MS	TM	TI
17	2	1	69	13	24	75	53
18	2	1	60	58	56	80	42
19	2	1	100	38	57	98	32
20	2	2	35	66	99	41	20
21	1	2	15	68	75	40	29
22	2	3	24	1	34	44	57
23	1	3	36	19	36	56	95
24	1	3	51	55	55	55	83
25	1	1	59	22	50	74	45
26	2	1	68	18	1	94	24
27	1	3	35	1	52	46	70
28	2	3	25	34	35	38	69
29	2	1	73	32	54	71	48
30	2	1	69	35	49	70	9
31	2	2	39	64	84	66	47
32	2	3	49	16	29	13	77
33	1	1	100	59	36	79	39
34	2	1	61	58	44	92	5
35	1	2	47	83	97	47	22
36	2	3	41	58	47	48	59
37	2	2	41	100	63	11	26
38	1	3	38	35	47	55	59
39	1	2	15	71	70	17	35
40	1	1	93	47	33	78	31
41	2	2	38	100	66	41	30
42	1	2	55	70	67	61	55
43	1	1	68	21	29	96	40
44	2	2	45	83	69	62	46

Id.	H/F	Age	OM	EE	MS	TM	TI
45	2	3	54	26	55	42	70
46	2	1	74	38	41	100	48
47	1	2	54	66	69	50	35
48	1	3	18	40	26	27	69
49	1	3	36	43	39	47	67
50	1	3	40	37	18	25	62
51	2	1	76	55	30	83	51
52	2	2	41	82	77	33	52
53	2	3	47	58	30	50	89
54	1	1	69	52	52	70	31
55	2	2	55	74	80	51	1
56	2	2	28	71	66	34	42
57	2	1	66	1	16	96	52
58	1	1	70	31	58	85	57
59	2	2	45	68	100	50	57
60	2	2	30	77	79	1	56
61	2	1	73	56	11	69	41
62	2	1	78	25	22	76	38
63	1	1	63	15	59	66	2
64	1	2	48	91	91	27	32
65	2	3	29	60	35	1	100
66	2	3	56	53	28	19	81
67	2	3	32	9	50	28	79
68	1	3	12	26	50	29	99
69	2	2	25	78	97	20	16
70	2	3	43	62	45	47	67
71	2	2	1	87	100	45	42
72	1	1	71	1	51	71	33

Id.	H/F	Age	OM	EE	MS	TM	TI
73	1	1	83	27	31	67	26
74	1	2	48	84	79	34	52
75	1	2	9	75	60	65	27
76	2	3	32	37	28	4	98
77	2	3	29	25	36	40	90
78	1	3	2	55	49	53	93
79	2	3	48	58	26	12	92
80	2	2	10	76	100	56	45
81	2	2	15	66	83	52	53
82	1	1	69	27	50	80	24
83	2	1	75	8	49	69	40
84	1	2	45	63	72	1	29
85	1	2	42	66	65	57	1
86	2	1	70	23	22	100	20
87	1	3	46	34	28	39	96
88	1	1	69	23	25	66	23
89	1	2	49	63	91	27	55
90	1	2	32	87	65	49	54
91	1	1	90	55	26	68	46
92	2	1	59	61	51	68	40
93	1	2	58	69	75	14	3
94	1	2	35	87	66	16	1
95	1	1	58	5	43	79	16
96	2	3	42	47	9	36	57
97	1	2	45	75	83	43	54
98	1	3	24	9	52	50	62
99	2	2	16	75	90	42	56
100	1	3	32	46	38	55	66

Vérification des conditions de validité

- Les valeurs aberrantes doivent être détectées pour chaque groupe. On identifie six groupes à partir des modalités des variables indépendantes :

		Genre	
		1 : Hommes	2 : Femmes
Tranche d'âge	1 : 18-21 ans	Groupe 1	Groupe 4
	2 : 22-32 ans	Groupe 2	Groupe 5
	3 : 33-48 ans	Groupe 3	Groupe 6

- Valeurs aberrantes univariées

Les variables dépendantes sont normalement distribuées :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
OM	0,55	0,60	0,30	0,61
EE	0,04	0,06	0,04	0,19
MS	0,08	0,15	0,30	0,26
TM	0,12	0,47	0,35	0,32
TI	0,13	0,44	0,39	0,50

Le test de Grubbs révèle la présence de valeurs aberrantes :

		Groupes					
		1	2	3	4	5	6
OM	G	2,01	2,12	2,10	3,01	2,05	1,88
	Val. crit.	2,50	2,62	2,58	2,68	2,58	2,65
	p	0,42	0,39	0,38	0,007	0,45	0,87
EE	G	1,55	1,75	1,68	1,95	1,71	2,06
	Val. crit.	2,50	2,62	2,58	2,68	2,58	2,65
	p	0,49	0,85	0,70	0,76	0,79	0,51
MS	G	1,48	1,85	2,13	2,11	1,48	2,51
	Val. crit.	2,50	2,62	2,58	2,68	2,58	2,65
	p	0,21	0,67	0,34	0,47	<0,0001	0,09
TM	G	2,46	2,25	1,75	1,38	2,00	1,69
	Val. crit.	2,50	2,62	2,58	2,68	2,58	2,65
	p	0,06	0,24	0,92	<0,0001	0,52	0,57
TI	G	2,16	1,73	1,67	2,29	2,24	1,57
	Val. crit.	2,50	2,62	2,58	2,68	2,58	2,65
	p	0,23	0,78	0,67	0,25	0,23	0,05

Le test indique trois cas de valeurs aberrantes univariées :

(a) la valeur 100 de l'Identifiant 19 pour la variable Ouverture multiculturelle au sein du Groupe 4 ;

(b) la valeur 100 de l'Identifiant 8 pour la variable Attirance pour les technologies mobiles au sein du Groupe 4 ; et

(c) la valeur 63 de l'Identifiant 37 pour la variable Attirance pour les médias sociaux au sein du Groupe 5.

Ces trois valeurs sont retirées de l'échantillon pour la suite des analyses.

- Valeurs aberrantes multivariées

Les distances de Mahalanobis (D) s'établissent comme suit :

Identifiant	Groupe	D	p-value
14	1	4,11	0,53
25	1	4,22	0,52
33	1	5,14	0,40
40	1	3,01	0,70
43	1	8,32	0,14
54	1	4,42	0,49
58	1	6,15	0,29

Identifiant	Groupe	D	p-value
63	1	6,54	0,26
72	1	5,85	0,32
73	1	2,76	0,74
82	1	1,68	0,89
88	1	5,11	0,40
91	1	4,33	0,50
95	1	3,37	0,64
10	2	5,50	0,36
16	2	8,47	0,13
21	2	2,83	0,73
35	2	5,80	0,33
39	2	4,23	0,52
42	2	3,36	0,64
47	2	1,81	0,87
64	2	4,61	0,47
74	2	2,94	0,71
75	2	5,13	0,40
84	2	5,24	0,39
85	2	5,49	0,36
89	2	5,66	0,34
90	2	5,46	0,36
93	2	4,92	0,43
94	2	6,98	0,22
97	2	1,59	0,90
2	3	5,45	0,36
7	3	4,63	0,46
9	3	4,13	0,53

Identifiant	Groupe	D	p-value
12	3	3,91	0,56
23	3	7,12	0,21
24	3	5,78	0,33
27	3	4,33	0,50
38	3	2,65	0,75
48	3	4,19	0,52
49	3	1,08	0,96
50	3	6,45	0,26
68	3	5,30	0,38
78	3	6,90	0,23
87	3	5,88	0,32
98	3	5,18	0,39
100	3	1,99	0,85
3	4	1,82	0,87
5	4	2,02	0,85
6	4	1,44	0,92
8	4	13,20	0,02
17	4	2,80	0,73
18	4	2,63	0,76
19	4	15,12	0,01
26	4	5,97	0,31
29	4	2,45	0,78
30	4	5,98	0,31
34	4	6,47	0,26
46	4	3,12	0,68
51	4	3,10	0,68
57	4	5,36	0,37

Identifiant	Groupe	D	p-value
61	4	6,91	0,23
62	4	1,13	0,95
83	4	5,17	0,40
86	4	3,05	0,69
92	4	2,25	0,81
11	5	4,47	0,48
20	5	3,57	0,61
31	5	3,29	0,66
37	5	11,08	0,05
41	5	4,68	0,46
44	5	4,15	0,53
52	5	2,55	0,77
55	5	7,36	0,19
56	5	2,43	0,79
59	5	4,84	0,44
60	5	7,60	0,18
69	5	4,67	0,46
71	5	5,73	0,33
80	5	3,08	0,69
81	5	3,86	0,57
99	5	1,63	0,90
1	6	1,98	0,85
4	6	9,00	0,11
13	6	7,01	0,22
15	6	1,61	0,90
22	6	6,81	0,23
28	6	1,91	0,86

Identifiant	Groupe	D	p-value
32	6	5,15	0,40
36	6	4,60	0,47
45	6	4,96	0,42
53	6	7,98	0,16
65	6	5,95	0,31
66	6	3,12	0,68
67	6	3,02	0,70
70	6	3,36	0,64
76	6	4,00	0,55
77	6	6,29	0,28
79	6	2,82	0,73
96	6	5,42	0,37

Aucun groupe ne comporte d'observation dont la distance serait associée à une p-value inférieure à 0,001. Il n'y a donc pas de valeur aberrante multivariée.

- Multinormalité

Les résultats du test du coefficient d'aplatissement de Mardia indiquent :

kurt	35,28
z-stat	0,17
p-value	0,86

La p-value du test est supérieure à 5%, donc la condition de multinormalité est remplie.

- Homogénéité des matrices de variances-covariances

Les résultats du test M de Box indiquent :

Statistique du test M de Box	122,11
DDL 1	75
DDL 2	15832
F	1,41
p-value	0,010

La p-value du test est supérieure à 0,001. La condition d'homogénéité des variances-covariances est donc remplie.

Analyse des résultats

	p			
	Wilks	Hotelling -Lawley	Pillai	Roy
Genre	0,40	0,40	0,40	0,37
Age	<0,0001	<0,0001	<0,0001	<0,0001
Interaction entre genre et âge	0,10	0,09	0,10	0,02

La tranche d'âge d'appartenance influence significativement le profil des répondants. En revanche

aucune influence significative n'est imputable au genre ou à la combinaison genre-âge. On applique un test de comparaisons multiples de Tukey afin d'identifier les profils en fonction de l'âge.

Contrastes entre établissements selon la variable-réponse

	Contrastes entre âges	Différence standardisée*	p	Significatif
OM	1 vs 2	11,44	<0,0001	Oui
	1 vs 3	11,18	<0,0001	Oui
	3 vs 2	0,34	0,93	Non
EE	2 vs 1	10,45	<0,0001	Oui
	2 vs 3	10,07	<0,0001	Oui
	3 vs 1	0,45	0,892	Non
MS	2 vs 1	11,62	<0,0001	Oui
	2 vs 3	11,31	<0,0001	Oui
	3 vs 1	0,38	0,92	Non
TM	1 vs 3	10,97	<0,0001	Oui
	1 vs 2	9,84	<0,0001	Oui
	2 vs 3	1,05	0,54	Non
TI	3 vs 1	11,26	<0,0001	Oui
	3 vs 2	11,19	<0,0001	Oui
	2 vs 1	0,06	0,99	Non

* Valeur critique : 2,38

Moyennes des variables pour chaque tranche d'âge

	Modalité	Moyenne	Classification des tranches d'âge	
OM	1	71,868	A	
	3	35,521		B
	2	34,387		B
EE	2	77,000	A	
	3	36,237		B
	1	34,392		B
MS	2	79,334	A	
	3	39,821		B
	1	38,469		B
TM	1	78,631	A	
	2	39,123		B
	3	34,907		B
TI	3	77,232	A	
	2	35,608		B
	1	35,365		B

On constate que chaque tranche d'âge a un profil particulier au regard des critères fixés par l'institution de formation :

- les 18-21 ans (tranche d'âge 1) se distinguent des autres par une ouverture multiculturelle et une

attirance pour les technologies mobiles significativement plus marquées ;

- les 22-32 ans (tranche 2) se distinguent des autres par des scores significativement plus élevés aux critères « Esprit d'entreprise, être autonome, jouer un rôle actif » et « Attirance pour les médias sociaux » ;

- Les 33-48 ans (tranche 3) se démarquent des autres par une préférence significativement plus forte pour le travail individuel.

Troisième partie : Modéliser et analyser des relations entre variables

L'analyse des données permet de détecter l'existence de relations entre variables. La modélisation vise à caractériser plus avant les relations mises en évidence. Modéliser consiste à exprimer de façon formelle et calculable la relation entre plusieurs variables. La modélisation permet de préciser la forme de la relation, d'en spécifier les paramètres, d'analyser de façon plus fine la dynamique des variables et les interactions entre elles, et de déterminer la portée de la relation.

Plusieurs procédures de modélisation sont présentées dans cette partie. Elles se distinguent par la nature (quantitative et/ou qualitative, manifeste ou latente) des variables qu'elles impliquent, et par la forme (linéaire ou non) des relations dont elles traitent.

Ici également, la vérification des conditions de validité est essentielle : les modèles, les analyses et les prédictions qu'ils permettent ne sont fiables que s'ils respectent les conditions sur lesquelles repose leur validité. Vérifier les conditions de validité, corriger les sources de fragilité et en rendre compte dans le rapport d'analyse constitue un aspect crucial de la crédibilité du travail du chercheur quantitatif.

Chapitre 16. Modéliser une relation entre des facteurs qualitatifs et une variable-réponse quantitative

On suppose une anova dont les résultats montrent que l'appartenance à tel ou tel groupe des variables qualitatives indépendantes exerce un effet sur la variable quantitative dépendante. Soit par exemple une anova dont les résultats montrent qu'en matière d'intégration scolaire, l'action pédagogique est plus efficace dans telle classe d'âge que des mesures socioculturelles, le suivi psychologique ou la médicalisation pour amener l'élève en difficulté à réaliser au mieux son potentiel. Ou encore une anova qui montre que les structures d'accueil qui privilégient la prise en charge individuelle favorisent davantage le développement cognitif du jeune enfant que celles qui donnent la priorité à la stabilité de l'encadrement ou à la durée de la période de garde. Etc. Ce chapitre présente les principes permettant d'établir un modèle synthétique rendant compte des relations entre les variables indépendantes et la variable dépendante, d'en spécifier les paramètres, d'identifier la significativité de ceux-ci, et de vérifier la validité du modèle.

16.1. DÉFINITION DU MODÈLE

Dans le cas général (anova multifactorielle) on dispose d'un échantillon dont les n individus $(1,2,\ldots,n)$ sont

caractérisés au regard de ω variables qualitatives indépendantes $(X_1, X_2, \dots, X_\omega)$ et une variable quantitative dépendante Y. Le nombre de modalités d'une variable est noté ξ, donc la variable X_1 comporte ξ_1 modalités, la variable X_2 comporte ξ_2 modalités, et ainsi de suite jusqu'à la variable X_ω qui comporte ξ_ω modalités. Par exemple, pour une anova à trois facteurs comportant chacun au moins quatre modalités, on peut se représenter comme suit les cinq premiers membres de l'échantillon :

Identifiants des individus	Profil des individus			Valeurs dans Y
	Modalité dans X_1	Modalité dans X_2	Modalité dans X_3	
$i1_{314}$	3	1	4	$y1_{314}$
$i2_{442}$	4	4	2	$y2_{442}$
$i3_{413}$	4	1	3	$y3_{413}$
$i4_{413}$	4	1	3	$y4_{413}$
$i5_{112}$	1	1	2	$y5_{112}$
⋮	⋮	⋮	⋮	⋮

Chaque individu est caractérisé par la modalité qui lui correspond dans chaque variable indépendante, et par la valeur qui lui correspond dans la variable quantitative dépendante. Plusieurs individus peuvent présenter la même combinaison de modalités, c'est-à-

dire le même profil (par exemple ici les individus 3 et 4).

On peut prévoir, avec un certain degré de confiance et en acceptant le risque de commettre une erreur ε, une valeur $\hat{y}_i$ approchée de la valeur y_i réelle de la variable *Y* pour un individu *i* présentant un profil donné :

$$\hat{y}_i = \alpha + (\beta_1 + \beta_2 + \cdots + \beta_\omega) + \gamma$$

où

- α est une constante ;
- β_1 est le coefficient qui caractérise la modalité de X_1 correspondant au profil de l'individu *i* ;
- β_2 est le coefficient qui caractérise la modalité de X_2 correspondant au profil de l'individu *i* ;
- β_ω est le coefficient qui caractérise la modalité de X_ω correspondant au profil de l'individu *i* ;
- γ est le coefficient qui caractérise l'interaction entre les modalités correspondant au profil de l'individu *i*.

Les logiciels statistiques fournissent des estimations $(\hat{\alpha}, \hat{\beta}, et\, \hat{\gamma})$ des valeurs des paramètres. Voici par exemple les paramètres obtenus à partir d'une anova à deux facteurs, X_1 (de modalités D1, D2, D3, D4) et X_2 (de modalités S1, S2, S3, S4) :

Paramètres du modèle							
Libellés des paramètres (1)		VP* (2)	ES** (3)	t (4)	Pr > |t| (5)	Borne inf. (95%) (6)	Borne sup. (95%) (7)
Constante ($\hat{\alpha}$)		0,41	0,01	38,38	<0,0001	0,39	0,43
Coefficients des modalités ($\hat{\beta}$)	D1	0,02	0,01	1,26	0,219	-0,01	0,05
	D2	-0,01	0,02	-0,59	0,560	-0,06	0,03
	D3	-0,02	0,01	-1,57	0,129	-0,05	0,00
	D4	0,02	0,02	0,96	0,346	-0,02	0,07
	S1	0,07	0,02	3,67	0,001	0,03	0,12
	S2	0,43	0,02	20,27	<0,0001	0,38	0,47
	S3	-0,28	0,02	-13,56	<0,0001	-0,32	-0,24
	S4	0,10	0,02	4,59	0,000	0,05	0,15
Coefficients des interactions ($\hat{\gamma}$)	D1S1	-0,01	0,02	-0,50	0,621	-0,07	0,04
	D1S2	0,06	0,03	2,03	0,058	-0,002	0,12
	D1S3	0,02	0,02	0,97	0,340	-0,03	0,08
	D1S4	-0,09	0,03	-2,59	0,017	-0,17	-0,01
	D2S1	-0,03	0,04	-0,79	0,436	-0,12	0,05
	D2S2	0,07	0,03	2,06	0,052	-0,001	0,14
	D2S3	-0,03	0,04	-0,77	0,448	-0,12	0,05
	D2S4	-0,00	0,03	-0,05	0,960	-0,07	0,07
	D3S1	0,02	0,03	0,67	0,509	-0,04	0,08
	D3S2	-0,07	0,03	-2,22	0,037	-0,13	-0,005
	D3S3	-0,01	0,03	-0,38	0,703	-0,07	0,05
	D3S4	0,07	0,03	2,27	0,033	0,007	0,14
	D4S1	0,02	0,04	0,60	0,553	-0,07	0,13
	D4S2	-0,04	0,04	-1,53	0,139	-0,17	0,02
	D4S3	0,01	0,04	0,20	0,837	-0,09	0,11
	D4S4	0,05	0,04	1,31	0,202	-0,03	0,14

* VP : Valeurs des paramètres ** ES : Erreur standard

Ce tableau des paramètres présente :

- en colonne (2), la valeur du paramètre estimé ;
- en colonne (3), l'erreur standard de chaque coefficient. L'erreur standard (ou erreur-type) d'un coefficient mesure la précision de l'estimation de ce coefficient, et est représentée par l'écart-type de ce coefficient dans l'échantillon. Plus petite est l'erreur standard, plus précise est l'estimation ;
- en colonne (4), la statistique de Student, obtenue en divisant la valeur du paramètre [colonne (2)] par l'erreur standard [colonne (3)] ;
- en colonne (5), la probabilité p associée à la valeur calculée de la statistique de Student, étant donné la valeur critique pertinente de la table de Student[150]. Au seuil de significativité de 5%, le paramètre est significatif si $p < 0{,}05$;
- en colonnes (6) et (7), les bornes des intervalles de confiance. Les intervalles de confiance fournissent, sous une autre forme, la même information que la probabilité p. Au seuil de significativité de 5%, le paramètre est significatif si sa valeur se situe entre les

[150] C'est la probabilité *p* que la valeur calculée de la statistique de Student soit supérieure à la valeur critique. Obtenir une valeur calculée supérieure à la valeur critique est d'autant moins probable que la valeur calculée est élevée, et plus la probabilité est faible, plus la significativité est forte.

bornes (il y a 95% de chances que la vraie valeur du paramètre pour l'ensemble de la population se situe entre les bornes), à condition que zéro ne soit pas compris dans l'intervalle. Par exemple pour l'interaction D2S1, la valeur du paramètre est comprise dans l'intervalle de confiance $[-0{,}127\ ;\ 0{,}057]$, donc l'intervalle contient 0, ce qui signifie que le paramètre n'est pas significatif, ce que montre aussi la probabilité *p*. En revanche, pour l'interaction D3S4, la valeur du paramètre est comprise dans l'intervalle $[0{,}007\ ;\ 0{,}147]$ qui ne contient pas 0, donc le paramètre est significatif.

On peut donc, connaissant les paramètres estimés, calculer la valeur $\widehat{y_i}$ de la variable *Y* pour un individu *i* dont on connait le profil. Par exemple, à partir du tableau des paramètres ci-dessus, si le profil de l'individu correspond à la modalité 3 de la première variable indépendante (D3) et à la modalité 1 de la deuxième variable indépendante (S1), la valeur estimée $\widehat{y_i}$ de la variable *Y* sera pour cet individu :

$$\begin{aligned}\widehat{y_i} &= \widehat{\alpha} + (\widehat{\beta_1} + \widehat{\beta_2}) + \widehat{\gamma} \\ &= constante + (D3 + S1) + D3S1 \\ &= 0{,}412 + (-0{,}025 + 0{,}077) \\ &\quad + 0{,}022 \\ &= 0{,}486\end{aligned}$$

ce qui représente une approximation à ε près de la vraie valeur y_i de *Y* pour l'individu :

$$y_i = \hat{y}_i + \varepsilon$$

La question est alors de déterminer la fiabilité et la validité de cette prévision.

Pour que la prévision soit fiable, il faut que les paramètres sur lesquels elle s'appuie soient le soient, c'est-à-dire qu'ils soient significatifs. Dans l'exemple utilisé ici, la constante et S1 sont significatifs, mais ni D3 ni D3S1 ne le sont. Les paramètres ne sont donc pas fiables.

Pour que la prévision soit valide, il faut que le modèle sur lequel elle s'appuie le soit. Pour que le modèle puisse être considéré comme valide, trois conditions doivent être remplies : normalité des résidus ; stabilité de la variance des résidus (condition d'*homoscédasticité*) ; et indépendance des résidus.

La vérification de ces trois conditions repose donc sur les "résidus". On appelle résidu (noté ε_i) l'erreur commise lors de l'estimation de $\hat{y}_i$ pour chaque individu i :

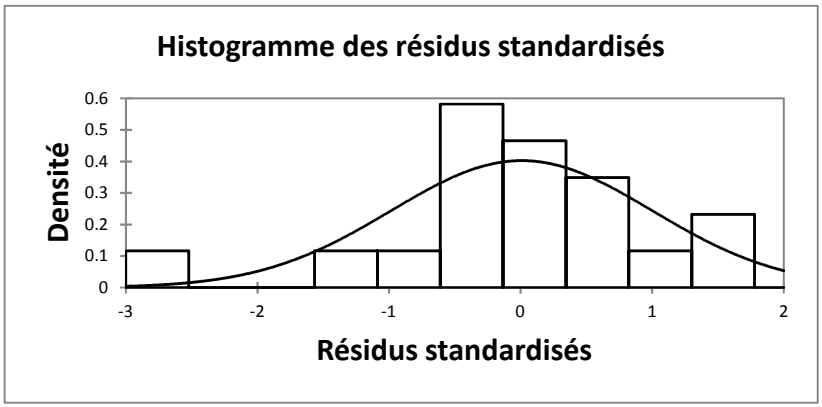
$$\varepsilon_i = y_i - \hat{y}_i$$

Lorsque le modèle est parfait, les résidus sont nuls. Plus la valeur des résidus s'écarte de zéro, plus on peut considérer que le modèle peine à traduire la relation entre X et Y. Il faut donc analyser les résidus pour voir

dans quelle mesure le modèle est pertinent pour décrire la relation entre *X* et *Y*.

16.2. NORMALITÉ DES RÉSIDUS

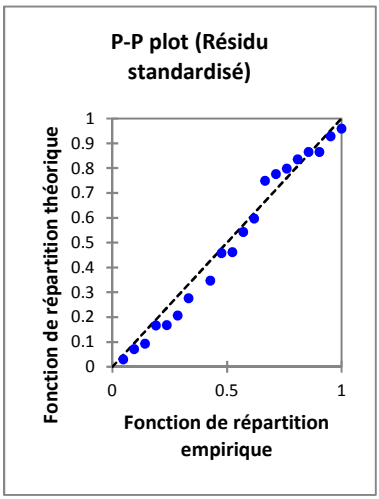
Les tests de significativité des paramètres et les intervalles de confiance sont définis sous hypothèse de normalité. Ils ne sont donc pas valides lorsque les séries ne sont pas normales. La vérification de la normalité s'effectue au niveau des résidus et au moyen des tests de normalité usuels. Des graphiques peuvent être utiles aussi, par exemple l'histogramme des résidus standardisés[151] et les diagrammes PP et QQ :

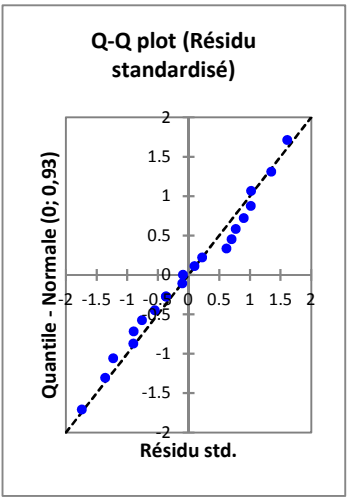


[151] Les logiciels utilisent indifféremment les expressions "résidus standardisés" et "résidus normalisés". Il s'agit généralement des résidus standardisés par centration-réduction (voir chapitre 3). L'expression "résidus normalisés" ne signifie pas que les résidus suivent une distribution normale.

Dans cet exemple, on peut considérer que les résidus standardisés sont normalement distribués.

Les diagrammes Probabilité-Probabilité (ou « diagrammes PP » ou « P-P plots ») et les diagrammes Quantile-Quantile (ou « diagrammes QQ », ou « Q-Q plots ») se lisent de la même façon : la distribution des résidus est d'autant plus proche de la distribution normale que les points sont alignés sur la diagonale.

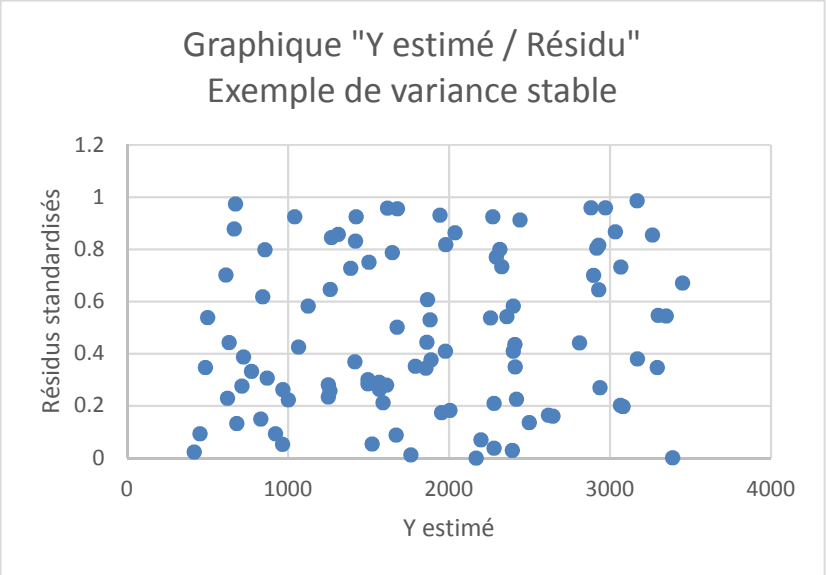




Une distribution en S nettement démarquée autour de la diagonale peut signifier que les résidus ne sont pas normalement distribués.

16.3. STABILITÉ DE LA VARIANCE DES RÉSIDUS : CONDITION D'HOMOSCÉDASTICITÉ

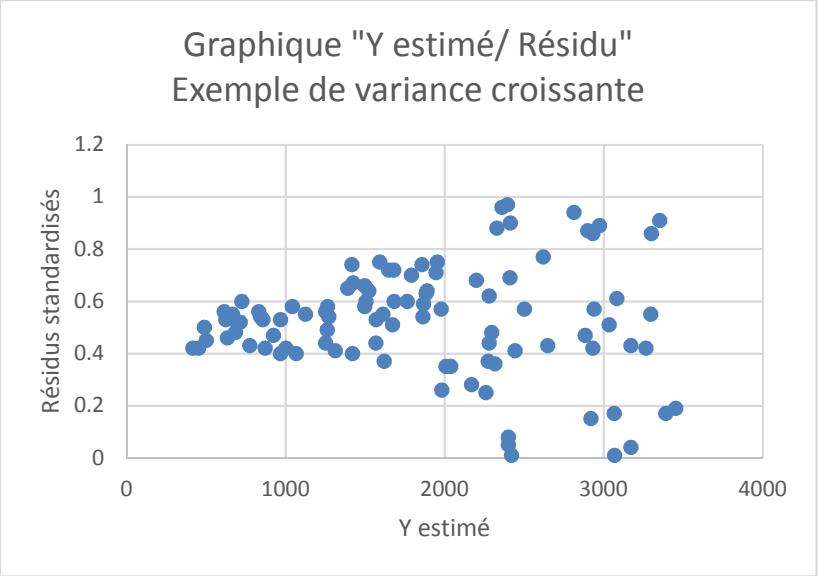
La condition d'homoscédasticité signifie que la variance des résidus est stable quelle que soit la valeur de la variable dépendante. Elle peut se vérifier au moyen du graphique « $\widehat{Y}$/Résidu »[152] sur lequel chaque individu *i* de l'échantillon est représenté par le point de coordonnées $(\hat{y}_i\ ;\ \varepsilon_i)$:



On voit sur ce graphique que la dispersion des résidus est de même ampleur que l'on se situe dans les valeurs basses ou dans les valeurs élevées de $\widehat{Y}$. La condition

[152] Qu'on appelle encore « Y estimé / Résidu » ou « Prédictions / Résidus ».

d'homoscédasticité est remplie. Elle ne l'est plus dans le graphique suivant :



On voit ici que la dispersion croît lorsque la valeur de la variable dépendante augmente. La condition d'homoscédasticité n'est pas remplie. On dit alors qu'il y a "*hétéroscédasticité*".

Cependant, les graphiques ne permettent pas toujours de trancher. On peut alors recourir au test de Breusch-Pagan ou au test de White, qui permettent de tester directement l'homoscédasticité[153]. À défaut, la

[153] Mais ces tests ne sont pas systématiquement accessibles sur tous les logiciels. Les logiciels sophistiqués permettent

variable indépendante étant qualitative dans une anova, on peut tester l'homogénéité des résidus répartis par catégories du facteur. On utilise à cette fin les tests usuels d'homogénéité des variances (tests de Levene, Bartlett, et Fisher si les séries sont normalement distribuées, test de Fligner-Killeen sinon).

16.4. INDÉPENDANCE DES RÉSIDUS

Il y a indépendance des résidus si :

(a) les résidus sont distribués de façon aléatoire ;

(b) les résidus sont indépendants de la variable explicative ; et

(c) les résidus ne sont pas corrélés entre eux (sinon il y a "autocorrélation des résidus", par exemple la n-*ième* erreur engendre la (n+1)-*ième* erreur).

16.4.1. Distribution aléatoire des résidus

Cette condition se vérifie elle aussi au moyen du graphique « *Y* estimé / Résidu ». Par exemple, la condition n'est pas remplie si les erreurs sont systématiquement du même signe (ce qui signifie que le modèle surestime ou sous-estime systématiquement la valeur de *Y*). Plus généralement, la condition de

---

d'effectuer directement des estimations suivant la procédure de White, robuste à l'hétéroscédasticité.

distribution aléatoire des résidus n'est pas remplie chaque fois que les erreurs ne semblent pas distribuées au hasard, mais semblent plutôt refléter un comportement systématique, une configuration, une structure ou une tendance particulière[154]. La vérification peut s'effectuer aussi au moyen du calcul des corrélations entre résidus et variable dépendante estimée, une forte corrélation signifiant évidemment que la distribution des résidus n'est pas aléatoire.

16.4.2. Indépendance des résidus par rapport à chaque variable indépendante

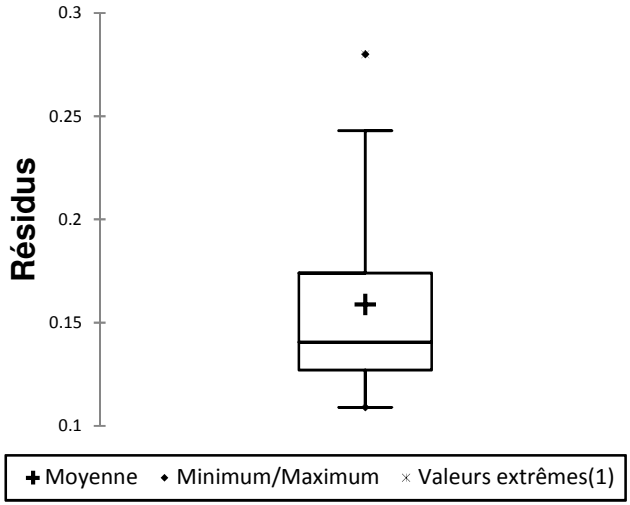
On peut vérifier cette condition par l'examen du nuage de points "Variable indépendante / Résidu". Dans une anova, la variable indépendante étant qualitative, on place en abscisse du nuage de points les individus de

[154] Une structure, une tendance ou toute autre configuration non purement aléatoire indique que les résidus contiennent des éléments de régularité ou de déterminisme, donc de prédictibilité. Un modèle $\hat{Y}$ correct, c'est-à-dire un modèle $\hat{Y}$ dont les variables indépendantes contiennent tous les prédicteurs importants de Y a pour conséquence que le résidu ne recèle plus que des erreurs aléatoires. Un résidu non-aléatoirement distribué indique donc qu'il manque parmi les variables indépendantes des facteurs explicatifs, par exemple d'autres variables ou des interactions entre variables, qui ont été reléguées dans les résidus, d'où il faut les extraire pour les réintégrer dans le modèle.

l'échantillon répartis par catégorie. Le résidu associé à chaque individu est reporté en ordonnée. On peut alors voir si la répartition des résidus est indifférente à la catégorie, ou si au contraire une ou plusieurs catégories se distinguent par une répartition spécifique.

Un autre mode de visualisation utile repose sur l'usage de *boites à moustaches*[155]. Les boites à moustaches

[155] Une boite à moustaches (*boxplot*) est un graphique qui présente sous forme synthétique les extrema, la médiane, et les quartiles d'une série :



Sur ce graphique synthétisant les caractéristiques des résidus standardisés d'une catégorie d'une variable qualitative, on voit le minimum de la série de résidus (le point inférieur : 0,11) ; son maximum (le point supérieur : 0,28), qui en l'occurrence est aussi une valeur aberrante ; le premier quartile (base du

permettent de visualiser, pour chaque catégorie de la variable indépendante, la dispersion et surtout ici les extrema et la moyenne des résidus. Il s'agit d'apprécier si une catégorie se distingue si radicalement des autres par sa moyenne et ses extrema qu'elle aurait pu influencer le coefficient de corrélation entre variable indépendante et résidu si on avait pu en calculer un (comme cela aurait été le cas si la variable indépendante avait été quantitative ou ordinale). Les boites à moustaches sont, pour cette raison, utiles pour apprécier l'indépendance des résidus par rapport aux catégories d'une variable indépendante qualitative.

Lorsque la nature de la variable indépendante s'y prête, on peut calculer le coefficient de corrélation bisérielle ponctuelle (si la variable indépendante est nominale binaire) ou les coefficients de Spearman et Kendall (si la variable indépendante est ordinale).

---

rectangle : 0,127), le deuxième quartile ou médiane (la ligne qui divise le rectangle en deux : 0,14), et le troisième quartile (le coté supérieur du rectangle : 0,174). Le graphique affiche en outre, sous forme de croix / signe plus, la moyenne de la série de résidus. Suivant le logiciel, il est aussi possible de faire varier la largeur du rectangle en fonction de la taille de la série, et de faire apparaitre l'intervalle de confiance autour de la médiane. Dans XLSTAT, les boites à moustaches sont accessibles à partir de la commande *Visualisation des données / Graphiques univariés*.

EXEMPLE 16.1.
Indépendance des résidus par rapport à la variable indépendante

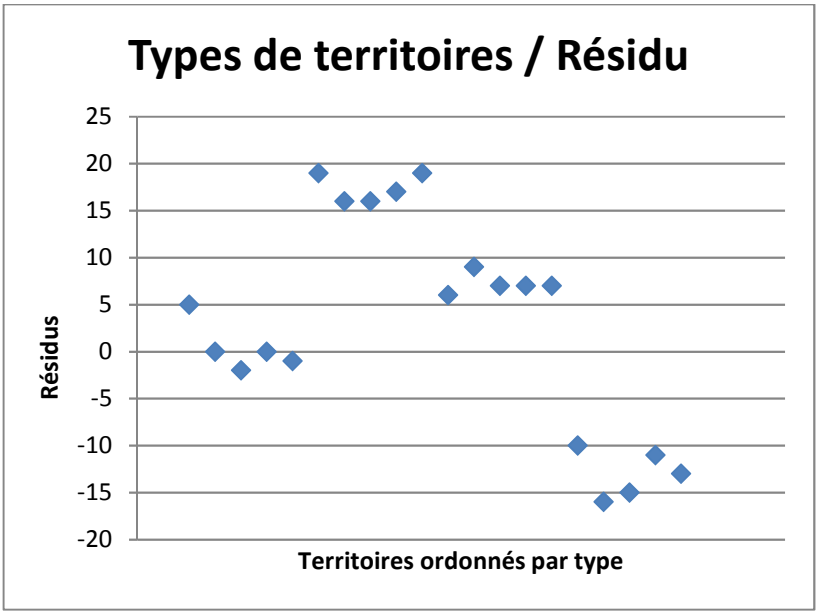
En vue de la mise en place d'actions de formation pour la reconversion de chômeurs dans les territoires, une estimation du nombre de candidats potentiels à ces formations a été effectuée par type de territoire (centres-villes, périphéries urbaines, zones rurales, zones frontalières). L'analyse de variance montre que la variabilité interclasses du nombre de candidats potentiels est significativement supérieure à la variabilité intraclasses. Il y a donc un effet significatif du type de territoire sur le nombre de candidats potentiels. On en tire un modèle de prévision du nombre de places de formation par type de territoire. On cherche à vérifier la validité du modèle, et on s'interroge sur l'indépendance des résidus par rapport à la variable indépendante. Le Tableau 16.1 montre les résidus par type de territoire.

Tableau 16.1.
Résidus par territoire et type de territoire.

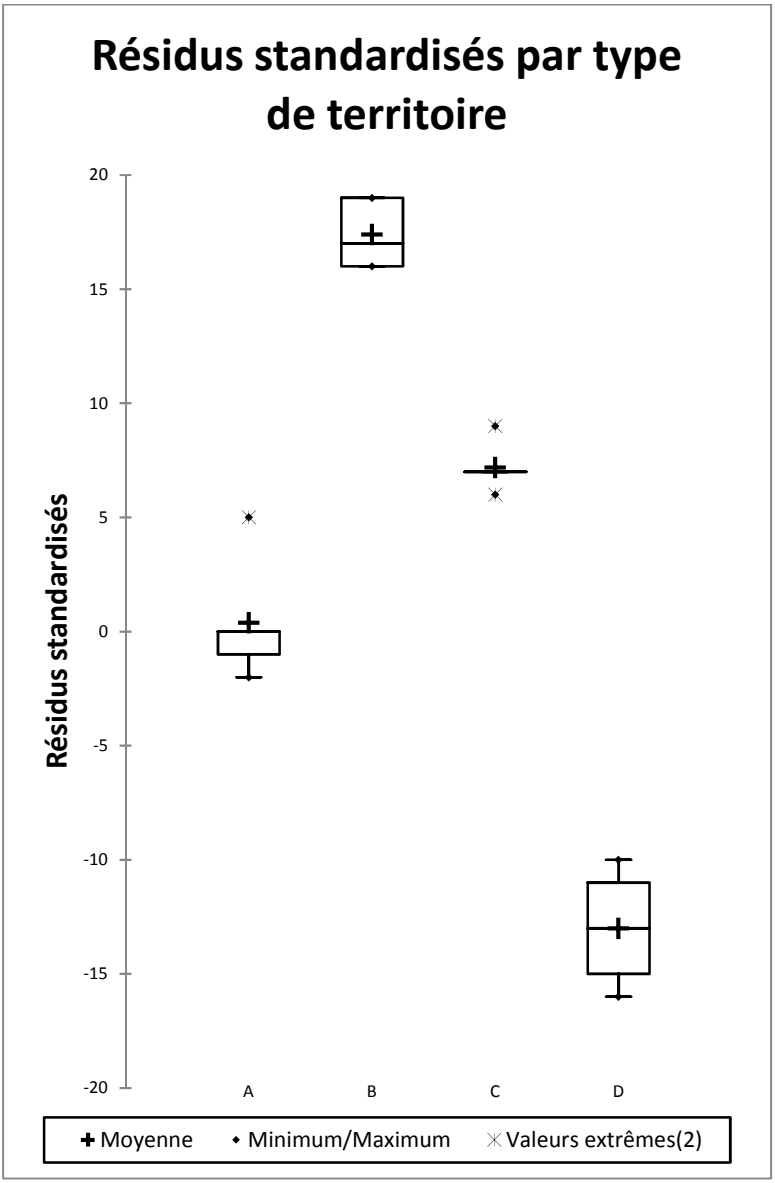
Note
Types de territoires : A – Centre-ville; B – Périphérie; C – Zone rurale; D – Zone frontalière

Territoire	Type	Résidus
T1	A	5
T2	A	0
T3	A	-2
T4	A	0
T5	A	-1
T6	B	19
T7	B	16
T8	B	16
T9	B	17
T10	B	19
T11	C	6
T12	C	9
T13	C	7
T14	C	7
T15	C	7
T16	D	-10
T17	D	-16
T18	D	-15
T19	D	-11
T20	D	-13

Le graphique Variable indépendante / Résidu exprime cette information :



On peut constater que les résidus s'établissent entre -5 et 5 pour les centres-villes ; entre 15 et 20 pour les périphéries urbaines ; entre 5 et 10 pour les zones rurales ; et sont systématiquement négatifs pour les zones frontalières. On ne peut manifestement pas prétendre que les résidus sont indépendants de la catégorie. La condition d'indépendance des résidus par rapport à la variable indépendante n'est pas remplie. Les boites à moustaches conduisent à la même conclusion :


Résidus standardisés par type de territoire
Résidus standardisés
20
15
10
5
0
-5
-10
-15
-20
A
B
C
D
Moyenne
Minimum/Maximum
Valeurs extrêmes(2)

16.4.3. Non-autocorrélation des résidus

L'autocorrélation des résidus[156] s'analyse généralement au moyen de la statistique de Durbin-Watson[157], que fournissent aussi les logiciels parmi les résultats de l'anova. La statistique de Durbin-Watson porte sur l'autocorrélation d'ordre 1 (c'est-à-dire l'autocorrélation dans laquelle la n-*ième* erreur engendre la (n+1)-*ième* erreur). La valeur de la statistique de Durbin-Watson s'établit toujours entre 0 et 4. Une valeur proche de 2 signifie qu'il n'y a pas d'autocorrélation des résidus. Les

---

[156] La problématique de l'autocorrélation se pose surtout dans l'analyse de séries temporelles, c'est-à-dire lorsque les observations se réfèrent à des périodes de temps différentes. C'est le cas par exemple si l'analyse porte sur les valeurs de variables au cours d'années successives. Dans ce cas, il est possible que les valeurs observées pour l'année N ne soient pas totalement indépendantes de celles observées pour l'année N-1, ou N-2, ou N-3, etc.

[157] La statistique de Durbin-Watson est définie par :

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{Tn} e_i^2}$$

où e_i est l'erreur (résidu) commise dans l'estimation de $\widehat{y_i}$, et n la taille de l'échantillon. Étant donné que le test vise à détecter l'existence d'erreurs imputables aux erreurs immédiatement précédentes, il va de soi que l'ordre dans lequel les données sont rangées importe : pour qu'une suspicion d'autocorrélation puisse être écartée ou confirmée, il importe que les données soient rangées dans l'ordre dans lequel elles sont suspectées de s'influencer.

valeurs proches de 0 indiquent une autocorrélation positive, et celles proches de 4 une autocorrélation négative. L'autocorrélation est positive si le fait qu'un résidu est positif a pour conséquence qu'un autre résidu sera également positif. L'autocorrélation est négative si le fait qu'un résidu est d'un signe donné implique qu'un autre résidu sera du signe contraire (un résidu positif rend négatif un autre résidu, et inversement). Il y a cependant des zones pour lesquels le test ne conclut pas. Pour une interprétation précise du DW, on utilise une table de Durbin-Watson (voir Annexe 5). L'Encadré 16.1 explique comment utiliser la table de Durbin-Watson.

ENCADRÉ 16.1 – UTILISATION DE LA TABLE DE DURBIN-WATSON

La table de Durbin-Watson se présente comme suit (une version plus complète figure en Annexe 5) :

	k'=1		k'=2		k'=3		…	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0,61	1,4	–	–	–	–	–	–
7	0,7	1,3	0,4	1,8	–	–	–	–
8	0,7	1,3	0,5	1,7	0,3	2,2	–	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

k' représente le nombre de variables indépendantes, par exemple 4 pour un modèle avec un seul facteur composé de 4 catégories, 9 pour un modèle à deux facteurs dont 4 catégories pour le premier facteur et 5 catégories pour le deuxième, etc. (on ne compte pas la constante α du modèle $\widehat{Y}$).

n est le « nombre d'observations », c'est-à-dire la taille de l'échantillon observé. C'est le nombre total d'individus (toutes catégories confondues) sur lesquels on dispose de données pour les variables indépendantes et la variable dépendante.

d_L , limite inférieure (*Lower limit*) et d_U, limite supérieure (*Upper limit*) sont les bornes qui permettent d'interpréter la valeur calculée de la statistique DW de Durbin-Watson.

Si $DW = 2$, il n'y a pas d'autocorrélation. Pour les autres valeurs de DW, on vérifie séparément l'autocorrélation positive et l'autocorrélation négative.

Vérification de l'autocorrélation positive

- Si $0 \leq DW < d_L$: il y a autocorrélation positive

- Si $DW > d_U$: il n'y a pas autocorrélation positive
- Si $d_L \leq DW \leq d_U$: il y a incertitude sur l'autocorrélation positive. On est dans une zone d'indétermination, on ignore s'il y a ou non autocorrélation positive.

Vérification de l'autocorrélation négative

- Si $4 - d_L < DW \leq 4$: il y a autocorrélation négative
- Si $DW < 4 - d_U$: il n'y a pas d'autocorrélation négative
- Si $4 - d_U \leq DW \leq 4 - d_L$: il y a incertitude sur l'autocorrélation négative. On est dans une zone d'indétermination, on ignore s'il y a ou non autocorrélation négative.

Par exemple, on obtient $DW = 2{,}6$ pour une anova à un facteur explicatif avec 3 catégories et échantillon de 8 individus. La table indique :
$d_L = 0{,}368$
$d_U = 2{,}287$

On peut calculer que :
$4 - d_U = 1{,}713$
$4 - d_L = 3{,}632$

Autocorrélation positive : $2{,}6 > d_U$. Il n'y a pas d'autocorrélation positive.

Autocorrélation négative : $4 - d_U < 2{,}6 < 4 - d_L$. Il y a incertitude sur l'autocorrélation négative.

Conclusion : Il n'y a pas d'autocorrélation positive d'ordre 1, et on ne peut pas conclure avec certitude sur l'autocorrélation négative.

16.5. CONDITIONS DE VALIDITÉ NON REMPLIES : RESTRUCTURER ET/OU REPARAMÉTRER LE MODÈLE

Il n'est pas rare que l'une ou l'autre des conditions relatives aux résidus ne soit pas remplie, ou même qu'aucune de ces conditions ne le soit. Il peut être alors utile de restructurer le modèle en y ajoutant des variables explicatives, et/ou de transformer les données, puis d'en recalculer les paramètres.

16.5.1. La condition de normalité des résidus n'est pas remplie

Si la condition de normalité n'est pas remplie, tout dépend du degré de non-normalité. Si l'écart par rapport à la normalité est faible, il n'est pas indispensable de

prendre des mesures car le modèle linéaire est assez robuste / insensible à un faible degré de non-normalité, surtout si l'échantillon est de grande taille. Si en revanche la distribution des résidus s'écarte fortement de la normalité, et si en outre l'échantillon est petit, la validité du modèle peut être remise en cause. Dans ce cas, une solution peut consister à retirer des échantillons les valeurs aberrantes, qui pourraient être à l'origine de la non-normalité. Une autre solution peut consister à procéder à une transformation des variables. Quant à la variable dépendante, une méthode de transformation courante d'une variable quantitative consiste à remplacer cette variable par son logarithme[158]. Il importe cependant de garder à l'esprit

[158] Il s'agit de remplacer chaque valeur $v_1, v_2, \cdots, v_n$ d'une variable V par son logarithme $log(v_1), log(v_2), \cdots, log(v_n)$. Cependant, étant donné que les logarithmes ne se calculent que pour les nombres strictement positifs, il importe de remplacer v_i par $(v_i + c)$ lorsque la série comporte des valeurs négatives. c est alors choisi de telle sorte que $(v_i + c)$ soit strictement positif même pour la valeur la plus fortement négative de v_i. On peut alors substituer à *V* sa transformée $log(V + c)$, dont les valeurs sont $log(v_1 + c), log(v_2 + c), \cdots, log(v_n + c)$. Bien d'autres méthodes de transformation existent, consistant notamment à remplacer v_i par :

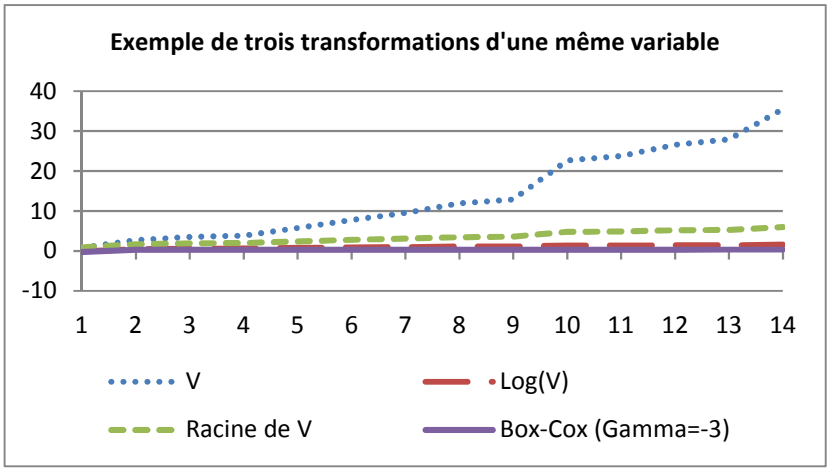
- $v_i^* = \sqrt{v_i + c}$; ou par
- $v_i^* = \frac{[(v_i)^\gamma]-1}{\gamma}$ $avec\ \gamma \neq 0$ (transformation Box-Cox).

qu'une transformation des données n'est pas neutre : contrairement à une simple standardisation, une transformation modifie la structure des données et peut par conséquent modifier la relation entre variables, alors que cette relation est l'objet même de l'analyse.

16.5.2. Situation d'hétéroscédasticité

La validité du modèle linéaire est très sensible à l'hétéroscédasticité. Une solution à l'hétéroscédasticité peut éventuellement consister à standardiser ou transformer les valeurs de la variable quantitative.

Par exemple, le graphique ci-dessous illustre trois transformations d'une même variable *V* :



16.5.3. La condition d'indépendance des résidus n'est pas remplie

Il y a dans ce cas distribution non-aléatoire des résidus, et/ou dépendance des résidus à la variable explicative, et/ou autocorrélation des résidus.

16.5.3.1. Distribution non-aléatoire des résidus et/ou dépendance des résidus par rapport à la variable explicative

Les résidus contiennent un élément explicatif (variable supplémentaire, interactions entre variables) qui aurait dû figurer parmi les variables indépendantes. Il faut donc restructurer le modèle en y intégrant cet élément explicatif comme variable explicative supplémentaire.

16.5.3.2. Autocorrélation des résidus

Il est possible, si un facteur d'autocorrélation est repéré, de rajouter ce facteur comme variable indépendante supplémentaire dans le modèle afin d'éliminer l'autocorrélation des résidus. On peut également intégrer au modèle une variable explicative additionnelle, jusque-là omise, mais dont on pense qu'elle pourrait expliquer au moins en partie la variable dépendante.

EXEMPLE 16.2.
Analyse de validité de modèle

On reprend les données de l'exemple 15.1 : afin de mesurer l'efficacité d'un enseignement, un test d'évaluation des connaissances a été administré à des étudiants avant le début puis après la fin de l'enseignement. Les résultats du test initial avaient permis de distinguer trois catégories d'étudiants : faibles, moyens, forts. Le Tableau 15.1 repris ci-dessous montre la différence entre la note au test initial et la note au test final pour chaque étudiant :

Tableau 15.1.
Effets de l'enseignement sur la performance aux tests d'évaluation

Étudiants	Catégorie d'appartenance (X)	Écart entre test initial et test final (Y)
Étudiant-e 1	Faible	4
Étudiant-e 2	Faible	5
Étudiant-e 3	Faible	9
Étudiant-e 4	Faible	6
Étudiant-e 5	Faible	6
Étudiant-e 6	Faible	8
Étudiant-e 7	Faible	3
Étudiant-e 8	Faible	-2
Étudiant-e 9	Faible	-3
Étudiant-e 10	Faible	11

Étudiants	Catégorie d'appartenance (X)	Écart entre test initial et test final (Y)
Étudiant-e 11	Faible	8
Étudiant-e 12	Faible	7
Étudiant-e 13	Faible	3
Étudiant-e 14	Faible	10
Étudiant-e 15	Faible	4
Étudiant-e 16	Moyen	3
Étudiant-e 17	Moyen	2
Étudiant-e 18	Moyen	2
Étudiant-e 19	Moyen	2
Étudiant-e 20	Moyen	-3
Étudiant-e 21	Moyen	4
Étudiant-e 22	Moyen	8
Étudiant-e 23	Moyen	3
Étudiant-e 24	Moyen	2
Étudiant-e 25	Moyen	6
Étudiant-e 26	Moyen	7
Étudiant-e 27	Moyen	8
Étudiant-e 28	Moyen	6
Étudiant-e 29	Moyen	8
Étudiant-e 30	Moyen	7
Étudiant-e 31	Fort	4
Étudiant-e 32	Fort	9
Étudiant-e 33	Fort	-5
Étudiant-e 34	Fort	10
Étudiant-e 35	Fort	-4

Étudiants	Catégorie d'appartenance (X)	Écart entre test initial et test final (Y)
Étudiant-e 36	Fort	3
Étudiant-e 37	Fort	-5
Étudiant-e 38	Fort	-2
Étudiant-e 39	Fort	-3
Étudiant-e 40	Fort	9
Étudiant-e 41	Fort	-4
Étudiant-e 42	Fort	5
Étudiant-e 43	Fort	3
Étudiant-e 44	Fort	6
Étudiant-e 45	Fort	-3
Progression moyenne de la catégorie Faible		*5,26*
Progression moyenne de la catégorie Moyen		*4,33*
Progression moyenne de la catégorie Fort		*1,53*
Progression moyenne de l'ensemble des 45 étudiants		*3,71*

On constate une progression de l'ensemble des notes en moyenne, mais la progression moyenne par catégorie est variable, la progression de la moyenne étant d'autant plus forte que la catégorie était initialement faible. L'analyse effectuée dans l'exemple 15.1 concluait à l'absence de différence significative entre catégories en matière de progression. On peut donc

écrire un modèle général, c'est-à-dire applicable aux trois catégories d'étudiants, et permettant de déterminer la progression à partir de la catégorie d'appartenance :

Progression = 3,71 + paramètre de la catégorie d'appartenance

où

- paramètre = 1,556 pour la catégorie Faible ;
- paramètre = 0,622 pour la catégorie Moyen ;
- paramètre = -2,178 pour la catégorie Fort.

On cherche à savoir si ce modèle est valide.

Indépendance des échantillons

Concernant tout d'abord la condition d'indépendance des catégories, les trois catégories étaient manifestement indépendantes puisqu'aucun étudiant ne pouvait appartenir à plus d'une catégorie à la fois.

Normalité des résidus

On teste ensuite la normalité des résidus. Les résidus et les résidus standardisés s'établissent comme suit :

Observation	Résidu	Résidu standardisé
Étudiant-e 1	-1,267	-0,292
Étudiant-e 2	-0,267	-0,062
Étudiant-e 3	3,733	0,862
Étudiant-e 4	0,733	0,169

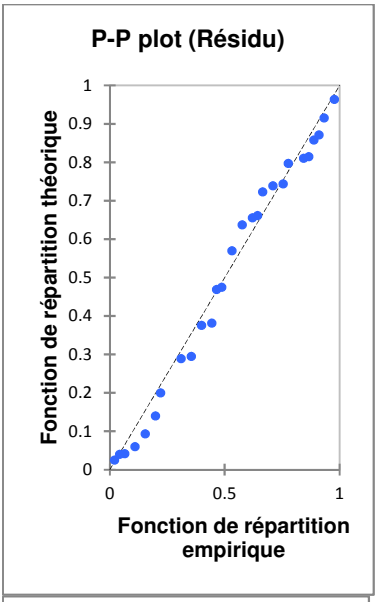
Observation	Résidu	Résidu standardisé
Étudiant-e 5	0,733	0,169
Étudiant-e 6	2,733	0,631
Étudiant-e 7	-2,267	-0,523
Étudiant-e 8	-7,267	-1,678
Étudiant-e 9	-8,267	-1,908
Étudiant-e 10	5,733	1,324
Étudiant-e 11	2,733	0,631
Étudiant-e 12	1,733	0,400
Étudiant-e 13	-2,267	-0,523
Étudiant-e 14	4,733	1,093
Étudiant-e 15	-1,267	-0,292
Étudiant-e 16	-1,333	-0,308
Étudiant-e 17	-2,333	-0,539
Étudiant-e 18	-2,333	-0,539
Étudiant-e 19	-2,333	-0,539
Étudiant-e 20	-7,333	-1,693
Étudiant-e 21	-0,333	-0,077
Étudiant-e 22	3,667	0,847
Étudiant-e 23	-1,333	-0,308
Étudiant-e 24	-2,333	-0,539
Étudiant-e 25	1,667	0,385
Étudiant-e 26	2,667	0,616
Étudiant-e 27	3,667	0,847
Étudiant-e 28	1,667	0,385
Étudiant-e 29	3,667	0,847
Étudiant-e 30	2,667	0,616
Étudiant-e 31	2,467	0,569

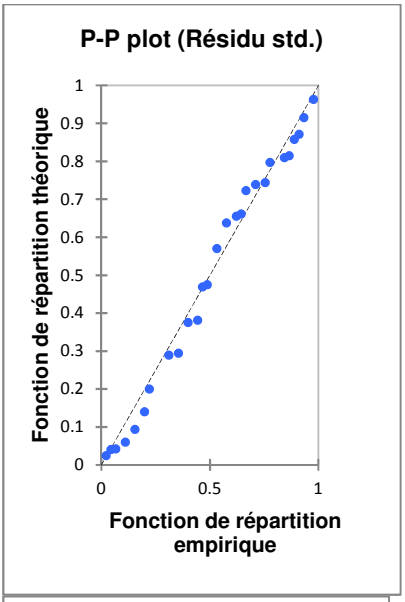
Observation	Résidu	Résidu standardisé
Étudiant-e 32	7,467	1,724
Étudiant-e 33	-6,533	-1,508
Étudiant-e 34	8,467	1,955
Étudiant-e 35	-5,533	-1,277
Étudiant-e 36	1,467	0,339
Étudiant-e 37	-6,533	-1,508
Étudiant-e 38	-3,533	-0,816
Étudiant-e 39	-4,533	-1,047
Étudiant-e 40	7,467	1,724
Étudiant-e 41	-5,533	-1,277
Étudiant-e 42	3,467	0,800
Étudiant-e 43	1,467	0,339
Étudiant-e 44	4,467	1,031
Étudiant-e 45	-4,533	-1,047

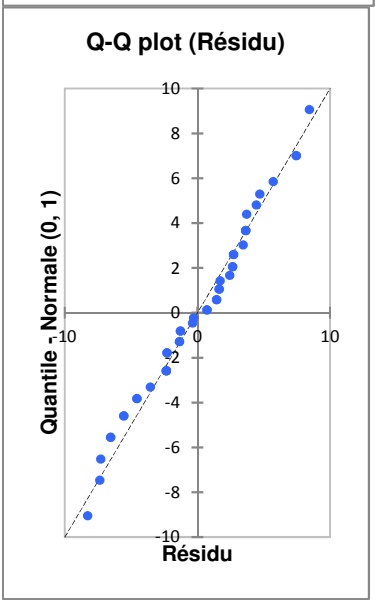
Les tests de normalité indiquent que les résidus sont normalement distribués :

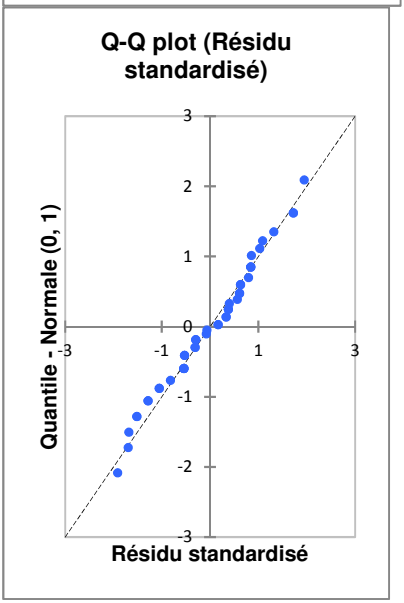
	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Résidu	0,409	0,382	0,281	0,592
Résidu standardisé	0,409	0,382	0,281	0,592

Ce qu'illustrent les diagrammes PP et QQ :


P-P plot (Résidu)
Fonction de répartition théorique
Fonction de répartition empirique
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1

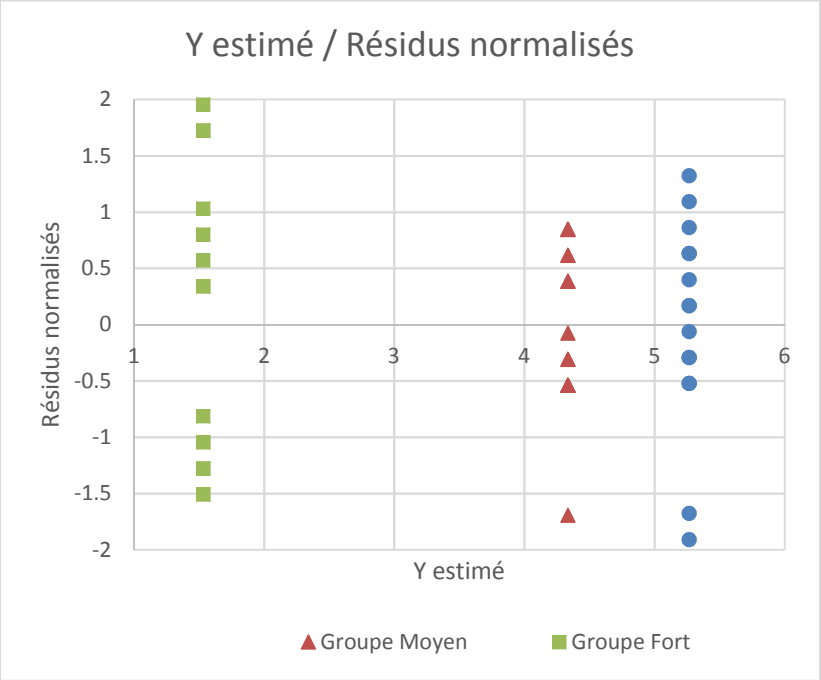

P-P plot (Résidu std.)
Fonction de répartition théorique
Fonction de répartition empirique
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1


Q-Q plot (Résidu)
Quantile - Normale (0, 1)
Résidu
10
8
6
4
2
0
-2
-4
-6
-8
-10
-10
0
10


Q-Q plot (Résidu standardisé)
Quantile - Normale (0, 1)
Résidu standardisé
3
2
1
0
-1
-2
-3
-3
-1
1
3

Homoscédasticité

On examine ensuite la stabilité de la dispersion des résidus sur le graphique $\widehat{Y}$/résidu :



Le graphique suggère que la dispersion des résidus diffère suivant le niveau estimé de progression scolaire (qui lui-même dépend du groupe). Pour confirmation, on teste l'homogénéité des variances des résidus des trois groupes d'étudiants :

	p	
	Levene	Bartlett
Comparaison simultanée des trois groupes	0,012	0,118

La comparaison simultanée des trois groupes ne permet pas de trancher puisque chacun des deux tests conduit à une conclusion différente. On effectue donc une comparaison des groupes deux-à-deux :

	p		
	Fisher	Levene	Bartlett
Comparaison Groupe Faible / Groupe Moyen	0,390	0,576	0,390
Comparaison Groupe Moyen / Groupe Fort	0,045	0,003	0,045
Comparaison Groupe Faible / Groupe Fort	0,238	0,037	0,238

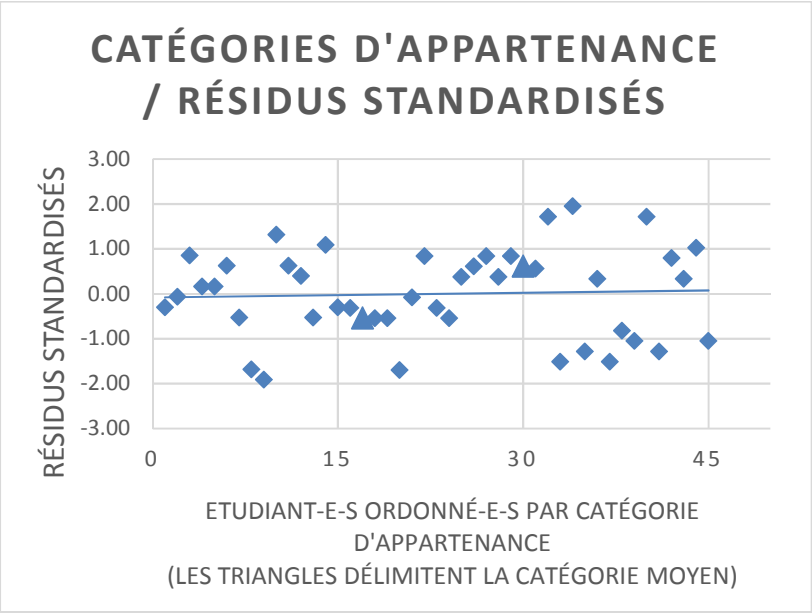
Les trois tests indiquent une différence de variances entre les résidus du groupe Moyen et les résidus du groupe Fort. La condition d'homoscédasticité n'est pas remplie.

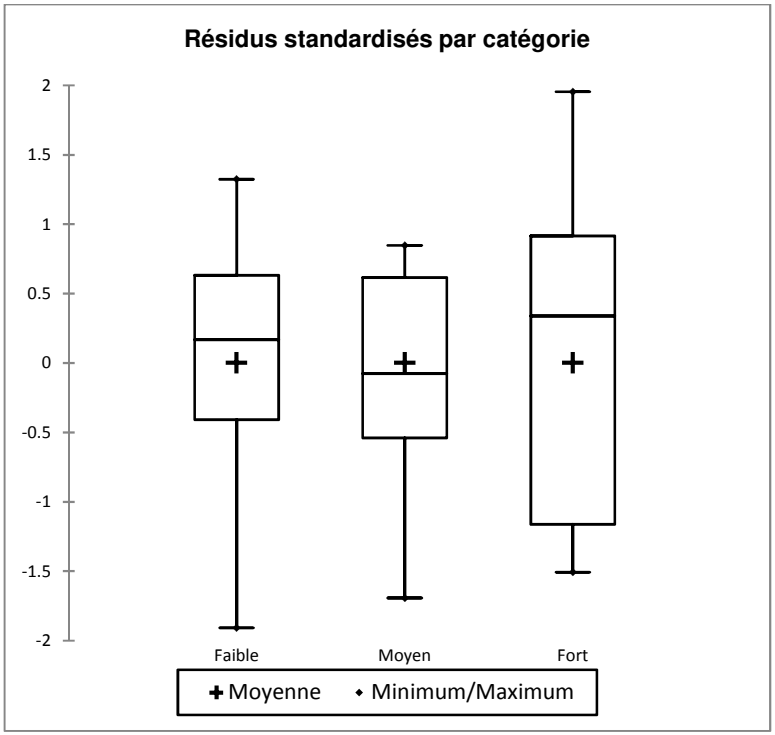
Distribution aléatoire des résidus

On calcule le coefficient de corrélation linéaire *r* entre résidus et variable dépendante estimée : $r<0{,}0001$, avec p-value=1. La condition de distribution aléatoire des résidus est donc remplie.

Indépendance par rapport à la variable indépendante

Le graphique Variable indépendante / Résidu et la boite à moustaches montrent que les niveaux de résidu peuvent être plus élevés pour le groupe fort (la droite de tendance est croissante) :





Donc on ne peut pas exclure une dépendance entre la catégorie d'appartenance et le niveau du résidu. Par conséquent, on peut considérer que la condition d'indépendance des résidus par rapport à la variable indépendante n'est pas remplie.

Autocorrélation des résidus

Enfin, la valeur calculée de la statistique de Durbin-Watson s'élève à 2,26 donc est proche de 2 ce qui laisse penser qu'il n'y a pas d'autocorrélation des résidus.

Effectivement, la table de Durbin-Watson indique que pour $k' = 3$ et $n = 45$:

- $d_L = 1{,}383$
- $d_U = 1{,}666$

Donc :

- $4 - d_U = 2{,}33$
- $4 - d_L = 2{,}61$
- Il n'y a pas d'autocorrélation positive car $2{,}26 > d_U$.
- Il n'y a pas d'autocorrélation négative car $2{,}26 < 4 - d_U$.

En conclusion, les conditions d'homoscédasticité et d'indépendance à la variable indépendante ne sont pas remplies. Le modèle déterminant la progression à partir de la catégorie d'appartenance ne peut pas être considéré comme valide à ce stade.

EXEMPLE 16.3.
Analyse de validité de modèle

On reprend les données de l'exemple 15.2. Le Tableau 15.2 repris ci-dessous montre la répartition des effectifs d'enseignants dans le premier degré public par académie et département :

Tableau 15.2.
Effectifs d'enseignants dans le premier degré public par Académie et Département pour vingt Académies métropolitaines – 2013-2014

Source : MENESR (2014), Repères et références statistiques sur les enseignements, la formation et la recherche, page 299
http://www.enseignementsup-recherche.gouv.fr/cid81900/reperes-et-references-statistiques-edition-septembre-2014.html

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Alpes-Haute-Provence	Aix-Marseille	917
Hautes-Alpes	Aix-Marseille	770
Bouches-du-Rhône	Aix-Marseille	9 618
Vaucluse	Aix-Marseille	2 839
Aisne	Amiens	2 918
Oise	Amiens	4 648
Somme	Amiens	2 817
Doubs	Besançon	2 932
Jura	Besançon	1 316
Haute-Saône	Besançon	1 321
Territoire de Belfort	Besançon	734
Dordogne	Bordeaux	1 776
Gironde	Bordeaux	6 930

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Landes	Bordeaux	1 818
Lot-et-Garonne	Bordeaux	1 514
Pyrénées-Atlantiques	Bordeaux	2 631
Calvados	Caen	3 308
Manche	Caen	2 175
Orne	Caen	1 276
Allier	Clermont-Ferrand	1 624
Cantal	Clermont-Ferrand	761
Haute-Loire	Clermont-Ferrand	959
Puy-de-Dôme	Clermont-Ferrand	2 864
Corse-du-Sud	Corse	656
Haute-Corse	Corse	782
Seine-et-Marne	Créteil	8 009
Seine-Saint-Denis	Créteil	9 838
Val-de-Marne	Créteil	6 933
Côte-d'Or	Dijon	2 624
Nièvre	Dijon	1 033
Saône-et-Loire	Dijon	2 742
Yonne	Dijon	1 815

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Ardèche	Grenoble	1 379
Drôme	Grenoble	2 534
Isère	Grenoble	6 425
Savoie	Grenoble	2 086
Haute-Savoie	Grenoble	3 795
Nord	Lille	12 946
Pas-de-Calais	Lille	7 926
Ain	Lyon	3 272
Loire	Lyon	3 383
Rhône	Lyon	8 436
Aude	Montpellier	1 703
Gard	Montpellier	3 538
Hérault	Montpellier	4 999
Lozère	Montpellier	426
Pyrénées-Orientales	Montpellier	2 272
Meurthe-et-Moselle	Nancy-Metz	3 640
Meuse	Nancy-Metz	1 064
Moselle	Nancy-Metz	5 381
Vosges	Nancy-Metz	2 029
Loire-Atlantique	Nantes	5 087
Maine-et-Loire	Nantes	2 962
Mayenne	Nantes	1 230

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Sarthe	Nantes	2 723
Vendée	Nantes	1 815
Cher	Orléans-Tours	1 460
Eure-et-Loir	Orléans-Tours	2 384
Indre	Orléans-Tours	1 035
Indre-et-Loire	Orléans-Tours	2 722
Loir-et-Cher	Orléans-Tours	1 616
Loiret	Orléans-Tours	3 419
Charente	Poitiers	1 590
Charente-Maritime	Poitiers	2 832
Deux-Sèvres	Poitiers	1 664
Vienne	Poitiers	2 017
Ardennes	Reims	1 641
Aube	Reims	1 614
Marne	Reims	2 875
Haute-Marne	Reims	1 056
Côtes-d'Armor	Rennes	2 267
Finistère	Rennes	3 092
Ille-et-Vilaine	Rennes	3 743
Morbihan	Rennes	2 134
Ariège	Toulouse	735
Aveyron	Toulouse	1 137

Département	Académie d'appartenance (X)	Effectifs d'enseignants (Y)
Haute-Garonne	Toulouse	5 999
Gers	Toulouse	874
Lot	Toulouse	775
Hautes-Pyrénées	Toulouse	1 044
Tarn	Toulouse	1 653
Tarn-et-Garonne	Toulouse	1 304

L'analyse menée dans l'exemple 15.2 montrait que l'effectif d'enseignants par Département diffère significativement d'une académie à l'autre. En tenant compte de ces différences, le modèle général déterminant l'effectif d'enseignants en fonction de l'académie s'écrit :

Effectif d'enseignants = 2885,744 + paramètre de l'académie

où les paramètres des académies s'établissent comme suit :

Aix-Marseille	650,256
Amiens	575,256
Besançon	-1309,994
Bordeaux	48,056

Caen	-632,744
Clermont-Ferrand	-1333,744
Corse	-2166,744
Créteil	5374,256
Dijon	-832,244
Grenoble	358,056
Lille	7550,256
Lyon	2144,589
Montpellier	-298,144
Nancy-Metz	142,756
Nantes	-122,344
Orléans-Tours	-779,744
Poitiers	-859,994
Reims	-1089,244
Rennes	-76,744
Toulouse	-1195,619

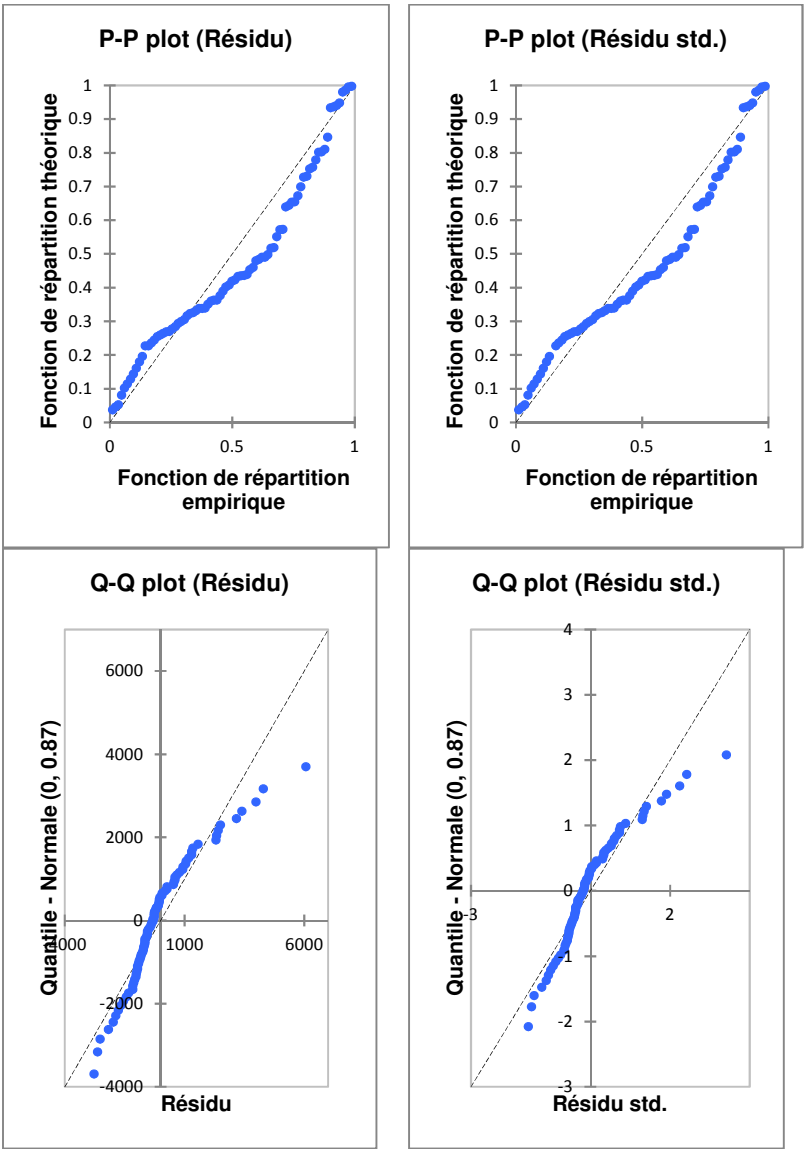
On cherche à savoir si ce modèle est valide.

Indépendance des échantillons

La condition est remplie puisque chaque département ne relève que d'une seule académie.

Normalité des résidus

Les diagrammes PP et QQ suggèrent que les résidus ne sont pas normalement distribués :



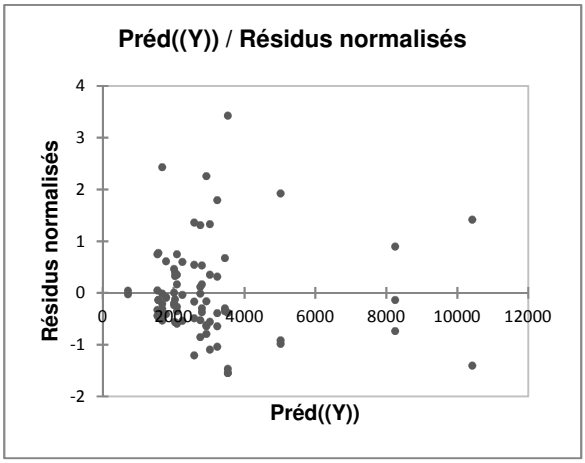
ce que confirment les tests de normalité :

	p			
	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Résidu standardisés	< 0,0001	< 0,0001	< 0,0001	< 0,0001

Il est clair qu'au seuil de significativité de 5%, la condition de normalité n'est pas remplie.

Homoscédasticité

Le graphique $\widehat{Y}$ / Résidus standardisés se présente comme suit :



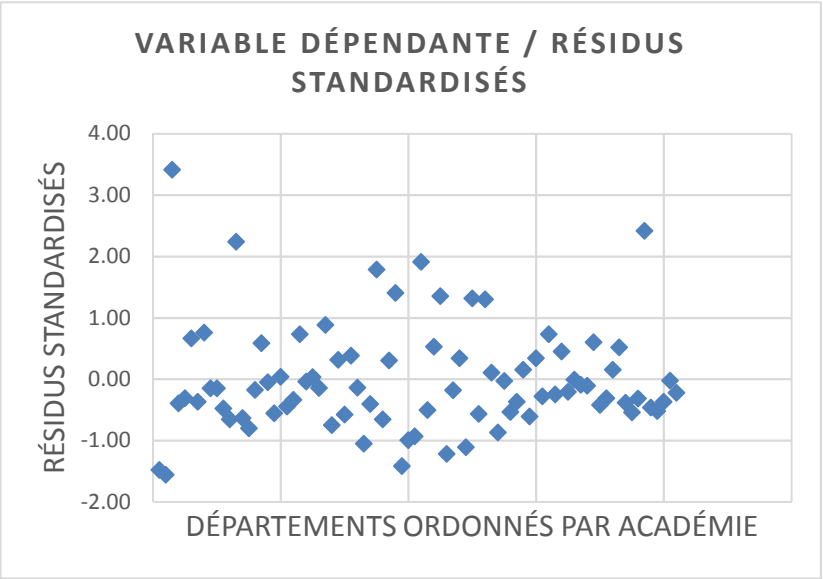
Il n'y a pas constance de la dispersion des résidus tout au long des valeurs que prend la variable dépendante prédite. La condition d'homoscédasticité n'est pas remplie elle non plus.

Distribution aléatoire des résidus

Le calcul du coefficient de corrélation de Pearson ($r < 0{,}0001$, $p = 1$) entre variable dépendante prédite et résidus standardisés conduit à conclure que la condition de distribution aléatoire des résidus est remplie.

Indépendance des résidus par rapport à la variable indépendante

Le graphique variable indépendante / Résidu montre que les niveaux de résidus dépendent des académies :



La condition d'indépendance des résidus à la variable indépendante n'est donc pas remplie.

Autocorrélation des résidus

La valeur calculée de la statistique de Durbin-Watson s'élève à 2,56. La table de Durbin-Watson indique que pour $k' = 20\ et\ n = 82$:

- $d_L = 1{,}08$
- $d_U = 2{,}26$

Donc :

- $4 - d_U = 1{,}74$
- $4 - d_L = 2{,}92$

On peut conclure que :

- Il n'y a pas d'autocorrélation positive car $2{,}56 > d_U$;
- Il y a incertitude sur l'autocorrélation négative car $4 - d_U < 2{,}56 < 4 - d_L$.

Conclusion

Mises à part les conditions d'indépendance des échantillons et d'indépendance des résidus par rapport à la variable dépendante, aucune des conditions de validité du modèle n'est remplie.

Chapitre 17. Régression linéaire simple : modéliser une relation de dépendance entre deux variables quantitatives

Soit un échantillon de n individus $i = 1, 2, \cdots, n$. On dispose de données caractérisant ces individus au regard de deux variables quantitatives X et Y. On suppose que les variables X et Y sont liées par une relation de dépendance, et que cette relation est de type linéaire. La *régression linéaire simple*[159] permet de modéliser la relation, d'estimer les paramètres du modèle, d'en tester la significativité, d'identifier le sens (positif ou négatif) de l'effet qu'exerce la variable indépendante sur la variable dépendante, de caractériser la portée explicative du modèle et, sur la base de ce dernier, de tenter d'éclairer l'avenir par la prévision.

17.1. MODÉLISER LA RELATION

La régression linéaire simple postule que la relation entre X et Y est de la forme

$$Y = \alpha + (\beta \times X)$$

où

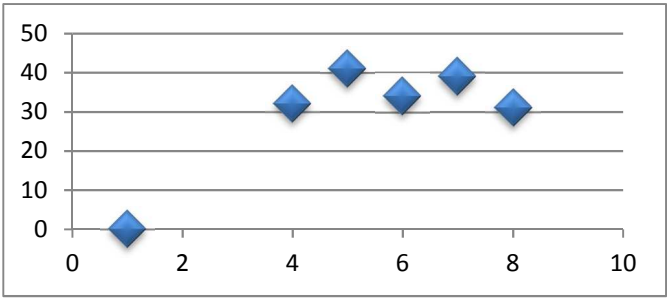
[159] Accessible dans XLSTAT via la commande *Modélisation des données / Régression linéaire*.

- X est la variable explicative (on dit encore « variable indépendante » ou *« régresseur »* ou « *prédicteur* ») ;
- Y, la variable expliquée ou variable dépendante ;
- α, une constante, qui indique la valeur de Y lorsque $X = 0$;
- β, le « coefficient de régression ».

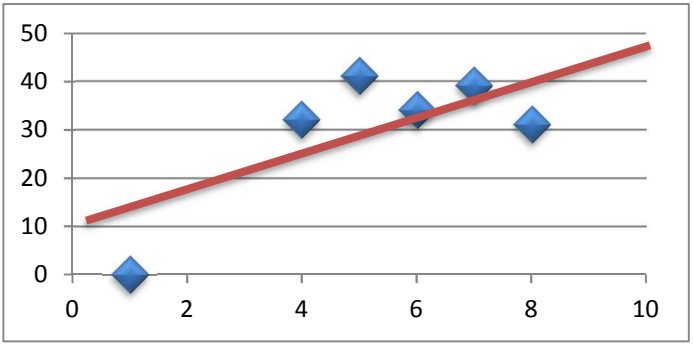
Pour que le modèle soit opérationnel, il faut calculer les paramètres α et β. La méthode de base du calcul des paramètres est la méthode des « *moindres carrés ordinaires* » (Encadré 17.1).

ENCADRÉ 17.1 – ESTIMATION DES PARAMÈTRES DE RÉGRESSION

La méthode de base (méthode des moindres carrés ordinaires – MCO) repose sur le principe de minimisation des écarts à la droite de régression. L'idée de départ est que la relation entre X et Y est linéaire, donc représentable par un nuage de points dont la forme d'ensemble est de type linéaire, par exemple :

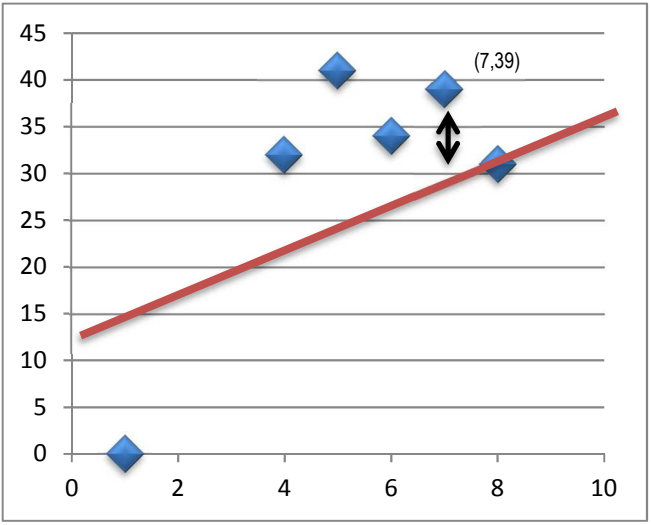


On peut donc représenter ce nuage par une droite "moyenne" qui passerait au plus près de l'ensemble de ses points :



On appelle cette droite "droite de régression". Pour qu'elle passe au plus près des points du nuage, il faut minimiser, pour chaque point du nuage, l'écart entre les coordonnées de ce point et la droite. L'écart se mesure par rapport à l'ordonnée. Il est par exemple de 10 (soit 39-29)

pour le point de coordonnées $x = 7, y = 39$ ci-dessous :



L'équation de la droite de régression s'écrit :

$$\hat{y}_i = \alpha + \beta x_i$$

où
$\hat{y}_i$ est l'ordonnée du point d'abscisse x_i ;
α, l'ordonnée à l'origine ;
β, la pente.

L'écart entre un point i du nuage et la droite de régression s'écrit :

$$y_i - \hat{y}_i$$

Ce qui équivaut à

$$y_i - (\alpha + \beta x_i)$$

Minimiser tous les écarts entre les n points $i = 1, 2, \cdots, N$ du nuage et la droite de régression revient à minimiser l'expression

$$\sum_{i=1}^{N} [y_i - (\alpha + \beta x_i)]^2 \qquad (1)$$

Il s'agit donc de minimiser les carrés des écarts à la droite de régression (et non pas directement les écarts eux-mêmes car ces écarts peuvent être positifs ou négatifs et se compenseraient).

Il est possible de calculer les valeurs des paramètres α et β qui permettent de minimiser l'expression (1). Les valeurs de α et β sont données par :

$$\begin{cases} \alpha = \bar{y} - \beta \bar{x} \\ \beta = \dfrac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \end{cases}$$

où
$\bar{x}$ est la moyenne des x_i ; et
$\bar{y}$, la moyenne des y_i.

On peut observer que β peut encore s'écrire

$$\beta = \frac{cov_{XY}}{\sigma_X^2}$$

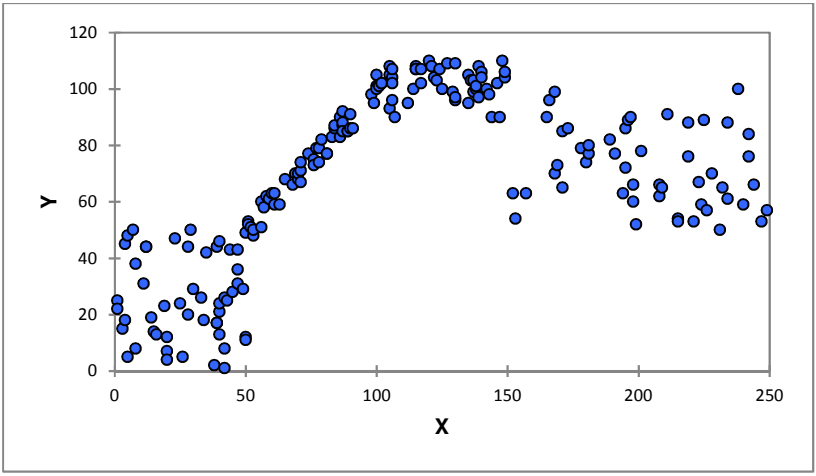
c'est-à-dire comme le rapport entre covariance de X et Y au numérateur, et variance de X au dénominateur, puisque :

$$cov_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$et \quad \sigma_X^2 = \frac{\sum_{i=1}^{N}(x - \bar{x})^2}{n}$$

Il importe d'observer que l'application des méthodes de régression linéaire n'exige pas que la relation analysée soit linéaire sur la totalité de son ensemble de définition. Une relation peut être linéaire sur certains seulement de ses segments, comme le montre par exemple le Graphique 17.1 :

Graphique 17.1.
Exemple de relation linéaire par segments



Dans cet exemple, la relation est définie sur l'ensemble des x compris entre 0 et 250. Sur cet ensemble de définition, la relation n'est clairement pas linéaire. Cependant, il est tout-à-fait possible de rechercher un modèle linéaire sur l'intervalle [50 ; 100] de la variable *X*, et un autre sur l'intervalle [100 ; 150] de *X*. Les deux modèles seront différents (c'est-à-dire caractérisés par des paramètres différents), mais chacun pourra rendre compte de façon précise des relations entre *X* et *Y* sur son segment spécifique de pertinence. En ce sens, les méthodes de régression linéaire sont utiles aussi pour analyser, au moins en partie, des relations non-linéaires.

17.1.1. Estimation des paramètres et opérationnalisation du modèle

Les logiciels statistiques fournissent des estimations ($\widehat{\alpha}\ et\ \widehat{\beta}$) des valeurs des paramètres. Par exemple, à partir de données sur les moyennes obtenues l'an passé par vingt élèves au dernier bac blanc et des moyennes effectivement obtenues par ces élèves au bac (Tableau 17.1), on obtient les paramètres de régression présentés au Tableau 17.2 :

Tableau 17.1.
Moyennes de vingt élèves au dernier bac blanc et au bac

Identifiants des élèves	Moyennes au dernier bac blanc	Moyennes au bac
1	4	6
2	8	11
3	2	10
4	5	6
5	10	12
6	4	5
7	6	5
8	8	2
9	15	18
10	6	7
11	17	19
12	5	8
13	1	3

Identifiants des élèves	Moyennes au dernier bac blanc	Moyennes au bac
14	6	3
15	10	12
16	7	9
17	5	8
18	5	6
19	1	3
20	6	8

Tableau 17.2.
Paramètres de régression

Libellés des paramètres	VP*	ES	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Constante (α)	1,94	1,18	1,64	0,117	-0,54	4,43
Coefficient de régression (β)	0,93	0,15	6,02	<0,0001	0,60	1,25

* VP : Valeur des paramètres
** ES : Erreur standard

Le tableau des paramètres s'interprète comme indiqué page 469. Connaissant les paramètres estimés, le modèle opérationnel s'écrit, de façon générale :

$$\hat{y}_i = \hat{\alpha} + (\hat{\beta} \times x_i)$$

où

- $\hat{\alpha}$ et $\hat{\beta}$ sont les paramètres estimés ;
- x_i est la valeur de la variable indépendante pour l'individu *i* ; et
- $\hat{y}_i$, la valeur estimée de la variable dépendante pour l'individu *i* ;

soit, dans cet exemple :

$$\hat{y}_i = 1{,}947 + (0{,}932 \times x_i)$$

où

- x_i est la moyenne obtenue par l'élève *i* au dernier bac blanc ; et
- $\hat{y}_i$, l'estimation de la moyenne effectivement obtenue par l'élève *i* au bac.

L'estimation implique évidemment un risque de commettre une erreur par rapport à la vraie valeur y_i de *Y* pour l'individu *i* :

$$y_i = \hat{y}_i + \varepsilon$$

où ε est le terme d'erreur.

La question est alors de déterminer la fiabilité et la validité du modèle.

17.1.2. Fiabilité et validité du modèle

Le modèle est fiable si ses paramètres sont significatifs. Dans l'exemple utilisé ici, au seuil de significativité de 5%, le coefficient de régression est significatif mais pas la constante. Le modèle n'est donc qu'en partie fiable.

Le modèle est valide si les conditions de normalité des résidus, homoscédasticité et indépendance des résidus (distribution aléatoire, indépendance à la variable indépendante, non-autocorrélation)[160] sont remplies.

La condition de normalité des résidus se vérifie au moyen des tests de normalité usuels, ou le cas échéant, au moyen de graphiques tels que l'histogramme des résidus standardisés et les diagrammes PP et QQ.

La condition d'homoscédasticité se vérifie au moyen du graphique « $\widehat{Y}$/Résidu » ou des tests de Breusch-Pagan et White.

La condition de distribution aléatoire des résidus se vérifie en calculant la corrélation entre résidus et variable dépendante estimée. La vérification peut s'effectuer aussi au moyen du graphique « $\widehat{Y}$/Résidu ».

[160] Voir chapitre 16.

On peut vérifier l'indépendance des résidus par rapport à la variable indépendante en calculant les corrélations entre les deux, ou sur base de l'examen du nuage de points "Variable indépendante / Résidu".

La condition de non-autocorrélation des résidus se vérifie généralement au moyen de la statistique de Durbin-Watson. Cependant, la statistique de Durbin-Watson ne teste que l'autocorrélation d'ordre 1. Un outil plus général – quoique moins fréquemment utilisé et plus rarement disponible sur les logiciels statistiques[161] – est le test de Breusch-Godfrey. Le test de Breusch-Godfrey détecte les autocorrélations d'ordre 1 ou plus : autocorrélations d'ordre 2, 3, 4, etc.

---

[161] Le test de Breusch-Godfrey est cependant disponible dans le « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9) via la fonction BGTEST. Supposons par exemple que l'on veuille tester la présence d'autocorrélation d'ordre λ parmi les résidus d'une régression linéaire. Il y a 30 observations dans l'échantillon. On inscrit les valeurs de la variable indépendante (X) dans la plage A1:A30 de la feuille Excel, puis les valeurs de la variable dépendante (Y) dans la plage B1:B30. Puis on sélectionne la cellule C1 et on inscrit dans la fenêtre la formule

=BGTEST(A1:A30,B1:B30,λ)

où λ est l'ordre d'autocorrélation auquel on s'intéresse (λ=1, ou 2, ou 3, etc.).

Enfin on appuie sur la touche Entrée. Dans la cellule C1 s'affiche la p-value du test de Breusch-Godfrey. Il y a autocorrélation de l'ordre examiné si la p-value est inférieure à 0,05. Il n'y a pas autocorrélation de l'ordre examiné sinon.

Il est aussi plus directement interprétable que le test de Durbin Watson.

Lorsqu'une de ces conditions n'est pas remplie, il peut parfois y être remédié en restructurant le modèle par ajout de variables explicatives supplémentaires, et/ou par standardisation ou transformation des données.

17.2. PORTÉE EXPLICATIVE DU MODÈLE

La question ici est de savoir dans quelle mesure le modèle permet d'expliquer *Y*. En d'autres termes, le modèle permet-il d'appréhender la totalité ou au moins une part majeure du phénomène Y que l'on tente d'expliquer ?

La mesure de la portée explicative du modèle est fournie par le *coefficient de détermination*, noté R^2. Il exprime la proportion de la variance de Y qui est « expliquée par » – c'est-à-dire « imputable à » – la variance de X. Sa valeur est comprise entre 0 et 1. Plus élevé est le R^2, plus élevée est la part de la variable dépendante qui est expliquée par la variable indépendante, et plus grande est donc la portée explicative du modèle dans l'analyse du phénomène étudié.

Il importe de noter que le coefficient de détermination est sensible au nombre de variables explicatives : le R^2 peut augmenter quand on introduit dans le modèle des variables explicatives supplémentaires, même si ces dernières ne sont pas significatives. Une mesure plus exacte de la portée explicative est donc le « R^2 ajusté », dont le calcul tient compte du nombre de variables indépendantes et du nombre d'observations.

17.3. FONCTION D'ANALYSE ET PRÉVISION

On reprend les paramètres présentés dans le Tableau 17.2 :

Tableau 17.2.
Paramètres de régression

Libellés des paramètres	VP*	ES	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Constante (α)	1,94	1,18	1,64	0,117	-0,54	4,43
Coefficient de régression (β)	0,93	0,15	6,02	<0,0001	0,60	1,25

* VP : Valeur des paramètres
** ES : Erreur standard

La constante, tout d'abord, renseigne sur le niveau plancher du phénomène expliqué. Elle indique la valeur de $\hat{y}_i$ lorsque $x_i = 0$. Dans l'exemple, la constante indique que, en règle générale, 1,94 points de la moyenne au bac ne dépendent pas de la moyenne au dernier bac blanc. Évidemment, l'observation ne vaut que si la constante est significative, et s'il y a du sens à ce que le modèle comporte une constante.

En effet, il n'y a parfois aucun sens à avoir une constante dans le modèle. C'est le cas lorsque la valeur de la variable dépendante est nécessairement nulle si la valeur de la variable indépendante est nulle. Par exemple, dans un modèle linéaire expliquant le taux de réussite au bac par le nombre de candidats présents issus de l'établissement, il est clair que si aucun des candidats venant d'un établissement n'est présent aux épreuves, le taux de réussite sera nul. Le modèle dans ce cas ne devrait pas comporter de constante, donc le taux de réussite y_i de l'établissement devrait dépendre uniquement du nombre de présents x_y :

$$y_i = \beta \times x_i$$

Le coefficient de régression β, ensuite, mesure l'effet qu'une variation de x_i exerce sur y_i, et indique si cet effet est positif ou négatif : de façon générale, quand x_i augmente d'une unité, y_i varie de β unité(s), et cette variation est positive (augmentation) si le signe de β est positif, et négative (diminution) sinon. Dans l'exemple,

si la moyenne au dernier bac blanc augmente d'un point, la moyenne au bac augmente de 0,93 points. Le coefficient étant significatif, cette partie du modèle pourrait, dans le cadre de l'exemple, être considérée comme fiable.

Enfin, la modélisation permet la prévision. Si les paramètres du modèle sont significatifs, si les conditions de validité du modèle sont remplies, et si la portée explicative du modèle est suffisante, on peut alors considérer que, au moins dans un futur proche, si la nature linéaire de la relation ne change pas, et si les paramètres α et β sont stables, il suffira d'avoir une bonne estimation de la valeur future de x_i pour pouvoir prévoir de façon fiable la valeur future de y_i. Cette condition de stabilité (au moins à court terme) de la linéarité de la relation et des paramètres de régression est habituellement synthétisée par l'expression « toutes choses égales par ailleurs ». L'ensemble de ces conditions (significativité, validité, portée et stabilité) souligne le luxe de précautions dont doit s'entourer l'exercice de prévision.

EXEMPLE 17.1

On dispose des résultats au Brevet professionnel par académie en 2014 (Tableau 17.3). On cherche à savoir si le nombre de présents aux épreuves détermine le taux de réussite de l'académie.

Tableau 17.3.
Présents et pourcentage d'admis au Brevet professionnel par académie en 2014

Source : MENESR (2014), Repères et références statistiques sur les enseignements, la formation et la recherche, Tableau 8.8, page 241.
http://www.education.gouv.fr/cid57096/reperes-et-references-statistiques.html#Données_publiques

	Brevet professionnel	
	Présents	% admis
Aix-Marseille	3 161	72,5
Amiens	1 646	81,2
Besançon	1 360	80,4
Bordeaux	3 890	78,2
Caen	1 992	79,1
Clermont-Ferrand	1 592	76,3
Corse	192	84,4
Créteil	2 742	74,0
Dijon	1 659	80,2
Grenoble	4 230	78,3
Lille	5 002	79,7
Limoges	886	77,8
Lyon	3 278	84,7
Montpellier	3 256	71,6
Nancy-Metz	2 306	80,4
Nantes	6 018	77,7
Nice	1 773	70,2
Orléans-Tours	2 482	80,1

	Brevet professionnel	
	Présents	% admis
Paris	900	71,6
Poitiers	2 877	77,7
Reims	1 803	82,0
Rennes	4 709	81,9
Rouen	1 734	83,2
Strasbourg	1 314	77,5
Toulouse	2 987	81,7
Versailles	3 398	80,1
Guadeloupe	738	66,3
Guyane	400	68,5
Martinique	324	69,1
Mayotte	111	57,7
La Réunion	1 967	81,5
Ensemble	*70 727*	*78,4*

On examine d'abord la significativité des paramètres de régression et la validité du modèle, puis on synthétise les résultats.

1. Significativité des paramètres de régression

Pour des raisons évidentes, on teste un modèle sans constante. Les paramètres de régression s'établissent comme suit :

	VP*	ES**	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Constante	0,000					
Présents	0,024	0,003	9,088	< 0,0001	0,019	0,030

* VP : Valeur des paramètres
** ES : Erreur standard

Le coefficient de la variable indépendante est significatif et positif : plus le nombre de présents aux épreuves du Brevet professionnel est élevé, plus le taux de réussite de l'académie à cet examen est élevé. Un présent de plus aux épreuves signifie un taux de réussite plus élevé de 0,024 points de pourcentage.

2. Vérification des conditions de validité du modèle

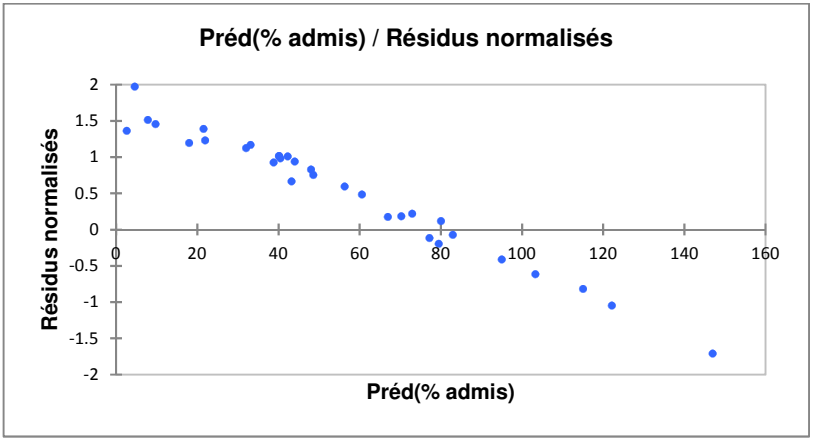
- Normalité des résidus

Les tests de normalité montrent que les résidus standardisés suivent une distribution gaussienne :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Résidu standardisé	0,164	0,123	0,173	0,244

- Homoscédasticité

Le graphique « $\widehat{Y}$/Résidu » montre que la condition d'homoscédasticité n'est pas remplie (la largeur de la bande décroissante n'est pas stable, mais tend au contraire à s'amenuiser quand les valeurs prédites augmentent)[162] :



[162] Ce que confirme le test de White, exécuté sur le logiciel STATA :

estat imtest, white
White's test for Ho: homoskedasticity
against Ha: unrestricted heteroskedasticity
chi2(2) = 26.98
Prob > chi2 = 0.0000

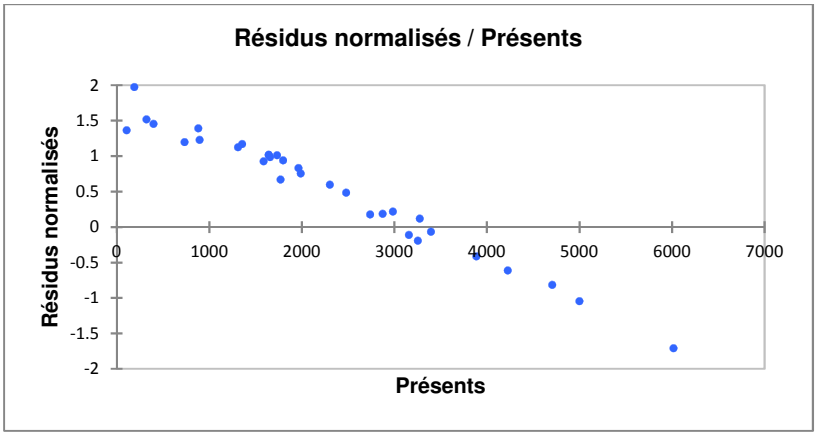
Source	*chi2*	*df*	*p*
Heteroskedasticity	*26.98*	*2*	*0.0000*

- Distribution aléatoire des résidus

La condition de distribution aléatoire des résidus n'est pas remplie : le graphique « $\widehat{Y}$/Résidu » ci-dessus montre clairement une corrélation (négative) entre résidus et prédictions.

- Indépendance des résidus par rapport à la variable indépendante

La condition d'indépendance par rapport à X n'est pas remplie elle non plus, comme le montre la corrélation entre résidus et variable indépendante :



- Non-autocorrélation des résidus

Il n'y a pas de raison particulière de suspecter de l'autocorrélation sur ces données. On effectue cependant par précaution une vérification sur la présence d'autocorrélations d'ordres 1 à 10. Le test de Breusch-Godfrey montre qu'il n'y a pas d'autocorrélation des résidus d'ordres 1 à 10 :

	p-value
Autocorrélation d'ordre 1	0,84
Autocorrélation d'ordre 2	0,47
Autocorrélation d'ordre 3	0,41
Autocorrélation d'ordre 4	0,30
Autocorrélation d'ordre 5	0,25
Autocorrélation d'ordre 6	0,34
Autocorrélation d'ordre 7	0,26
Autocorrélation d'ordre 8	0,37
Autocorrélation d'ordre 9	0,46
Autocorrélation d'ordre 10	0,57

3. Conclusion

- Quant à la relation entre variable indépendante et variable dépendante

Les résultats obtenus indiquent une relation significative entre effectif des présents aux épreuves et taux de réussite de l'académie.

- Quant au modèle

Mais la portée de ce modèle est limitée car le coefficient de détermination est faible : $R^2 = 0{,}14$, ce qui signifie que 86% (c'est-à-dire 1-0,14) de la variance du taux de succès n'est pas expliquée. Il faut prendre en compte bien d'autres variables si l'objectif est de comprendre ce qui détermine le taux de réussite. C'est aussi ce que suggère le fait que les conditions d'indépendance des résidus à la variable dépendante et à la variable indépendante ne sont pas remplies : les résidus incorporent vraisemblablement une composante déterministe qui devrait être extraite et intervenir comme variable indépendante pour améliorer le modèle.

Chapitre 18. Régression linéaire multiple : modéliser une relation entre deux (ou plus) variables quantitatives indépendantes et une variable quantitative dépendante

La régression linéaire multiple[163] prolonge la régression linéaire simple en introduisant la possibilité de prendre en compte plusieurs variables explicatives, et non plus une seule. Cela étant, comme dans la régression linéaire simple, toutes les variables sont quantitatives et il n'y a qu'une variable dépendante.

Comme dans la régression simple, l'objectif est de modéliser la relation entre variables, d'identifier la portée du modèle ainsi que la significativité et le sens (positif ou négatif) d'éventuels liens, et de se doter d'une base pour la prévision.

Cependant, dans la régression multiple, il y a au moins deux variables indépendantes, ce qui implique que l'effet sur la variable dépendante peut provenir non seulement des variables indépendantes elles-mêmes, mais aussi de leurs interactions. Les interactions peuvent donc être prises en compte également parmi les termes du modèle. Au-delà, cette pluralité des variables indépendantes pose aussi la question de l'importance

[163] Accessible dans XLSTAT via la commande *Modélisation des données / Régression linéaire*.

respective de ces variables. C'est aussi l'un des objectifs de la régression multiple que de déterminer l'importance respective des variables indépendantes. En règle générale, l'importance d'une variable indépendante est mesurée par l'effet qu'exerce cette variable sur la variable dépendante, effet lui-même représenté par le coefficient standardisé associé à cette variable indépendante, à condition que ce coefficient soit significatif.

L'absence de multicolinéarité entre variables indépendantes doit être préalablement vérifiée (et l'éventuelle multicolinéarité traitée)[164]. Les conditions de validité du modèle sont les mêmes que pour l'anova et la régression linéaire simple : le modèle est valide si les conditions de normalité, homoscédasticité et indépendance des résidus (distribution aléatoire, indépendance par rapport aux variables indépendantes, non-autocorrélation[165]) sont remplies.

---

[164] Voir Section 15.2.

[165] Le test de Breusch-Godfrey disponible dans le « Real Statistics Resource Pack » s'utilise comme indiqué en note 161 : si par exemple les variables indépendantes (X_1, X_2) sont placées dans la plage A1:B30, et la variable dépendante (Y) dans la plage C1:C30, on inscrit en D1 :

=BGTEST(A1:B30,C1:C30,λ)

où λ est l'ordre d'autocorrélation auquel on s'intéresse.

EXEMPLE 18.1

On cherche à identifier les conditions qui favorisent la violence scolaire. L'une des théories candidates met en avant la taille de la population scolaire et le taux d'encadrement des élèves. On dispose de données sur le nombre d'incidents, le nombre d'établissements ainsi que l'effectif et le taux d'encadrement par établissement dans trente villes (Tableau 18.1). On applique une régression multiple afin d'identifier d'éventuels liens significatifs.

Tableau 18.1.
Incidents de violence scolaire, démographie scolaire et encadrement institutionnel dans trente villes

Identifiant de la ville	Nombre d'établis-sements	Nombre moyen d'élèves par établis-sement	Nombre d'encadrants par élève	Nombre moyen d'incidents graves pour 1000 élèves
1	10	320	0,03	20
2	3	168	0,09	4
3	16	230	0,05	20
4	19	167	0,03	29
5	2	271	0,03	13
6	18	256	0,03	27
7	12	180	0,07	10
8	8	261	0,04	17
9	1	329	0,08	4

Identifiant de la ville	Nombre d'établis-sements	Nombre moyen d'élèves par établis-sement	Nombre d'encadrants par élève	Nombre moyen d'incidents graves pour 1000 élèves
10	16	350	0,04	26
11	5	213	0,08	3
12	11	180	0,04	14
13	5	257	0,07	2
14	10	252	0,07	13
15	5	235	0,06	9
16	2	242	0,10	5
17	14	182	0,06	17
18	5	253	0,10	17
19	7	183	0,03	8
20	4	201	0,05	8
21	9	343	0,06	19
22	4	347	0,09	8
23	2	217	0,10	17
24	3	326	0,03	8
25	16	225	0,05	14
26	13	254	0,05	20
27	14	251	0,03	28
28	9	251	0,05	16
29	7	331	0,07	9
30	10	300	0,09	8

1. Vérification de la condition préalable de non-multicolinéarité des variables indépendantes

Statistique	Nombre d'établis-sements	Nombre moyen d'élèves par établissement	Nombre d'encadrants par élève
R^2	0,29	0,03	0,27
Tolérance	0,71	0,97	0,73
VIF	1,40	1,04	1,36

Le VIF est largement inférieur à 5 pour toutes les variables indépendantes, donc il n'y a pas de multicolinéarité des variables indépendantes[166].

2. Significativité des paramètres

On estime un modèle sans constante, considérant qu'il n'y a pas de raison théorique de fixer un plancher au nombre d'incidents (s'il n'y a pas d'établissement

[166] Le R^2 mentionné dans le tableau est le coefficient de détermination obtenu en régressant la variable considérée sur les autres variables indépendantes. On peut observer que :

$$tolérance = 1 - R^2$$

scolaire dans la ville, $x_1 = x_2 = x_3 = 0$, et il n'y a pas d'incident).

Les paramètres de régression s'établissent comme suit :

Termes de la régression	VP*	ES**	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Constante	0,00					
Nombre d'établis-sements	1,47	0,55	2,67	0,01	0,33	2,61
Nombre moyen d'élèves par établis-sement	0,01	0,02	0,63	0,53	-0,02	0,04
Nombre d'encadrants par élève	100,08	83,62	1,20	0,24	-72,5	272,65
Nombre d'établis-sements * Nombre moyen d'élèves par	0,002	0,002	1,09	0,29	0,00	0,01

Termes de la régression	VP*	ES**	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
établis-sement						
Nombre d'établis-sements * Nombre d'encadrants par élève	-21,95	7,72	-2,8	0,01	-37,8	-6,01
Nombre moyen d'élèves par établis-sement * Nombre d'encadrants par élève	-0,16	0,36	-0,4	0,66	-0,90	0,58

* VP : Valeur des paramètres
** ES : Erreur standard

On constate que, dans cet échantillon :

- le nombre d'établissements exerce une influence significative sur le nombre d'incidents de violence scolaire (la p-value du paramètre de cette variable est égale à 0,01). Le signe de ce paramètre indique que

plus il y a d'établissements, plus le nombre d'incidents est élevé ;

- la combinaison entre nombre d'établissements et nombre d'encadrants par élève influence significativement elle aussi la violence scolaire (p=0,01). Mais ici la relation est négative : plus le produit du nombre d'établissements et du nombre d'encadrants est élevé, moins il y a d'incidents.

La valeur des paramètres standardisés permet de hiérarchiser les variables indépendantes en fonction de l'ampleur de leur effet sur la variable dépendante. Ici par exemple, les paramètres standardisés s'établissent comme suit :

Termes de la régression	VP*	ES**	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Nombre d'établis-sements	0,95	0,35	2,67	0,01	0,22	1,68
Nombre moyen d'élèves par établis-sement	0,16	0,26	0,63	0,53	-0,37	0,70

Termes de la régression	VP*	ES**	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Nombre d'encadrants par élève	0,41	0,34	1,20	0,24	-0,29	1,10
Nombre d'établis-sements * Nombre moyen d'élèves par établis-sement	0,36	0,33	1,09	0,29	-0,32	1,05
Nombre d'établis-sements * Nombre d'encadrants par élève	-0,7	0,25	-2,8	0,01	-1,23	-0,19

Termes de la régression	VP*	ES**	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Nombre moyen d'élèves par établis-sement * Nombre d'encadrants par élève	-0,1	0,38	-0,4	0,66	-0,95	0,61

* VP : Valeur des paramètres
** ES : Erreur standard

On peut donc voir que le nombre d'établissements est la variable qui exerce le poids le plus élevé sur le nombre d'incidents, juste devant la combinaison nombre d'établissements*nombre d'encadrants.

On peut enfin préciser que si tous les paramètres avaient été significatifs, il aurait été possible de prédire, avec un degré de confiance de 95% et en acceptant une erreur ε, le nombre d'incidents en s'appuyant sur le modèle :

$$Nombre\ d'incidents = (\beta_1 \times x_1) + (\beta_2 \times x_2) + (\beta_3 \times x_3) + (\gamma_1 \times x_1 \times x_2) + (\gamma_2 \times x_1 \times x_3) + (\gamma_3 \times x_2 \times x_3)$$

où

x_1 est le nombre d'établissements ;
x_2, le nombre moyen d'élèves par établissement ;
x_3, le nombre d'encadrants par élève ;
β_1 est le coefficient non-standardisé de x_1 (soit 1,47) ;
β_2, le paramètre non-standardisé de x_2 (soit 0,01) ;
β_3, le paramètre non-standardisé de x_3 (soit 100,08) ;
γ_1, γ_2, et γ_3, les paramètres non-standardisés respectifs des combinaisons de variables.

3. Vérification des conditions de validité

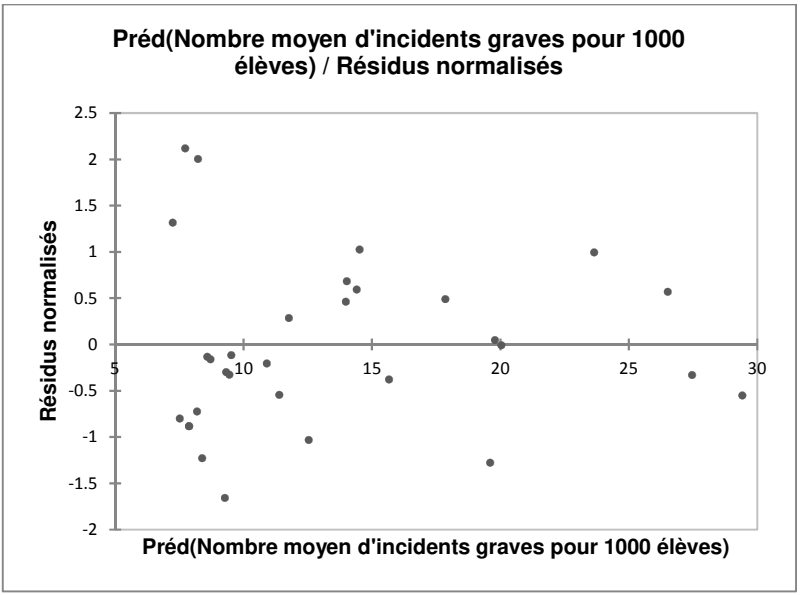
- Normalité des résidus

Tous les tests montrent que les résidus sont normalement distribués :

	Shapiro -Wilk	Anderson -Darling	Lilliefors	Jarque -Bera
Résidu standardisé	0,52	0,53	0,37	0,50

- Homoscédasticité

L'examen du graphique « $\widehat{Y}$/Résidu » montre que la condition d'homoscédasticité n'est pas remplie :



- Indépendance des résidus par rapport à la variable dépendante estimée

Le graphique « $\widehat{Y}$/Résidu » ci-dessus ne suggère pas de relation particulière entre résidu et variable dépendante estimée. Le calcul du coefficient de corrélation linéaire (r=0,01, p-value=0,94) confirme que la condition de distribution aléatoire est remplie.

- Indépendance des résidus par rapport aux variables indépendantes

Les coefficients de corrélation entre résidus standardisés et chacune des variables indépendantes montrent que la condition est remplie :

Variables	Résidus standardisés	
	Coefficient de corrélation	p-value
Nombre d'établissements	0,01	0,95
Nombre moyen d'élèves par établissement	0,03	0,87
Nombre d'encadrants par élève	0,02	0,93

- Non-autocorrélation des résidus

Le test de Breusch-Godfrey montre qu'il n'y a pas d'autocorrélation des résidus d'ordres 1 à 10 :

	p-value
Autocorrélation d'ordre 1	0,65
Autocorrélation d'ordre 2	0,65
Autocorrélation d'ordre 3	0,60
Autocorrélation d'ordre 4	0,76
Autocorrélation d'ordre 5	0,69
Autocorrélation d'ordre 6	0,78
Autocorrélation d'ordre 7	0,51
Autocorrélation d'ordre 8	0,58
Autocorrélation d'ordre 9	0,69
Autocorrélation d'ordre 10	0,70

4. Conclusion

- Quant à la relation entre variable indépendante et variables dépendantes :
  Les estimations montrent que le nombre d'établissements d'abord, et la combinaison entre nombre d'établissements et nombre d'encadrants par élève ensuite, exercent une influence significative sur le nombre d'incidents de violence scolaire. Ces deux facteurs font respectivement augmenter (le premier) et diminuer (le second) la violence scolaire.

- Quant au modèle :
  La portée du modèle est substantielle : les variables indépendantes (interactions comprises) expliquent près des trois-quarts de la variance de la variable dépendante ($R^2 = 0{,}73$). Cependant la validité du modèle n'est pas totalement établie car les résidus sont hétéroscédastiques.

Chapitre 19. Prendre en compte l'influence de variables quantitatives et qualitatives sur une variable-réponse quantitative

L'objectif ici est d'analyser l'effet que des variables indépendantes qualitatives et quantitatives exercent simultanément sur une variable dépendante quantitative. Deux approches peuvent être distinguées : la régression linéaire multiple avec variables qualitatives, et l'analyse de covariance (Ancova). Chacune se distingue par ses objectifs et conditions de validité.

19.1. RÉGRESSION LINÉAIRE MULTIPLE AVEC VARIABLES QUALITATIVES

La régression linéaire multiple avec une ou plusieurs variables qualitatives[167] permet de formuler la relation entre variables de façon précise sous la forme d'un modèle calculable ; d'identifier la portée du modèle, la significativité et le sens (positif ou négatif) des éventuels liens, et l'importance respective des variables indépendantes ; et de se doter d'une base pour la prévision. Ici également, les effets des interactions

[167] Accessible dans XLSTAT via la commande *Modélisation des données / Régression linéaire*. L'outil de calcul est en fait le même que pour l'Ancova, qui s'active dès que la présence de variables qualitatives est mentionnée.

entre variables indépendantes peuvent être pris en compte.

La régression linéaire multiple avec variables qualitatives est généralement utilisée avec pour objectif de prendre en compte l'ensemble le plus complet possible de variables explicatives pertinentes (qu'elles soient qualitatives ou quantitatives) afin de comprendre le mieux possible comment est déterminée la variable expliquée.

Les conditions de validité sont les mêmes que pour la régression linéaire multiple : l'absence de multicolinéarité entre variables indépendantes doit être préalablement vérifiée ; et le modèle est valide si les conditions de normalité, homoscédasticité et indépendance des résidus (distribution aléatoire, indépendance à la variable indépendante, non-autocorrélation) sont remplies.

EXEMPLE 19.1

On cherche à comprendre les déterminants de la satisfaction des enseignants-chercheurs dans l'exercice de leur métier. L'une des théories candidates met en avant l'âge, la discipline et le genre. En vue de tester cette théorie, on administre à trente universitaires un questionnaire destiné à mesurer leur degré de satisfaction sur une échelle de 1 à 100. Le Tableau 19.1

présente les valeurs des variables pour les trente membres de l'échantillon.

Tableau 19.1.
Age, genre, discipline (D) et indice de satisfaction (IS) dans l'exercice du métier pour trente enseignants-chercheurs

Note
Disciplines : 1 – Droit, économie, gestion ; 2 – Lettres et sciences humaines ; 3 – Sciences

Identifiant de l'enseignant-chercheur	Age	Genre	D	IS
1	51	F	1	34
2	40	F	3	30
3	58	H	2	61
4	57	F	3	94
5	42	H	3	17
6	33	F	2	82
7	35	H	3	79
8	44	H	3	40
9	35	H	1	98
10	64	F	3	94
11	51	F	2	21
12	64	F	3	92
13	46	H	1	59
14	44	H	1	97
15	60	H	2	82
16	57	H	1	95

Identifiant de l'enseignant-chercheur	Age	Genre	D	IS
17	34	H	3	50
18	43	F	1	95
19	37	F	1	97
20	42	F	2	68
21	64	F	3	82
22	54	H	1	82
23	56	F	1	94
24	49	F	2	19
25	63	H	3	93
26	32	H	2	95
27	50	F	2	70
28	59	H	1	38
29	43	F	1	96
30	60	F	1	94

1. Vérification de la condition préalable de non-multicolinéarité

La question de la multicolinéarité ne se pose pas puisqu'une seule des variables indépendantes est quantitative ou ordinale.

2. Significativité des paramètres

On estime un modèle sans constante car n'y a pas de raison théorique en l'espèce de supposer un niveau minimal de satisfaction[168]. Les paramètres de régression s'établissent comme suit :

Termes du modèle	VP	ES	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Cste	0,00					
Age	1,37	0,11	12,6	<0,0001	1,15	1,60
Genre-1	32,15	30,4	1,06	0,30	-31,16	95,47
Genre-2	-28,1	26,6	-1,06	0,30	-83,54	27,27
Discipl. -1	38,50	36,8	1,04	0,31	-38,17	115,1
Discipl. -2	61,66	45,9	1,34	0,19	-33,98	157,2
Discipl. -3	-95,5	40,3	-2,37	0,03	-179,4	-11,65

[168] Le modèle est en outre ici spécifié avec somme des paramètres nulle : dans XLSTAT, *Contraintes : Somme (ni.ai)=0* du menu Options de la fonction *Modélisation des données / Régression linéaire*.

Termes du modèle	VP	ES	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Age * Genre-1	-0,53	0,61	-0,87	0,39	-1,81	0,74
Age * Genre-2	0,47	0,53	0,87	0,39	-0,64	1,58
Age * Discipl. -1	-0,53	0,74	-0,72	0,48	-2,06	1,00
Age * Discipl. -2	-1,33	0,95	-1,41	0,17	-3,30	0,64
Age * Discipl. -3	1,70	0,79	2,16	0,04	0,07	3,34
Genre-1 * Discipl. -1	-9,94	7,28	-1,36	0,19	-25,08	5,20
Genre-1 *Discipl. -2	7,29	9,83	0,74	0,47	-13,16	27,73
Genre-1*	6,10	8,79	0,69	0,50	-12,18	24,38

Termes du modèle	VP	ES	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Discipl. -3						
Genre-2 * Discipl. -1	8,70	6,37	1,36	0,19	-4,55	21,95
Genre-2 * Discipl. -2	-6,38	8,60	-0,74	0,47	-24,27	11,51
Genre-2 * Discipl. -3	-5,34	7,69	-0,69	0,50	-21,33	10,66

Il apparaît que trois des termes du modèle ont une influence sur la satisfaction :

- l'âge d'abord, dont le paramètre (1,37) est significatif ($p<0,0001$). Le paramètre de l'âge est positif, ce qui signifie que, dans cet échantillon, la satisfaction dans l'exercice du métier augmente avec l'âge ;
- la discipline ensuite, puisque le paramètre de la discipline 3 (-95,5) est significatif ($p=0,03$). La satisfaction dans l'exercice du métier dépend donc de

la discipline. En comparant les paramètres des trois disciplines, on voit que l'appartenance à la discipline 3 exerce un effet négatif sur la satisfaction ;

- la combinaison âge-discipline, enfin, puisque le paramètre de la combinaison âge-discipline 3 (1,70) est significatif (p=0,04). Le signe du paramètre montre que dans la discipline 3 la satisfaction augmente avec l'âge.

En revanche, les tests de Student n'indiquent aucun effet significatif du genre. L'analyse de cet échantillon contribue donc à une confirmation partielle de la théorie : il semble y avoir un effet de l'âge et de la discipline sur la satisfaction, mais on ne trouve pas trace d'effet du genre.

Enfin, on peut préciser que si tous les paramètres avaient été significatifs, il aurait été possible, avec un degré de confiance de 95% et en acceptant une erreur ε, de prédire la satisfaction (au sens du questionnaire utilisé) en s'appuyant sur le modèle :

$$Indice\ de\ satisfaction = (\beta_1 \times \hat{a}ge) + \beta_2 + \beta_3 + (\gamma_1 \times \hat{a}ge) + (\gamma_2 \times \hat{a}ge) + \gamma_3$$

où

$\beta_1 = 1{,}37$ est le paramètre de l'âge ;

β_2, le paramètre du genre (32,15 pour les hommes, -28,1 pour les femmes) :

β_3, le paramètre de la discipline ;

γ_1, le paramètre de la combinaison âge-genre ;

γ_2, le paramètre de la combinaison âge-discipline ; et

γ_3, le paramètre de la combinaison genre-discipline.

3. Vérification des conditions de validité

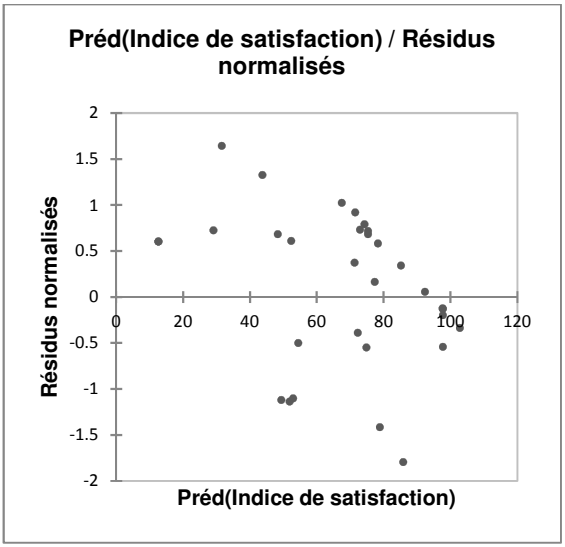
- Normalité

Tous les tests montrent que les résidus sont normalement distribués :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Résidus standardisés	0,34	0,17	0,07	0,53

- Homoscédasticité

Le graphique Prédictions/Résidus ne suggère pas de violation manifeste de la condition d'homoscédasticité :



Le modèle comporte deux facteurs catégoriels, donc on peut comparer les variances des résidus répartis par catégorie pour chaque facteur. Les tests confirment que la condition d'homoscédasticité est remplie :

	p		
	Fisher	Levene	Bartlett
Comparaison des variances des résidus des catégories Homme et Femme	0,91	0,92	0,90
Comparaison des variances des résidus des Disciplines 1, 2 et 3	–	0,97	0,94

- Distribution aléatoire des résidus

La corrélation entre prédictions et résidus est modérée et non-significative :

	Préd(Indice de satisfaction)	
	Coefficient de corrélation	p-value
Résidus standardisés	-0,31	0,09

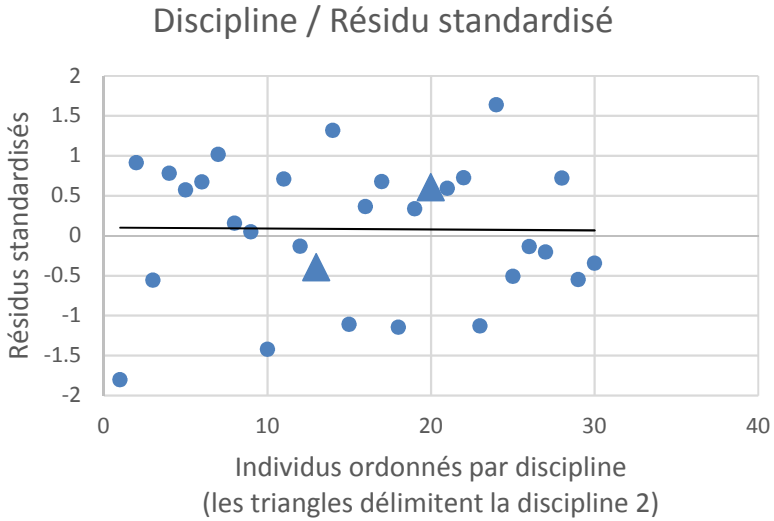
Donc la condition peut être considérée comme remplie.

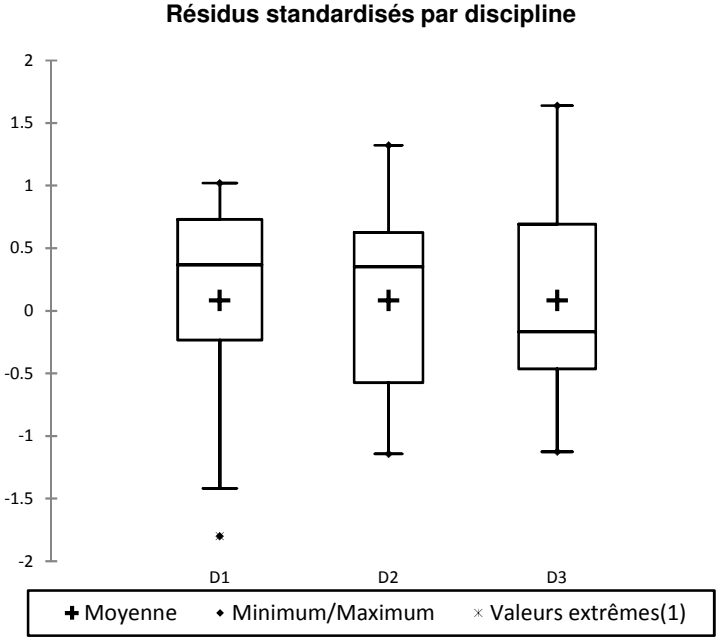
- Indépendance des résidus aux variables indépendantes

On peut calculer le coefficient de corrélation de Pearson entre l'âge et les résidus standardisés, et le coefficient de corrélation bisérielle ponctuelle entre le genre (variable qualitative binaire) et les résidus standardisés :

	Résidus standardisés	
	Coefficient de corrélation	p-value
Age	-0,48	0,01
Genre	0,00	0,99

On peut aussi visualiser la distribution des résidus pour les individus de chaque catégorie de la variable Discipline :





On peut voir que les résidus standardisés ne sont corrélés ni au genre ni à la discipline. En revanche, ils sont moyennement et significativement corrélés à l'âge. Donc la condition d'indépendance des résidus aux variables indépendantes n'est pas remplie.

- Non-autocorrélation des résidus

Les résultats d'estimation indiquent que la valeur calculée de la statistique de Durbin-Watson s'établit à 2,06. Donc il n'y a pas d'autocorrélation d'ordre 1.

4. Conclusion

Les estimations sur cet échantillon (fictif) montrent que l'âge, la discipline et la combinaison des deux influencent significativement la satisfaction des enseignants-chercheurs dans l'exercice de leur métier. L'effet de la discipline est négatif, celui des deux autres facteurs positifs.

Le coefficient de détermination $R^2 = 0{,}29$. Le modèle n'explique donc qu'une faible part de la variance de la variable dépendante. En outre, la condition d'indépendance des résidus aux variables indépendantes n'est pas remplie. Le modèle devrait donc être redéfini de façon à inclure parmi les variables indépendantes des facteurs éventuellement pertinents que pourraient avoir capturé les résidus.

19.2. L'ANALYSE DE COVARIANCE

L'analyse de covariance (ancova)[169] s'utilise lorsque, cherchant à identifier l'effet d'un « traitement » sur un objet, le chercheur doit prendre en compte l'éventuelle influence d'une circonstance extérieure (c'est-à-dire sans lien avec le traitement) mais susceptible d'exercer un effet déterminant sur l'objet. La problématique est

[169] Accessible dans XLSTAT via la commande *Modélisation des données / Ancova*.

celle de la démarche expérimentale avec comparaison entre groupe traité et groupe témoin, ou comparaison entre différents groupes de traitement. Ne pas tenir compte de la circonstance extérieure entrainerait un risque de commettre une erreur lors de la mesure de l'effet du traitement[170]. Lorsqu'on prend en compte la circonstance extérieure, on dit qu'on en « contrôle » l'effet. L'ancova vise à montrer si la moyenne de la variable dépendante diffère significativement d'un groupe à l'autre quand on tient compte de la circonstance extérieure.

Un exemple de ce cas de figure est l'analyse de l'effet d'une réforme éducative – par exemple la réforme de la formation des professeurs – sur l'efficacité pédagogique lorsque, de toute évidence, une circonstance extérieure telle que le quotient intellectuel de l'élève est aussi, par ailleurs, un déterminant potentiel de sa performance scolaire.

---

[170] Au contraire, faire figurer la circonstance extérieure parmi les variables indépendantes augmente la puissance du test F d'anova en retirant du dénominateur de la statistique F (variabilité résiduelle) la variance de la variable dépendante imputable à la circonstance extérieure. La valeur de F augmente, ce qui permet de mieux mesurer l'éventuelle significativité des différences entre catégories ou groupes (variabilité intraclasses) : voir Encadré 15.1.

La terminologie de l'analyse de covariance distingue parmi les variables indépendantes entre « covariable » et « facteur ». La covariable (*covariate*) ou « variable de contrôle » représente la circonstance extérieure. C'est toujours une variable quantitative. En général, il n'y a qu'une covariable par modèle. La covariable n'est pas elle-même la variable d'intérêt : elle n'intervient dans l'analyse que pour mieux faire ressortir l'effet du traitement. La variable d'intérêt est le « facteur » ou « prédicteur », c'est-dire la variable indépendante qualitative qui représente le traitement. Les modalités du facteur désignent les groupes traité(s) et témoin(s). Il peut y avoir plusieurs facteurs dans le modèle.

Il importe de noter que parmi les circonstances extérieures – que représentent les covariables – figure aussi le niveau qu'avait atteint la variable dépendante avant expérimentation. Dans certains cas, le niveau de la variable dépendante à l'issue de l'expérimentation ne peut être exclusivement attribué à l'expérimentation elle-même, mais dépend du niveau déjà atteint avant l'expérimentation. Le chercheur a donc le choix entre définir la variable dépendante en variation (écart entre niveau d'avant expérimentation et niveau d'après expérimentation), ou la définir en niveau, et dans ce dernier cas il peut être pertinent d'introduire le niveau d'avant expérimentation comme covariable.

19.2.1. Conditions de validité

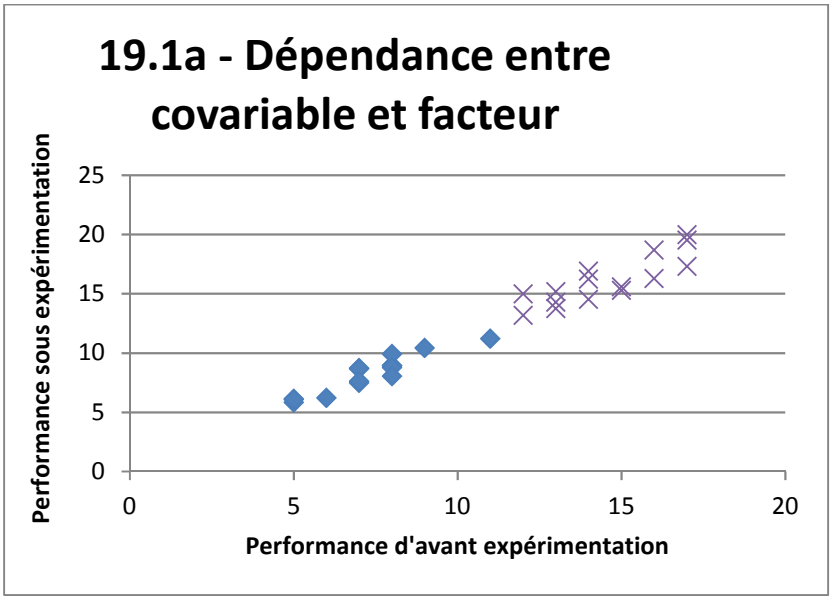
La validité d'une analyse de covariance nécessite d'abord que soient remplies les mêmes conditions que pour une régression linéaire multiple avec variable(s) qualitative(s) : absence de multicolinéarité entre variables indépendantes ; et normalité, homoscédasticité et indépendance des résidus. Mais trois conditions spécifiques supplémentaires sont aussi à vérifier préalablement : indépendance entre covariables et facteurs ; existence d'une relation linéaire entre chaque covariable et la variable dépendante ; et homogénéité des pentes des droites de régression.

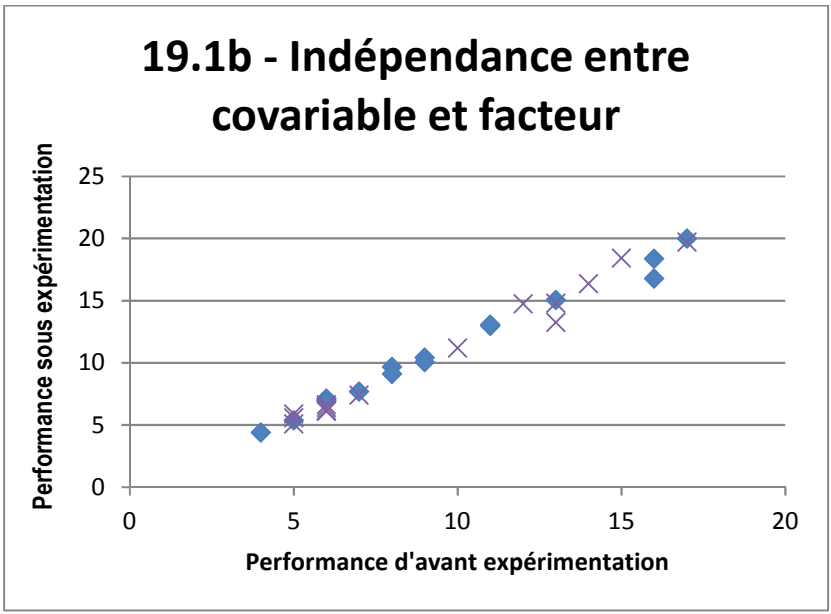
19.2.1.1. Indépendance entre covariables et facteurs

L'objectif de l'analyse étant d'évaluer l'effet d'un traitement en comparant différents groupes (les catégories du facteur), il est évident que ces groupes doivent être similaires. À défaut, la comparaison serait biaisée. La similarité entre groupes est appréciée au regard de la covariable : les groupes sont similaires s'ils ont la même distribution statistique en termes de la covariable. Lorsque c'est le cas, la condition d'indépendance entre la covariable et le facteur est remplie. La comparaison de distributions peut s'effectuer au moyen du test de Kolmogorov-

Smirnov[171]. On peut aussi l'effectuer visuellement au moyen du nuage de points entre la covariable et la variable dépendante. Supposons par exemple une expérimentation visant à analyser l'impact sur la performance scolaire lorsqu'on augmente la durée de la semaine d'école de quatre à cinq jours. La variable de contrôle est le niveau de performance d'avant expérimentation, qui est susceptible d'influencer aussi la performance lors de l'expérimentation, mais qui est indépendant de l'expérimentation elle-même. Le Graphique 19.1 ci-après illustre des cas de dépendance (19.1a) et d'indépendance (19.1b) entre covariable et facteur :

[171] Le test de Kolmogorov-Smirnov, généralement disponible sur les logiciels statistiques, permet de comparer des distributions deux-à-deux. L'hypothèse nulle du test est que les distributions sont égales, l'hypothèse alternative du test bilatéral est qu'elles sont différentes. On peut rejeter l'hypothèse nulle quand la p-value est inférieure à 0,05. Dans XLSTAT, le test de Kolmogorov-Smirnov est accessible sous l'onglet *Tests non paramétriques*.


19.1a - Dépendance entre covariable et facteur
Performance sous expérimentation
25
20
15
10
5
0
0
5
10
15
20
Performance d'avant expérimentation


19.1b - Indépendance entre covariable et facteur
Performance sous expérimentation
25
20
15
10
5
0
0
5
10
15
20
Performance d'avant expérimentation

Sur le Graphique 19.1a, le groupe expérimental (dont les membres sont représentés par des croix) et le groupe témoin (dont les membres sont représentés par des losanges) n'ont pas les mêmes niveaux de performance avant expérimentation : les membres du groupe expérimental ont une performance plus élevée (entre 11 et 17) que celle du groupe témoin (entre 5 et 11). Il y a donc une relation entre l'affectation à un groupe et la performance d'avant expérimentation, c'est-à-dire une dépendance entre facteur et covariable.

Au contraire, sur le Graphique 19.1b, il n'y a pas de différence entre groupes du point de vue de la performance d'avant expérimentation. Les deux distributions sont similaires. La condition d'indépendance entre covariable et facteur est remplie.

19.2.1.2. Existence d'une relation linéaire entre chaque covariable et la variable dépendante

La condition de linéarité se vérifie au moyen du coefficient de corrélation de Pearson. L'absence de linéarité implique que l'introduction de la covariable ne permettra pas d'améliorer la puissance du test F, alors même que cette amélioration est la raison d'être du recours à la procédure d'ancova. Donc, quand la condition de linéarité n'est pas remplie, il faut, par exemple, changer de covariable ou adopter une

procédure statistique plus appropriée, par exemple une anova.

19.2.1.3. Homogénéité des pentes des droites de régression

Cette condition signifie que la relation entre covariable et variable dépendante est la même dans toutes les catégories du facteur. Supposons une analyse de covariance avec une covariable X_1, un facteur X_2 de modalités m_1 et m_2, et une variable dépendante *Y*. Pour vérifier la condition d'homogénéité des pentes des droites de régression, on teste le modèle de régression $(Y, X_1, X_2, X_1 \times m_1, X_1 \times m_2)$, c'est-à-dire un modèle comprenant les interactions entre la covariable et les modalités du facteur, contre le même modèle mais sans interactions (Y, X_1, X_2). La condition d'homogénéité des pentes de régression est remplie si le terme d'interaction n'est pas significatif[172]. Lorsque la

[172] Dans XLSTAT, la p-value (notée *Pr > F*) du test de significativité du terme d'interaction est fournie dans le tableau *Analyse Type III Sum of Squares* lorsque sont cochées les options (a) *Interactions/Niveau* sous l'onglet *Options* et (b) *Type I/II/III SS* sous l'onglet *Sorties/Général*. Les interactions n'ont pas d'effet significatif (et la condition d'homogénéité des pentes est remplie) si la p-value est supérieure à 0,05.

Un outil alternatif pour effectuer le test de significativité du terme d'interaction est la fonction RSquareTest du « Real Statistics Resource Pack » (voir note de bas de page n°4 page 9).

condition n'est pas remplie, cela signifie que la relation entre variables indépendantes et variable dépendante diffère d'un groupe à l'autre. Dans ce cas, chaque catégorie du facteur doit faire l'objet d'une analyse séparée, suivant une procédure statistique plus appropriée que ne l'est alors l'ancova.

19.2.2. Exemple

On expérimente un programme du type *Devoirs faits* (les devoirs sont faits pendant le temps de présence dans l'établissement) sur 34 élèves. L'objectif est d'identifier l'éventuel effet de la durée des séances de Devoirs sur les résultats scolaires, de façon que l'établissement puisse calibrer au mieux la durée des

---

Dans l'exemple du modèle $(Y, X_1, X_2, X_1 \times m_1, X_1 \times m_2)$ considéré plus haut, à supposer qu'il y ait 30 observations dans l'échantillon, Y sera placée dans la plage A1:A30 de la feuille Excel, et X_1 dans la plage B1:B30. Dans la plage C1:C30, on inscrit 1 pour chaque observation qui vérifie la modalité m_1, et 0 pour les autres. Dans la plage D1:D30, on inscrit 1 pour chaque observation qui vérifie la modalité m_2, et 0 pour les autres. En E1:E30, on inscrit pour chaque observation le produit $X_1 \times m_1$. En F1:F30, on inscrit pour chaque observation le produit $X_1 \times m_2$. Puis on sélectionne la cellule G1 et on inscrit dans la fenêtre la formule

=RSquareTest(B1:F30,B1:D30,A1:A30)

Enfin on appuie sur la touche Entrée. En G1 s'inscrit la p-value du test. La condition d'homogénéité est remplie si la p-value est supérieure à 0,05.

séances. Trois formats sont définis : 1 heure ; 2 heures ; 3 heures. Les élèves sont assignés aléatoirement à l'un ou l'autre des formats. Le Tableau 19.2 synthétise les observations dont on dispose à l'issue de l'expérience.

Tableau 19.2.
Données de l'expérimentation *Devoirs faits*

Note
Format des séances – 1 : 1 heure ; 2 : 2 heures ; 3 : 3 heures

Identifiant de l'élève	Moyenne générale avant expérimentation	Format	Moyenne générale après expérimentation
1	7	1	7
2	15	1	14
3	8	1	9
4	7	3	9
5	12	3	15
6	8	2	9
7	14	1	15
8	16	2	18
9	4	1	4
10	9	1	10
11	11	1	10
12	8	2	10
13	16	3	19
14	10	3	13
15	16	1	16

Identifiant de l'élève	Moyenne générale avant expérimentation	Format	Moyenne générale après expérimentation
16	6	3	6
17	4	3	4
18	14	3	15
19	8	2	10
20	9	3	12
21	11	1	11
22	10	2	11
23	12	3	15
24	11	2	11
25	13	1	14
26	14	2	17
27	13	2	13
28	15	2	15
29	12	2	12
30	10	2	13
31	8	3	9
32	7	3	7
33	13	3	15
34	4	2	5

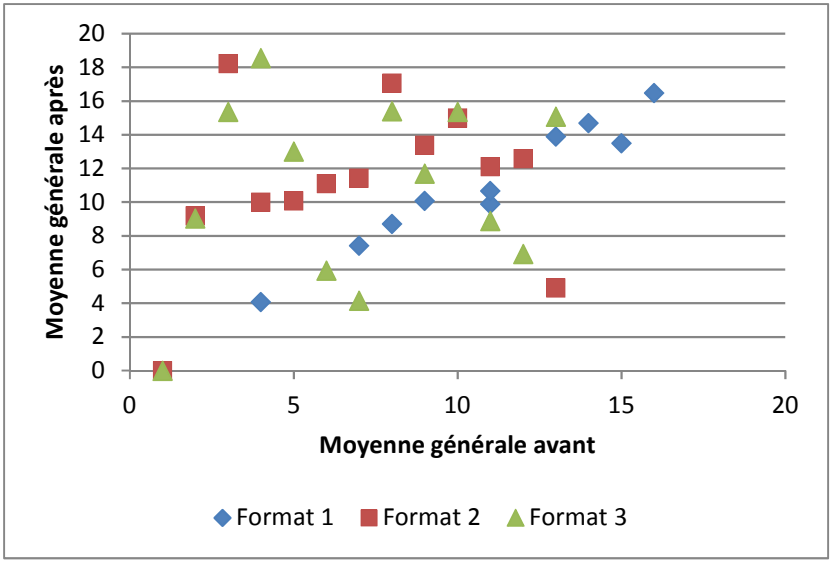
1. Multicolinéarité

Il n'y a qu'une variable indépendante quantitative ou ordinale, donc la question de la multicolinéarité des variables indépendantes ne se pose pas.

2. Conditions spécifiques à l'ancova

- Indépendance entre covariable et facteur

Le graphique covariable / variable dépendante montre qu'il n'y a pas dépendance entre covariable et facteur :



Le test de Kolmogorov-Smirnov confirme ce diagnostic :

Test de Kolmogorov-Smirnov

	p-value
Format 1 – Format 2	0,99
Format 2 – Format 3	0,82
Format 1 – Format 3	0,99

- Relation de linéarité entre covariable et variable dépendante

Le coefficient de corrélation linéaire entre covariable et variable dépendante s'établit à 94%, avec p-value < 0,0001.

- Homogénéité des pentes des droites de régression

Le test du modèle avec interactions contre le modèle sans interaction indique que le terme d'interaction n'est pas significatif : p-value = 0,06. La condition d'homogénéité des pentes des droites de régression est donc remplie.

3. Analyse des résultats

On estime le modèle sans interaction. Le tableau d'analyse de la variance, tout d'abord, indique que les variables indépendantes sélectionnées pour expliquer la variable dépendante dans ce modèle apportent une information significative :

Source	DDL	Somme des carrés	Moyenne des carrés	F	Pr > F
Modèle	3	469,70	156,56	133,55	<0,0001
Erreur	30	35,16	1,17		
Total corrigé	33	504,87			

Le tableau *Analyse Type III Sum of Squares*, ensuite, indique que le facteur[173] a un effet significatif sur la variable dépendante :

Analyse Type III Sum of Squares

Source	DDL	Somme des carrés	Moyenne des carrés	F	Pr > F
Moyenne générale avant	1	462,45	462,45	394,49	<0,0001
Format	2	16,21	8,10	6,91	0,003

[173] Il y a aussi, en l'occurrence, un effet significatif de la covariable, mais c'est le facteur qui est la variable d'intérêt en analyse de covariance.

Il faut donc comparer les catégories de facteur entre elles à l'aune de l'effet sur la variable dépendante. On examine les résultats des comparaisons multiples par paire :

Test de Tukey

Contraste	Différence	Différence standardisée*	Pr > Diff	Signifi-catif
3 vs 1	1,700	3,645	0,003	Oui
3 vs 2	0,496	1,115	0,513	Non
2 vs 1	1,205	2,599	0,037	Oui

* Valeur critique : 2,465

Moyennes des catégories en termes de la variable dépendante

Moda-lité	Moyennes estimées	ES*	Borne inf. (95%)	Borne sup. (95%)	Classes	
3	12,264	0,31	11,62	12,90	A	
2	11,769	0,31	11,12	12,40	A	
1	10,564	0,34	9,86	11,26		B

* ES : Erreur standard

On voit que les résultats scolaires des élèves soumis au format 1 (séances de devoirs d'une durée d'une heure) sont significativement inférieurs à ceux des élèves

soumis aux deux autres formats. Il n'y a pas de différence significative de performance entre élèves soumis aux formats 2 et 3.

4. Vérification des conditions de validité a posteriori

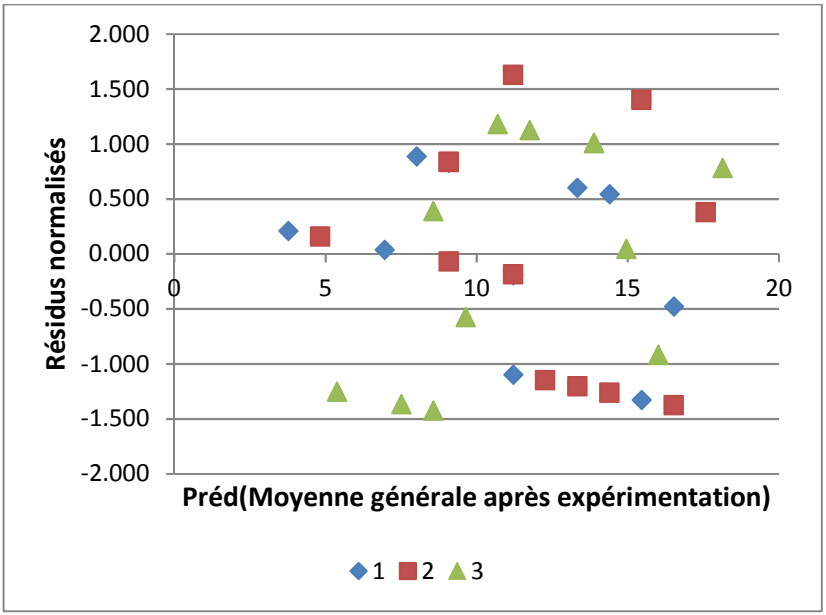
- Normalité des résidus

Tous les tests indiquent que les résidus sont normalement distribués :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Résidus std.	0,063	0,087	0,147	0,309

- Homogénéité des résidus

Le graphique prédictions/résidus ne révèle pas de différence évidente entre groupes quant à la dispersion des résidus :



Les tests d'homogénéité des variances des résidus répartis par catégories du facteur confirment que les variances sont homogènes :

	p-value	
	Levene	Bartlett
Comparaison des variances des résidus des catégories Formats 1, 2 et 3	0,12	0,37

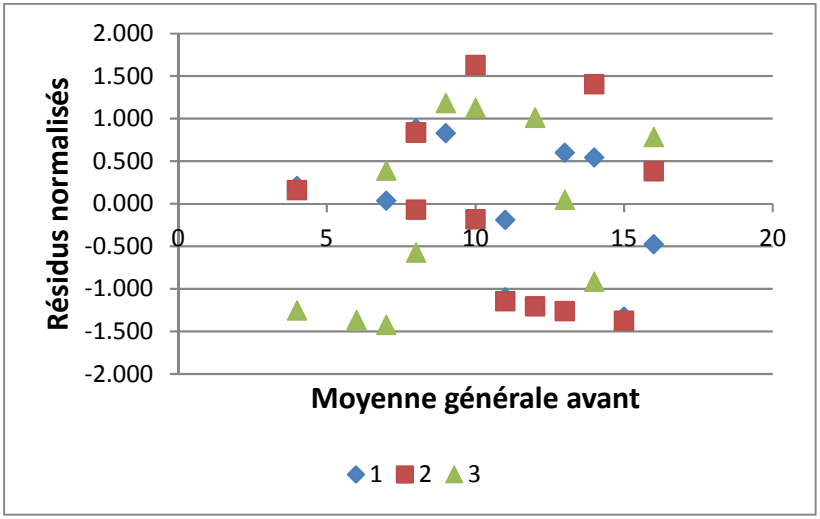
- Indépendance des résidus par rapport à la variable dépendante

Le coefficient de corrélation entre résidus et variable dépendante prédite est inférieur à 0,0001 avec une p-

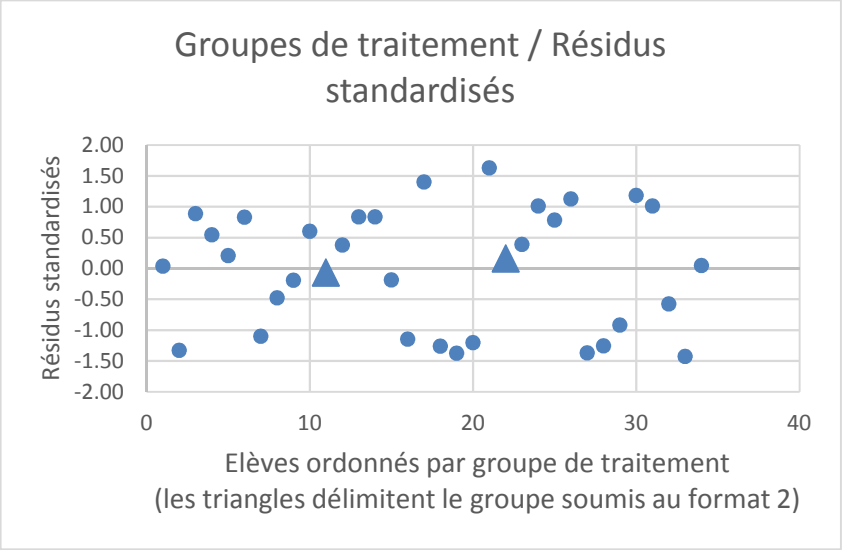
value de 1. Il y a donc indépendance des résidus par rapport à la variable dépendante.

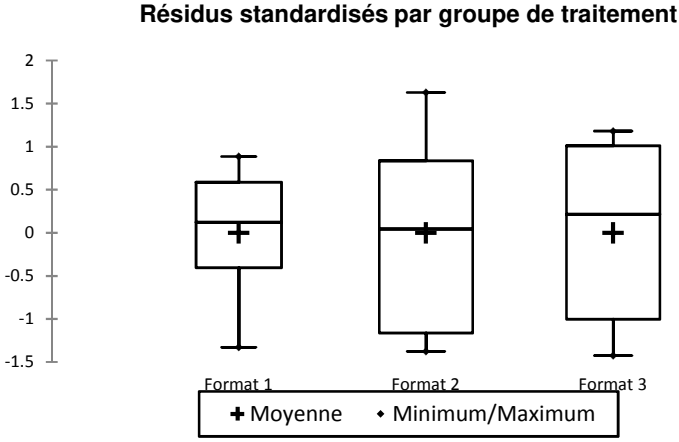
- Indépendance des résidus par rapport aux variables indépendantes

Les résidus sont indépendants de la covariable (coefficient de corrélation inférieur à 0,0001 avec une p-value de 1) :



Le graphique des résidus par catégories du facteur n'indique pas non plus de schéma particulier dans la répartition des résidus des trois catégories du facteur :





On peut donc considérer la condition d'indépendance des résidus aux variables indépendantes comme remplie.

- Non-autocorrélation des résidus

La statistique de Durbin-Watson s'établit à 1,99. Il n'y a donc pas d'autocorrélation d'ordre 1.

5. Conclusion

L'analyse montre donc que (dans le cadre de cette expérimentation fictive) le traitement (programme *Devoirs faits*) exerce un effet significatif sur les résultats scolaires des élèves. L'effet est significativement plus élevé lorsque la durée des séances de devoirs est de deux ou trois heures plutôt que d'une heure. Il n'y a pas de différence entre une durée de deux heures et une durée de trois heures.

Toutes les conditions de validité sont remplies. Le modèle peut donc être considéré comme valide, et les résultats bien établis.

Chapitre 20. Régression polynomiale et régression non-linéaire : prendre en compte des non-linéarités dans les relations entre variables quantitatives

Le manque de significativité du paramètre d'une variable explicative dans un modèle de régression linéaire simple ou multiple peut être dû non seulement au fait que le rôle de ladite variable est marginal ou nul dans l'explication de la variable indépendante, mais aussi au fait que la relation entre les deux variables en question n'est pas linéaire. Lorsqu'une variable dont on a toutes les raisons théoriques de penser qu'elle devrait exercer un effet significatif semble ne pas le faire, il importe de prendre en compte dans le modèle les éventuelles non-linéarités qui pourraient permettre de mieux appréhender la réelle significativité de la variable.

Deux méthodes permettent de prendre en compte les non-linéarités : la régression polynomiale (Section 20.1) et la régression non-linéaire (Section 20.2). Elles sont illustrées ci-après par un exemple (Section 20.3).

20.1. RÉGRESSION POLYNOMIALE

La régression polynomiale est une forme de régression linéaire multiple dans laquelle on prend en compte non seulement une ou plusieurs variables indépendantes

quantitatives ($X_1, X_2, \ldots, X_\omega$), mais aussi *au moins* l'une de ces variables élevée à une puissance supérieure à 1, comme par exemple dans le modèle ci-après :

$$y_i = a + b_1 x_{1_i} + b_2 x_{1_i}^2 + c_1 x_{2_i} + \varepsilon_i$$

où
a est la constante ;
x_{1_i}, la valeur de la variable X_1 pour l'individu i ;
b_1, le coefficient associé à X_1 ;
$x_{1_i}^2$, le carré de x_{1_i} ;
x_{2_i}, la valeur de la variable X_2 pour l'individu i ;
c_1, le coefficient associé à X_2 ;
ε_i, le terme d'erreur.

Une régression polynomiale est d'ordre *d*, où *d* est le degré le plus élevé associé à une variable dans le modèle. Une régression polynomiale peut ainsi être d'ordre 2 (comme dans l'exemple ci-dessus), d'ordre 3, d'ordre 4, etc., suivant le degré le plus élevé présent dans le modèle.

On distingue la régression polynomiale simple de la régression polynomiale multiple.

Dans la régression polynomiale simple, le modèle ne comprend qu'une seule variable indépendante *X* accompagnée (outre la constante et le terme d'erreur) de ses valeurs au carré, au cube, etc. Par exemple, un

modèle de régression polynomiale simple d'ordre 5 s'écrit :

$$y_i = a + b_1 x_{1_i} + b_2 x_{1_i}^2 + b_3 x_{1_i}^3 + b_4 x_{1_i}^4 + b_5 x_{1_i}^5 + \varepsilon_i$$

Dans la régression polynomiale multiple, il y a au moins deux variables indépendantes. Par exemple, un modèle de régression polynomiale multiple d'ordre 5 pourrait être :

$$y_i = a + b_1 x_{1_i} + b_2 x_{1_i}^2 + b_3 x_{1_i}^3 + b_4 x_{1_i}^4 + b_5 x_{1_i}^5 + c_1 x_{2_i} + \varepsilon_i$$

Bien évidemment, on peut aussi – s'il y a du sens à le faire (et si le nombre d'observations est suffisamment élevé) – prendre en compte des interactions entre variables indépendantes, par exemple :

$$y_i = a + b_1 x_{1_i} + b_2 x_{1_i}^2 + b_3 x_{1_i}^3 + b_4 x_{1_i}^4 + b_5 x_{1_i}^5 + c_1 x_{2_i} + d_1 x_{1_i} x_{2_i} + \varepsilon_i$$

De même, des variables qualitatives peuvent être introduites aussi, mais elles n'ont évidemment pas vocation à être élevées à une puissance supérieure à 1, ce qui n'aurait aucun sens.

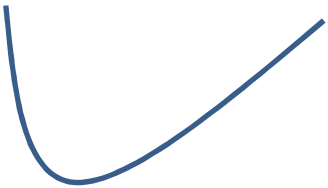
Dans la régression polynomiale multiple, il n'est pas indispensable que toutes les variables indépendantes

soient élevées au même degré. Seules les variables indépendantes dont la relation avec la variable dépendante est affectée par une non-linéarité justifient l'élévation à un degré supérieur à 1. De surcroit, en cas de non-linéarités impliquant plusieurs variables indépendantes, il n'y aucune raison *a priori* pour que ces non-linéarités soient toutes du même type et exigent une élévation à un même degré. Il appartient au chercheur d'adapter spécifiquement le modèle aux besoins du cas, et de s'en expliquer.

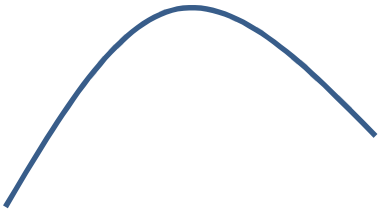
Le recours à la régression polynomiale se justifie lorsque les relations entre variables qui font l'objet de l'étude sont présumées non-linéaires dans le cadre théorique dans lequel s'inscrit l'étude, ou lorsque ces relations présentent des non-linéarités que le traitement par régression linéaire ne permet pas de prendre en compte (ce que révèle notamment une distribution non-aléatoire des résidus).

La mise en œuvre de la régression polynomiale suit la même procédure que celle de la régression linéaire multiple, les termes de puissance 2 et plus étant traités comme autant de variables indépendantes. Le choix du degré le plus élevé dépend du type de courbe que suggère le nuage de points. L'encadré 20.1 montre quelques exemples de ces types de courbes et quelques-uns des modèles correspondants.

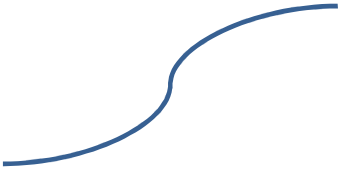
ENCADRÉ 20.1 – EXEMPLES DE COURBES DE MODÈLES POLYNOMIAUX



$Y = aX^2 + bX + c$
a,b,c>0

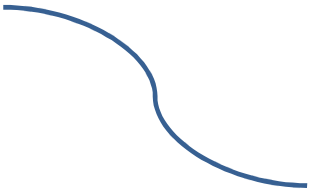


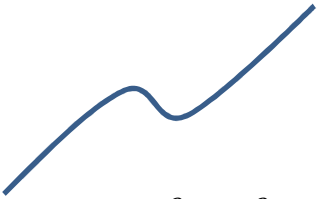
$Y = aX^2 + bX + c$
a<0 , b,c>0



$Y = aX^3 + bX^2 + cX + d$
a,b,c,d>0

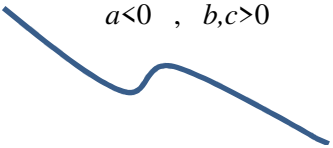
$Y = aX^3 + bX^2 + cX + d$
a,b,c<0 , d>0





$Y = aX^3 + bX^2 + c$
a,b,c>0

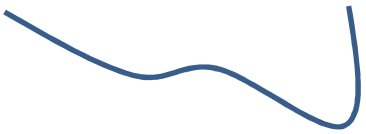
$Y = aX^3 + bX^2 + c$
a<0 , b,c>0



$Y = aX^3 + bX^2 + cX + d$
a,b<0 , c,d>0

$$Y = aX^5 + bX^4 + cX^3 + dX^2 + mX + n$$

a>0 , b,c,d,m,n<0



$$Y = aX^6 + bX^5 + cX^4 + dX^3 + mX^2 + nX$$

a>0 , b,c,d,m,n<0

Les conditions de validité de la régression polynomiale sont les mêmes que celles de la régression linéaire multiple[174].

L'approche pour interpréter les résultats dépend de l'objectif initial qui a conduit à opter pour la procédure de régression polynomiale.

Si la régression polynomiale a été choisie *a priori* comme étant la procédure la mieux adaptée au cadre théorique et à l'objet d'analyse, l'interprétation des résultats s'effectue de la même façon que pour une régression linéaire. L'attention est portée en priorité à la significativité des coefficients des variables indépendantes, quel que soit leur degré.

[174] La présence de multicolinéarité est cependant probable car on peut s'attendre à des corrélations entre les termes de degrés différents $(X, X^2, X^3, ...)$ d'une même variable X. Certains logiciels proposent une fonction « d'orthogonalisation des polynômes », qui permet de traiter ce problème.

Si au contraire la régression polynomiale a été entreprise pour pallier les insuffisances de la régression linéaire, un point crucial dans l'interprétation des résultats est de savoir si les variables indépendantes de degré 1 impliquées dans des non-linéarités et qui étaient non significatives dans la régression linéaire sont significatives dans la régression polynomiale. La significativité des termes de degrés supérieurs à 1 n'est pas le point d'intérêt majeur de l'interprétation. Cela étant, ces termes, s'ils sont significatifs eux-aussi, doivent être conservés dans le modèle puisqu'ils permettent de l'améliorer.

Lorsque la régression polynomiale ne permet pas d'améliorer la significativité des variables d'intérêt alors même que les conditions de validité du modèle sont remplies, on peut définitivement conclure que ces variables ne sont pas significatives.

En revanche, si les conditions de validité – et en particulier l'indépendance des résidus – ne sont pas remplies, on peut suspecter que le modèle polynomial ne permet pas de prendre convenablement en compte les non-linéarités. On peut alors tenter une régression non-linéaire.

20.2. RÉGRESSION NON-LINÉAIRE

Il importe de distinguer les concepts de *relation non-linéaire* et de *modèle de régression non-linéaire (*ou *modèle non-linéaire)*. Une relation est non-linéaire si elle est curviligne, au moins sur l'un de ses segments. Une relation non-linéaire peut être analysée par modèle de régression linéaire (au moins sur ceux de ses segments qui sont linéaires, comme indiqué au chapitre 17), ou par modèle de régression polynomiale. Comme indiqué en section 20.1, le modèle polynomial est lui-même une forme de modèle linéaire. Une relation non-linéaire peut également être analysée par modèle de régression non-linéaire. Un modèle de régression est non-linéaire lorsqu'il ne peut pas être écrit sous la forme

$$y_i = a + b_1 x_{1_i} + b_2 x_{2_i} + b_3 x_{3_i} + \cdots + b_\omega x_{\omega_i} + \varepsilon_i$$

où $x_{1_i}, x_{2_i}, \cdots, x_{\omega_i}$ sont les valeurs des ω variables indépendantes pour l'individu i, et $b_1, b_2, \cdots, b_\omega$ les paramètres associés à ces variables[175].

---

[175] En termes plus techniques, un modèle est dit « linéaire en ses paramètres » si, quel que soit le paramètre, une augmentation d'une unité de la valeur d'un paramètre entraine une variation Δ_{y_i} indépendante des paramètres du modèle. En d'autres termes, la dérivée partielle première de y_i par rapport à un paramètre ne dépend pas des paramètres du modèle. Par exemple, si le modèle s'écrit :

La régression non-linéaire[176] est simple ou multiple suivant que le modèle compte seulement une ou au contraire plusieurs variables indépendantes.

$$y_i = a + b_1 x_{1_i} + b_2 x_{2_i} + \varepsilon_i$$

et que le paramètre b_1 devient $(1 + b_1)$, le modèle devient :

$$y'_i = a + (1 + b_1) x_{1_i} + b_2 x_{2_i} + \varepsilon_i$$

La variation de y_i est :

$$\begin{aligned} \Delta_{y_i} &= y'_i - y_i \\ &= a + x_{1_i} + b_1 x_{1_i} + b_2 x_{2_i} + \varepsilon_i - a \\ &\quad - b_1 x_{1_i} - b_2 x_{2_i} - \varepsilon_i = x_{1_i} \end{aligned}$$

On voit que la variation de y_i, c'est-à-dire x_{1_i}, s'écrit sans référence aux paramètres du modèle. Au contraire, si un modèle est *non-linéaire en ses paramètres*, la variation de la variable réponse consécutive à une variation d'un paramètre est elle-même une fonction des paramètres du modèle. Par exemple si le modèle s'écrit :

$$y_i = b x_i^c$$

où les paramètres sont *b* et *c*, et que le paramètre *b* devient (1+*b*), le modèle devient :

$$y'_i = (1 + b) x_i^c$$

La variation de y_i est :

$$\Delta_{y_i} = y'_i - y_i = x_i^c + b x_i^c - b x_i^c = x_i^c$$

qui est une fonction du paramètre *c*.

[176] Accessible dans XLSTAT via la commande *Modélisation des données / Régression non linéaire*.

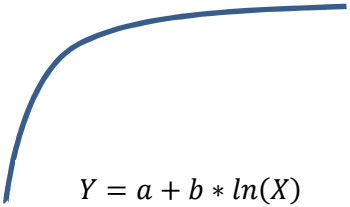
Le recours à la régression non-linéaire se justifie lorsque les relations entre variables qui font l'objet de l'étude sont présumées non-linéaires dans le cadre théorique dans lequel s'inscrit l'étude, ou lorsque ces relations présentent des non-linéarités que le traitement par régression linéaire ou polynomiale ne permet pas de prendre convenablement en compte (ce que révèle notamment une distribution non-aléatoire des résidus).

Il existe une infinité de modèles non-linéaires, mais le choix du modèle à ajuster n'est pas une question de hasard. Le choix d'un modèle non-linéaire peut s'effectuer en fonction du type de non-linéarité que suggère la représentation graphique des données à traiter. L'encadré 20.2 montre quelques exemples de modèles non-linéaires courants avec les allures de courbes correspondantes. A partir d'une idée de la courbe de la relation entre variables, le chercheur peut tenter d'ajuster le modèle correspondant[177]. Il n'y a pas nécessairement adéquation cependant : une courbe qui rappelle un modèle non-linéaire peut très bien être mieux analysable à partir d'un modèle polynomial.

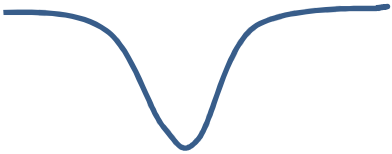
[177] XLSTAT permet d'effectuer des régressions non-linéaires à partir de la fonction *Modélisation des données / Régression non linéaire*.

ENCADRÉ 20.2 – EXEMPLES DE COURBES DE MODÈLES NON-LINÉAIRES

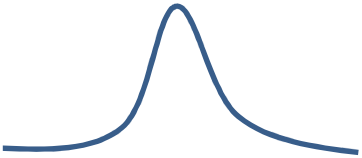
$Y = e^X$



$Y = a + b * ln(X)$
a,b>0



$Y = a * \left[1 - \left(b * e^{-c*X^2}\right)\right]$
a,b,c>0



$Y = a * \left[1 - \left(b * e^{-c*X^2}\right)\right]$
a<0 , b,c>0

$Y = \frac{ab + cX^d}{b + X^d}$,
$a, b, c, d > 0$

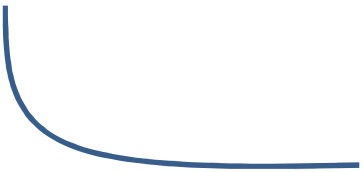
$Y = \frac{a}{1 + b * e^{-(cX + dX^2 + mX^3)}}$
a,b,c,d,m>0

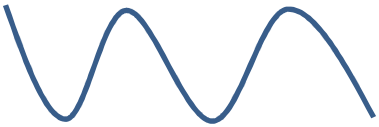
$$Y = \frac{a}{1 + b * e^{cX}}$$

$a,b>0$, $c<0$

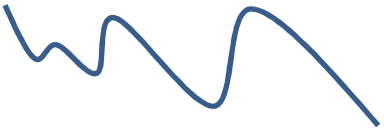
$$Y = \left(a^{1-b} - c * e^{-dX}\right)^{\frac{1}{1-b}}$$

a,b,c,d>0





$$\begin{cases} Y = a + sinus(X) \\ Y = a + cosinus(X) \end{cases}, \quad a > 0$$



$$a + b * cos(c * X) + d * sin(m * X) \quad , \quad a > 0$$

Les conditions de validité de la régression non-linéaire sont les mêmes que celles de la régression multiple.

Les résultats de la régression non-linéaire s'interprètent au regard de deux principaux critères :

- le coefficient de détermination R^2, qui indique la part de la variance de la variable dépendante qui est expliquée par les variables indépendantes. R^2 est compris entre 0 et 1. Plus R^2 est proche de 1, plus il y a de chances que les variables indépendantes retenues dans le modèle soient celles qui expliquent le mieux la variable dépendante ;

- l'erreur standard d'estimation, mesurée par la RMCE (racine de la moyenne des carrés des erreurs), qui indique la moyenne des résidus d'estimation. Plus faible est la RMCE, plus exactement le modèle reflète les relations entre variables indépendantes et variable dépendante.

Lorsque la régression non-linéaire a été entreprise pour pallier les insuffisances de la régression linéaire ou de la régression polynomiale, un aspect crucial de l'interprétation des résultats consiste à mesurer l'avantage (ou non) du recours à la régression non-linéaire. Il s'agit d'en comparer les résultats à ceux de la régression linéaire ou polynomiale préalablement tentée. La comparaison porte essentiellement sur trois points :

(a) le modèle non-linéaire permet-il, davantage que le modèle linéaire ou polynomial antérieur, d'expliquer la variance de la variable dépendante ? En d'autres termes, le coefficient de détermination R^2 est-il plus élevé avec le modèle non-linéaire qu'avec le modèle linéaire ou polynomial essayé au préalable ?

(b) l'erreur standard d'estimation (RMCE) est-elle plus faible avec le modèle non-linéaire qu'avec le modèle linéaire ou polynomial essayé au préalable ?

(c) la condition de distribution aléatoire des résidus est-elle mieux remplie avec le modèle non-linéaire qu'avec le modèle linéaire ou polynomial essayé au préalable ?

On peut conclure que le modèle non-linéaire est préférable si la réponse aux trois questions est positive.

20.3. EXEMPLE

On dispose de données sur le nombre d'heures de cours particuliers pris par 40 élèves pendant une année. On connait par ailleurs l'évolution de la moyenne générale de chacun de ces élèves entre le premier et le troisième trimestres (Tableau 20.1). On cherche à identifier l'éventuelle existence d'une relation entre les deux variables, et le cas échéant, la forme de cette relation.

Tableau 20.1.
Nombre d'heures de cours particuliers et évolution de la moyenne générale

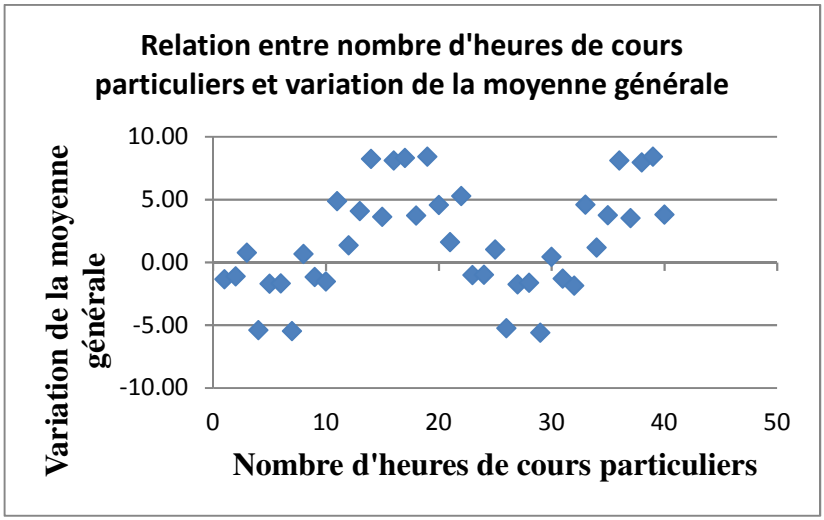
Identifiant	Nombre d'heures	Variation de la moyenne générale
1	29	-5,60
2	35	3,75
3	2	-1,12
4	19	8,40
5	36	8,09
6	21	1,61

Identifiant	Nombre d'heures	Variation de la moyenne générale
7	24	-1,00
8	3	0,78
9	17	8,29
10	13	4,08
11	27	-1,77
12	34	1,17
13	9	-1,18
14	6	-1,69
15	40	3,80
16	7	-5,47
17	12	1,35
18	18	3,73
19	20	4,57
20	26	-5,25
21	16	8,09
22	30	0,44
23	11	4,87
24	15	3,61
25	37	3,52
26	1	-1,36
27	4	-5,40
28	31	-1,29
29	14	8,23
30	5	-1,72
31	28	-1,62
32	38	7,94

Identifiant	Nombre d'heures	Variation de la moyenne générale
33	22	5,26
34	10	-1,52
35	23	-1,02
36	39	8,41
37	32	-1,85
38	8	0,66
39	33	4,60
40	25	1,01

1. Visualisation des données

Le nuage de points suggère qu'il existe une relation entre les deux variables, mais qu'elle est non-linéaire :



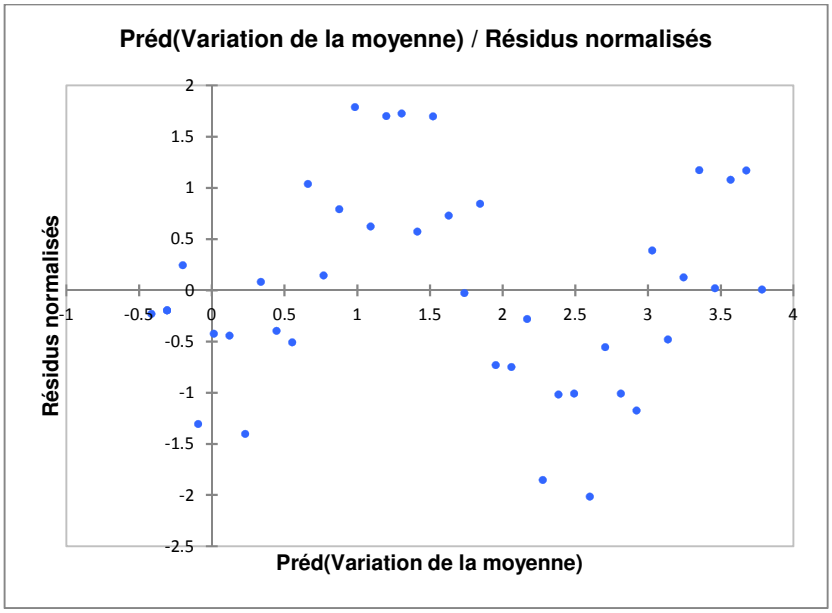
2. Modèle linéaire

On essaie d'abord d'ajuster aux données un modèle linéaire $y_i = a + bx_i + \varepsilon$.

Le tableau des paramètres de régression indique que la variable indépendante n'est pas significative :

Source	VP*	ES**	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Constante	-0,52	1,30	-0,39	0,69	-3,16	2,12
Nombre d'heures	0,10	0,05	1,93	0,06	-0,005	0,22

Le graphique Prédictions / Résidus normalisés montre que les résidus sont hétéroscédastiques et non-aléatoirement distribués :



Il est donc manifeste que le modèle linéaire ne permet pas de rendre compte de la relation entre les deux variables.

3. Modèle polynomial

On tente alors le modèle polynomial d'ordre 3 suivant :

$$y_i = a + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + \varepsilon$$

Le tableau des paramètres de régression montre que tous les termes indépendants sont significatifs :

Source	VP*	ES**	t	Pr > |t|	Borne inf. (95%)	Borne sup. (95%)
Constante	-7,54	2,46	-3,06	0,004	-12,54	-2,547
NbH (Nombre d'heures)	1,96	0,51	3,81	0,001	0,91	3,003
NbH^2	-0,10	0,02	-3,74	0,001	-0,16	-0,050
NBH^3	0,002	0,00	3,71	0,001	0,001	0,003

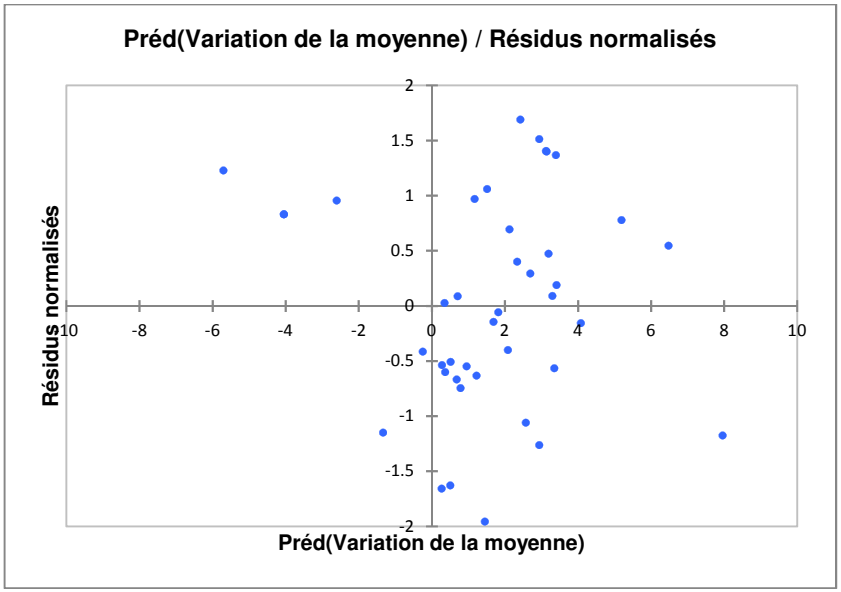
* VP : Valeur des paramètres
** ES : Erreur standard

Donc le nombre d'heures de cours particuliers pris apparait comme un facteur significatif de la variation de la moyenne générale. Reste cependant à vérifier les conditions de validité *a posteriori* du modèle.

Les résidus sont normalement distribués :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Résidus standardisés	0,472	0,601	0,613	0,528

Mais ils sont toujours hétéroscédastiques et non-aléatoirement distribués :



Le modèle polynomial représente une avancée par rapport au modèle linéaire, mais n'est pas pleinement satisfaisant. On se tourne donc une régression non-linéaire.

4. Modèle non-linéaire

Étant donné le caractère sinusoïdal que suggérait le nuage de points, on essaie d'ajuster le modèle non-linéaire suivant :

$$y_i = a + b * cosinus(c * x) + d * sinus(e * x)$$

où a, b, c, d, e sont les paramètres à estimer. L'estimation donne les paramètres suivants[178] :

Paramètre	Valeur	Erreur standard
a	1,671	0,701
b	-0,334	1,007
c	5,536	0,124
d	0,303	0,993
e	5,030	0,137

[178] L'estimation est effectuée sur XLSTAT. Avant de lancer les calculs, il faut d'abord :

(a) Déterminer les dérivées premières de la fonction y_i par rapport à chacun de ses paramètres (si besoin est, des calculateurs mathématiques gratuits en ligne permettent de calculer des dérivées, par exemple : https://www.solumaths.com/fr/calculatrice-en-ligne/calculer/deriver). Ces dérivées s'établissent ici comme suit :

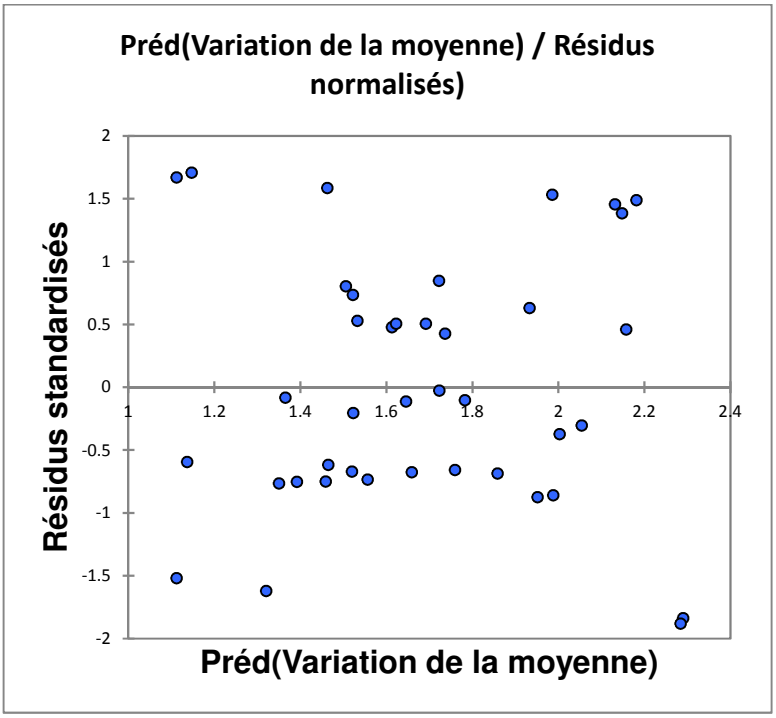
Dérivée de y_i par rapport à a =	1
Dérivée de y_i par rapport à b =	$Cos(c * x)$
Dérivée de y_i par rapport à c =	$-b * x * Sin(c * x)$
Dérivée de y_i par rapport à d =	$Sin(e * x)$
Dérivée de y_i par rapport à e =	$d * x * Cos(e * x)$

(b) fixer les valeurs initiales à partir desquelles les paramètres devront être estimés. La règle générale est de choisir des paramètres initiaux plausibles étant donné la relation à estimer. Les paramètres choisis ici sont : 1, 2, 3, 4, 5.

L'analyse des résidus montre que les résidus sont normalement distribués :

	Shapiro -Wilk	Anderson -Darling	Lilliefors	Jarque -Bera
Résidus standardisés	0,085	0,083	0,123	0,552

L'hétéroscédasticité a été sensiblement réduite :



Mais les résidus sont corrélés à la variable dépendante :

	Coefficients de corrélation linéaire	
	Variable dépendante prédite	Variable indépendante
Résidus standardisés	1 (p-value <0,0001)	0,094 (p-value = 0,56)

Enfin, on peut comparer les valeurs de R^2 et de RMCE obtenues pour les trois modèles :

	R^2	RMCE
Modèle linéaire	0,090	4,05
Modèle polynomial	0,345	3,53
Modèle non-linéaire	0,006	4,41

5. Conclusion

Le modèle linéaire apparait clairement inadapté aux données. Le modèle polynomial et le modèle non-

linéaire ne se départagent pas clairement au regard du critère d'indépendance des résidus par rapport à la variable dépendante. Le modèle non-linéaire respecte la condition d'homoscédasticité. En revanche, le modèle polynomial domine nettement le modèle non-linéaire en termes de contribution à l'explication de la variance de la variable dépendante (R^2), et en termes d'erreur moyenne d'estimation (RMCE). On peut donc considérer que c'est le meilleur des trois modèles. Il permet de conclure que le nombre d'heures de cours particuliers pris exerce un effet significatif (p-value=0,001) sur la variation de la moyenne générale. Mais cet effet est minoritaire : 67% de la variance de la moyenne générale dépendent d'autres facteurs que le nombre d'heures de cours particuliers. Du reste, l'effet mis en évidence reste à confirmer, la validité du modèle n'étant pas établie.

Chapitre 21. Régression logistique : analyser l'influence de facteurs quantitatifs et/ou qualitatifs sur une variable-réponse qualitative

La régression logistique vise à analyser l'influence d'une ou plusieurs variables indépendantes quantitatives et/ou qualitatives sur une unique variable dépendante qualitative (nominale ou ordinale, binaire ou polytomique).

L'objectif de la régression logistique est essentiellement de permettre de repérer l'existence de liens entre variables, de mesurer la significativité et le poids d'éventuels liens, et d'identifier leur sens (positif ou négatif).

En termes techniques, le modèle logistique établit une relation entre d'une part une combinaison linéaire des variables indépendantes, et d'autre part, la *probabilité* (ou plus précisément la *« cote » de la probabilité*) de réalisation de la modalité d'intérêt de la variable dépendante. On distingue différentes versions du modèle logistique suivant la fonction de probabilité utilisée pour calculer la probabilité de réalisation des

modalités. La version la plus fréquemment utilisée est le modèle « logit »[179], utilisé aussi dans ce chapitre[180].

---

[179] Soient des variables indépendantes $X_1, X_2, \dots, X_\omega$ et une variable dépendante *Y* qualitative. Le modèle logit est défini par :

$$\ln\left(\frac{P_{y_i}}{1-P_{y_i}}\right) = \alpha + \left(\beta_1 \times x_{1_i}\right) + \left(\beta_2 \times x_{2_i}\right) + \dots + \left(\beta_\omega \times x_{\omega_i}\right)$$

où

- P_{y_i} est la probabilité de réalisation de la modalité d'intérêt de la variable qualitative pour l'individu *i*. P_{y_i} est définie par :

$$P_{y_i} = \frac{e^{y_i}}{1+e^{y_i}}$$

où e^{y_i} est l'exponentielle de y_i ;

- $\frac{P_{y_i}}{1-P_{y_i}}$ est la « cote » (*odds*) de réalisation de la modalité d'intérêt pour l'individu *i* (c'est-à-dire le rapport entre la probabilité que la modalité se réalise et la probabilité qu'elle ne se réalise pas) ;
- $\ln\left(\frac{P_{y_i}}{1-P_{y_i}}\right)$ est le logarithme de la cote ou « logit de la probabilité P_{y_i} ».

Estimer le modèle consiste à calculer la valeur des coefficients $\alpha, \beta_1, \dots, \beta_\omega$.

[180] D'autres modèles existent, par exemple le modèle Probit, le modèle Log-log complémentaire, ou le modèle de Gompertz.

La régression logistique[181] est *dichotomique* (ou *binomiale*) si la variable dépendante est binaire, qu'elle soit nominale ou ordinale (section 21.2) ; *multinomiale* si la variable dépendante est nominale polytomique (section 21.3) ; et *ordinale* si la variable dépendante est ordinale polytomique (section 21.4). Quel que soit le type de régression logistique, des conditions de validité sont à respecter (section 21.1).

21.1. CONDITIONS DE VALIDITÉ

S'assurer que la nature des variables est adaptée à la procédure de régression logistique envisagée est évidemment primordial : les résultats seront contestables si, par exemple, la procédure de régression logistique multinomiale est utilisée alors que la variable dépendante est ordinale. D'autres précautions usuelles de base doivent également être prises, consistant notamment à assurer que les catégories d'une même variable qualitative sont mutuellement exclusives, et qu'un même individu ne peut pas figurer dans plus d'une catégorie d'une variable qualitative donnée (indépendance des observations).

Au-delà, cinq conditions essentielles doivent être remplies pour qu'une régression logistique soit valide.

---

[181] Accessible dans XLSTAT via la commande *Modélisation des données / Régression logistique*.

Quatre de ces conditions (absence de valeurs aberrantes, taille de l'échantillon, absence de multicolinéarité et effectif minimum des tableaux de contingence observé et théorique) se vérifient avant d'effectuer la régression. La cinquième condition (linéarité) se vérifie en cours de régression, de façon à affiner par étape la définition du modèle.

21.1.1. Conditions préalables

Avant même d'effectuer la régression logistique, on vérifie que ces conditions sont remplies, et on prend les mesures correctives (si possible) si ce n'est pas le cas. Selon ces conditions :

- Les variables indépendantes quantitatives ne doivent comporter aucune valeur aberrante, ni univariée ni multivariée[182]. L'inspection s'effectue pour l'échantillon pris tout ensemble ;

- l'échantillon doit être de taille suffisante. On sait déjà que les tests statistiques usuels nécessitent des échantillons d'une taille minimum de trente individus sans observation manquante. Ce minimum s'impose donc ici également. Mais il est évident que la taille de l'échantillon doit tenir compte aussi du nombre de variables à prendre en compte : plus le nombre de

[182] Voir section 15.3.1.

variables est élevé, plus la taille de l'échantillon doit être élevée. Il n'existe pas de règle théorique fixant de façon absolue la taille minimale de l'échantillon, mais une pratique admise consiste à fixer, par précaution, cette taille minimale à 10 ou 50 fois le nombre de variables indépendantes (y compris les interactions) suivant que la variable indépendante est qualitative ou quantitative :

Exemples de tailles minimales d'échantillons

Nombre de variables indépendantes et interactions			Taille minimale*
Qualitatives	Quantitatives	Interactions	
1 à 3	0	0	30 individus
4	0	0	40 individus
0	1	0	50 individus
1	1	0	60 individus
2	1	0	70 individus
3	1	0	80 individus
⋮	⋮	⋮	⋮
0	2	0	100 individus
1	2	0	110 individus
⋮	⋮	⋮	⋮

Nombre de variables indépendantes et interactions			Taille minimale*
Qualitatives	Quantitatives	Interactions	
0	3	0	150 individus
0	3	1 $interaction$: X_1X_2	200 individus
0	3	2 $interactions$: X_1X_2 ; X_1X_3	250 individus
0	3	3 $interactions$: X_1X_2 ; X_1X_3 ; X_2X_3	300 individus
1	3	3 $interactions$: X_1X_2 ; X_1X_3 ; X_2X_3	310 individus
⋮	⋮	⋮	⋮

* Sans observation manquante

- les variables indépendantes quantitatives ou qualitative ordinales doivent être exemptes de multicolinéarité[183] ;
- pour chaque paire de variables indépendantes qualitatives :
  - aucune case du tableau de contingence observé ne doit avoir un effectif nul ; et

[183] Voir section 15.2.

- le tableau de contingence théorique[184] ne doit pas contenir plus de 20% de cases ayant un effectif inférieur à 5[185]. Par exemple, soit deux variables indépendantes qualitatives décrivant la Tranche d'âge et le Niveau de diplôme dans un échantillon de 551 enseignants. Le tableau de contingence observé en croisant les deux variables est :

	Licence	Master	Doctorat	Total
31-40 ans	100	145	1	246
41-50 ans	115	85	2	202
51-60 ans	45	55	3	103
Total	260	285	6	551

Aucune des cases du tableau de contingence observé n'a un effectif nul. Le tableau de contingence théorique est :

	Licence	Master	Doctorat	Total
31-40 ans	116,08	127,24	2,68	246
41-50 ans	95,32	104,48	2,20	202
51-60 ans	48,60	53,28	1,12	103
Total	260	285	6	551

[184] XLSTAT fournit les tableaux de contingence observé et théorique de deux séries à partir de la commande *Créer un tableau de contingence* sous l'onglet *Préparation des données*.

[185] L'objectif est d'éviter de surestimer les statistiques de significativité fondées sur le Khi-carré : voir section 11.1.

On constate que 33% des cases (trois cases sur 9) du tableau de contingence théorique ont un effectif inférieur à 5. On y remédie en regroupant les colonnes Master et Doctorat dans le tableau de contingence observé[186] :

	Licence	Master et Doctorat	Total
31-40 ans	100	146	246
41-50 ans	115	87	202
51-60 ans	45	58	103
Total	260	291	551

Le tableau de contingence théorique devient :

	Licence	Master et Doctorat	Total
31-40 ans	116,08	129,92	246
41-50 ans	95,31	106,68	202
51-60 ans	48,60	54,39	103
Total	260	291	551

de sorte que la condition relative aux effectifs par case du tableau de contingence théorique est, elle aussi, remplie à présent.

[186] Un autre exemple en a été donné page 295.

21.1.2. Condition de linéarité

Il doit exister une relation linéaire entre chaque variable indépendante quantitative et la variable dépendante du modèle logit (c'est-à-dire la série des logits des probabilités associées aux individus de l'échantillon)[187]. Cette condition se vérifie en appliquant la transformation de Box-Tidwell, qui consiste à rajouter aux variables indépendantes du modèle, pour chaque variable indépendante quantitative, une variable « construite » : pour toute variable indépendante quantitative *X*, on ajoute une variable indépendante quantitative « *X* construite » telle que

$$X\ construite = X\ \times ln(X)$$

où *ln* désigne le logarithme népérien.

La variable construite est donc le produit de la variable *X* par son logarithme. La condition de linéarité est remplie si le coefficient de régression de la variable construite est non significatif. Si le coefficient est significatif, une transformation s'impose : *X* doit être remplacée par X^{λ} où

[187] Il ne s'agit pas ici de relation linéaire entre variable indépendante et variable dépendante (comme c'était par exemple le cas pour l'analyse de covariance), mais de relation linéaire entre variable indépendante et *logit* de la variable dépendante.

$$\lambda = 1 + \frac{\delta}{\beta}$$

avec

β représentant le coefficient de régression de *X* dans le modèle sans variable construite (le modèle initial) ; et

δ, le coefficient de régression de *X* construite.

Ainsi par exemple, si le modèle initial est représenté par :

$$ln\left(\frac{P_{y_i}}{1 - P_{y_i}}\right) = \alpha + \left(\beta_1 \times x_{1_i}\right) + \left(\beta_2 \times x_{2_i}\right)$$

avec X_1 variable qualitative et X_2 variable quantitative, on commence par estimer le modèle pour déterminer les valeurs des paramètres, ce qui fournit la forme calculable :

$$ln\left(\frac{P_{y_i}}{1 - P_{y_i}}\right) = \hat{\alpha} + \left(\hat{\beta}_1 \times x_{1_i}\right) + \left(\hat{\beta}_2 \times x_{2_i}\right) + \varepsilon$$

Puis on introduit dans le modèle la variable quantitative construite et on estime le nouveau modèle, ce qui donne :

$$ln\left(\frac{P_{y_i}}{1-P_{y_i}}\right) = \hat{\alpha}_{\succ} + \left(\hat{\beta}_{1\succ} \times x_{1_i}\right) + \left(\hat{\beta}_{2\succ} \times x_{2_i}\right) + \left[\hat{\delta} \times x_{2_i} \times ln\left(x_{2_i}\right)\right] + \varepsilon$$

où le signe $\succ$ désigne la nouvelle valeur du paramètre après introduction de la variable construite.

Si le paramètre $\hat{\delta}$ est non-significatif, la condition de linéarité est remplie. Sinon la variable X_2 doit être remplacée dans le modèle final par X_2^{λ} où

$$\lambda = 1 + \frac{\hat{\delta}}{\hat{\beta}_2}$$

21.2. RÉGRESSION LOGISTIQUE DICHOTOMIQUE

La régression logistique dichotomique s'utilise lorsque la variable réponse est nominale binaire. On code 1 la modalité d'intérêt (celle dont on cherche à connaitre les facteurs déterminants), et 0 l'autre.

Les éléments essentiels[188] pour interpréter les résultats d'une régression logistique sont :

---

[188] Les logiciels fournissent aussi des « pseudos R^2 » (R^2 de McFadden ; R^2 de Cox et Snell ; R^2 de Nagelkerke), mais ceux-ci n'expriment pas le pourcentage expliqué de la variance de la

- les paramètres de régression ;
- les effets marginaux ; et
- les coefficients standardisés.

Soit par exemple une régression logistique dichotomique de l'obtention d'une mention « très bien » au bac (variable dépendante, oui=1, non=0) sur la série (variable indépendante qualitative : L, ES, S) et la moyenne au contrôle continu annuel (variable indépendante quantitative). Les paramètres de régression, les effets marginaux et les coefficients standardisés se présentent comme suit :

Paramètres de régression					
	VP*	ES**	Khi² de Wald	Pr > Khi²	Odds ratio
Constante	-9,90	4,49	4,86	0,03	
Moyenne annuelle	0,79	0,35	5,17	0,02	2,21
Série-L	0,98	0,67	2,13	0,14	2,66
Série-ES	-1,73	0,84	4,25	0,04	0,18
Série-S	0,75	0,69	1,19	0,28	2,12

* VP : Valeur des paramètres
** ES : Erreur standard

variable dépendante, et par conséquent ne s'interprètent pas de la même façon que le R^2 d'une régression linéaire.

Effets marginaux à la moyenne						
Source	Effets marginaux	ES*	z	Pr > |z|	Borne inf. (95%)	Borne sup. (95%)
Moyenne annuelle	0,18	0,14	1,28	0,10	-0,10	0,47
Série-L	0,19	0,10	1,81	0,04	-0,02	0,39
Série-ES	-0,40	0,15	-2,5	1,00	-0,70	-0,10
Série-S	0,00					

* ES : Erreur standard

Coefficients standardisés				
	Valeur du paramètre	Erreur standard	Khi² de Wald	Pr > Khi²
Moyenne annuelle	0,75	0,33	5,17	0,02
Série-L	0,44	0,30	2,13	0,14
Série-ES	-0,78	0,38	4,25	0,04
Série-S	0,34	0,31	1,19	0,28

Le Khi-carré de Wald est l'équivalent des *t* de Student dans un tableau des paramètres de régression linéaire, et s'interprète de la même façon[189].

On voit ici que les deux variables indépendantes sont significatives.

Quant à la moyenne annuelle, le signe de son paramètre (0,79) étant positif, on peut dire que plus la moyenne annuelle est élevée, plus s'élève la probabilité d'obtention de la mention « très bien » au bac. Pour quantifier cet effet, on utilise l'effet marginal, présenté dans le tableau « Effets marginaux à la moyenne ». Dans ce tableau, l'effet marginal pour la variable indépendante « Moyenne annuelle » est de 0,18. Ce chiffre signifie qu'une augmentation d'un point de la moyenne annuelle augmente de 0,18 points de pourcentage la probabilité d'obtenir la mention « très bien » au bac. Par exemple, si la probabilité de mention est de 60% chez ceux qui ont 14 de moyenne annuelle, elle devient de 60,18% chez ceux qui ont 15 de moyenne.

Quant à la série, le signe du paramètre de la catégorie ES étant négatif, on peut dire que la probabilité

[189] Voir section 16.1. Le Khi-carré de Wald est égal au carré du rapport entre coefficient et erreur standard.

d'obtenir la mention « très bien » est plus faible pour les candidats de la série ES que pour les autres. Pour quantifier, on peut[190], ici également, se référer au

[190] Une manière alternative de formuler l'effet d'une variable qualitative est de se référer au « rapport des cotes » (*odds ratio*). Il importe de ne pas confondre les trois notions « *probabilités* », « *cotes* » (ou « chances relatives » ou « risques relatifs – *odds*) et « rapports de cotes ». Comparons par exemple les chances qu'ont les étudiants de deux Universités (A et B) d'être acceptés dans le programme Erasmus :

	Probabilité d'être sélectionné pour partir en Erasmus	Probabilité de ne pas être sélectionné	Cotes (odds)	Rapports des cotes (*odds ratio*)
Université A	90%	10%	0,90 / 0,10 = 9	9 / 1,5 = 6
Université B	60%	40%	0,60 / 0,40 = 1,5	1,5 / 9 = 0,16

Pour être sélectionné, la cote d'un candidat de l'Université A représente six fois celle d'un candidat de l'Université B, comme le montre le rapport des cotes. Les cotes des candidats de l'Université B représentent 16,66% de celles des candidats de l'Université A : elles leur sont de 83,34% inférieures ($0{,}1666 - 1 = -0{,}8334 = -83{,}34\%$). Dans un tableau des paramètres de régression logistique, le coefficient de régression est le logarithme du rapport des cotes, donc le rapport des cotes est l'exponentielle du coefficient de régression. Le rapport des cotes est parfois fourni avec les paramètres de régression

tableau des effets marginaux : le chiffre de -0,40 signifie que, par rapport aux candidats de la série S prise ici comme référence pour le calcul (ligne vide), la probabilité pour les candidats de ES d'obtenir la mention « très bien » est plus faible de 0,40 points de pourcentage. Ainsi par exemple, si la probabilité de mention est de 60% chez les candidats de la série S, elle sera de 59,6% (60% − 0,40%) chez ceux de la série ES. De même, la probabilité d'obtention de la mention est de 0,19 points de pourcentage plus élevée chez les candidats issus de L que chez ceux de la série S.

Enfin, le tableau des coefficients standardisés permet de classer les variables indépendantes en fonction de leur impact sur la variable réponse : on voit ici que le facteur ou modalité de facteur qui influence de la façon la plus décisive la probabilité d'obtention de la mention « très bien » est l'appartenance ou non à la série ES (paramètre standardisé = -0,78), immédiatement suivie par la moyenne annuelle (paramètre standardisé = 0,75).

(colonne Odds ratio). Dans le tableau des paramètres de régression présenté *supra*, l'odds ratio pour la série ES est de 0,18. Ce chiffre signifie que les cotes (chances relatives) des candidats de ES pour l'obtention de la mention « très bien » ne représentent que 18% des cotes des autres candidats, c'est-à-dire encore qu'elles sont de 82% plus faibles que les cotes des autres candidats ($0{,}18 - 1 = -0{,}82 = -82\%$).

La connaissance des paramètres estimés permet d'écrire le modèle sous forme calculable en termes de probabilités[191]. Ainsi, dans l'exemple présenté ici, le tableau des paramètres de régression indique que pour un élève *i*, la probabilité P_{y_i} d'obtenir une mention « très bien » au bac est égale à :

$$P_{y_i} = \frac{e^{y_i}}{1 + e^{y_i}}$$

avec
$y_i = -9{,}90 + (0{,}79 \times moyenne\ annuelle) + \beta_i$; et
$\beta_i = 0{,}98\ ,\ ou\ bien\ \beta_i = -1{,}73\ ,\ ou\ bien\ \beta_i = 0{,}75$
(suivant que *i* est en L, ou en ES, ou en S).

Ainsi par exemple, pour trois élèves issus des trois séries mais ayant tous la même moyenne annuelle de 12, les probabilités d'obtenir une mention « très bien » s'établiraient respectivement à 63%, 10% et 58% :

[191] On pourrait sans aucun doute écrire aussi le modèle sous forme calculable en termes de logits de probabilités, mais l'intérêt pratique n'en paraît pas vraiment évident.

	Élève 1 - Série L	Élève 2 - Série ES	Élève 3 - Série S
β_i	0,98	-1,73	0,75
Moyenne annuelle	12	12	12
y_i	0,56	-2,15	0,33
P_{y_i}	0,63	0,10	0,58

EXEMPLE 21.1

On dispose de données sur le sentiment de harcèlement scolaire dans un échantillon de 100 élèves (Tableau 21.1). On recherche l'effet de la taille du réseau d'amis et de l'appartenance ou non à une minorité ethnique sur le sentiment de harcèlement.

Tableau 21.1.
Taille du réseau d'amis, appartenance ou non à une minorité ethnique, et sentiment de harcèlement dans un échantillon de 100 élèves (Codage des colonnes 2 et 4 : 0=Non, 1=Oui)

Identifiant de l'élève (1)	Minorité ethnique (2)	Nombre d'amis (3)	Sentiment d'être harcelé(e) (4)
1	0	23	0
2	0	12	0
3	0	4	1

Identifiant de l'élève (1)	Minorité ethnique (2)	Nombre d'amis (3)	Sentiment d'être harcelé(e) (4)
4	1	12	1
5	0	6	1
6	0	29	0
7	0	2	0
8	0	13	0
9	1	5	1
10	1	7	1
11	1	8	1
12	0	30	0
13	0	20	0
14	0	30	0
15	1	21	1
16	0	12	0
17	1	15	1
18	0	3	0
19	1	25	1
20	0	20	0
21	0	14	1
22	0	6	1
23	0	5	0
24	0	1	0
25	1	12	1
26	1	23	1
27	0	25	0
28	1	5	0

Identifiant de l'élève (1)	Minorité ethnique (2)	Nombre d'amis (3)	Sentiment d'être harcelé(e) (4)
29	0	16	0
30	0	2	1
31	0	21	0
32	1	25	0
33	1	8	1
34	0	16	0
35	1	8	1
36	1	19	1
37	0	13	0
38	1	2	0
39	1	12	1
40	1	11	1
41	0	28	0
42	1	25	0
43	1	1	1
44	0	21	0
45	1	1	1
46	1	10	1
47	1	2	0
48	0	2	1
49	0	12	0
50	1	25	1
51	0	4	0
52	1	19	0
53	0	18	0

Identifiant de l'élève (1)	Minorité ethnique (2)	Nombre d'amis (3)	Sentiment d'être harcelé(e) (4)
54	1	9	1
55	1	2	1
56	0	11	0
57	0	1	1
58	1	24	0
59	0	25	0
60	1	4	0
61	1	19	1
62	0	12	1
63	0	10	0
64	0	4	1
65	0	0	1
66	1	12	1
67	0	19	0
68	1	12	1
69	1	19	0
70	0	10	0
71	1	22	0
72	0	25	0
73	0	22	0
74	1	28	1
75	0	23	0
76	0	22	0
77	1	26	1
78	0	25	0

Identifiant de l'élève (1)	Minorité ethnique (2)	Nombre d'amis (3)	Sentiment d'être harcelé(e) (4)
79	0	17	0
80	1	29	1
81	1	4	1
82	0	13	0
83	1	8	1
84	1	21	0
85	0	16	0
86	0	12	0
87	0	15	1
88	1	13	1
89	1	22	0
90	0	6	0
91	1	24	0
92	0	15	0
93	0	5	1
94	0	6	0
95	0	7	1
96	0	25	0
97	1	9	1
98	1	23	1
99	0	13	0
100	1	19	0

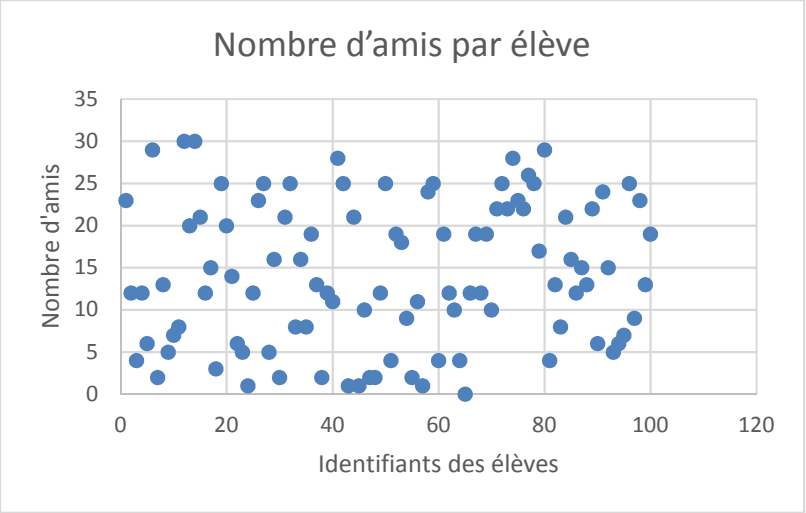
1. Vérification des conditions préalables

- Valeurs aberrantes parmi les variables indépendantes quantitatives

La variable Nombre d'amis est la seule variable indépendante quantitative. Les tests de normalité montrent qu'elle ne suit pas une distribution gaussienne :

	Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
Nombre d'amis	0,00	0,00	0,03	0,05

Donc appliquer le test de Grubbs ne serait pas éclairant. On se repose alors sur l'examen du nuage de points :



Aucune valeur ne se détache comme manifestement aberrante.

- Taille d'échantillon

L'échantillon comporte 100 observations pour deux variables indépendantes dont une quantitative. Il est donc de taille suffisante.

- Multicolinéarité des variables indépendantes quantitatives

Il n'y a qu'une seule variable indépendante quantitative ou ordinale, donc il n'y a pas de risque de multicolinéarité.

- Tableaux de contingence

Il n'y a qu'une seule variable indépendante qualitative, donc pas de tableau de contingence.

2. Mise en œuvre de la régression logistique dichotomique

2.1. Estimation initiale du modèle logit

Paramètres de régression

	Valeur	ES*	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds Ratio
Const.	0,14	0,47	0,08	0,76	-0,78	1,06	
Nbre d'amis	-0,11	0,03	11,4	0,001	-0,17	-0,04	0,89
Min. ethn.-0	0,00	0,00					
Min. ethn.-1	2,39	0,53	19,9	<0,0001	1,34	3,44	10,96

* ES : Erreur standard

On constate que les paramètres des deux variables indépendantes sont significatifs. Avant toute interprétation, il importe de s'assurer de la validité de ce résultat en vérifiant que la condition de linéarité est remplie.

2.2. Vérification de la condition de linéarité entre les variables indépendantes quantitatives et le logit de la variable réponse

On introduit la variable indépendante quantitative construite. Les observations à modéliser deviennent :

Identifiant de l'élève	Minorité ethnique	Nombre d'amis	NbAm * ln(NbAm)	Sentiment d'être harcelé(e)
1	0	23	72,12	0
2	0	12	29,82	0
3	0	4	5,55	1
⋮	⋮	⋮	⋮	⋮
98	1	23	72,12	1
99	0	13	33,34	0
100	1	19	55,94	0

On estime à nouveau le modèle, qui comprend à présent la variable quantitative indépendante construite. Le tableau des paramètres de régression indique :

Paramètres de régression

	VP*	ES**	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds Ratio
Constante	0,19	0,93	0,04	0,84	-1,64	2,02	
Nombre d'amis	-0,13	0,31	0,18	0,67	-0,73	0,47	0,88
NbAm * ln(NbAm)	0,01	0,09	0,00	0,95	-0,17	0,18	1,01
Minorité ethnique-0	0,00	0,00					
Minorité ethnique-1	2,39	0,54	19,93	<0,0001	1,34	3,45	10,96

* VP : Valeur des paramètres
** ES : Erreur standard

Le coefficient de la variable construite est non-significatif, donc la condition de linéarité est remplie[192]. Les paramètres de l'estimation initiale sont donc valides, et peuvent à présent être interprétés.

[192] Si le coefficient de la variable construite (0,01) avait été significatif, il aurait fallu remplacer chaque valeur x de la variable Nombre d'amis par $x^{0,9}$ dans le modèle initial (car $1 + \frac{0,01}{-0,11} = 0,9$).

2.3. Tableaux de résultats

Paramètres de la régression initiale

	VP*	ES**	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds Ratio
Cst.	0,14	0,47	0,08	0,76	-0,78	1,06	
Nbre amis	-0,11	0,03	11,4	0,001	-0,17	-0,04	0,89
Min. ethn. -0	0,00	0,00					
Min. ethn. -1	2,39	0,53	19,9	<0,0001	1,34	3,44	10,9

* VP : Valeur des paramètres
** ES : Erreur standard

Effets marginaux à la moyenne

	Effets marginaux	Erreur standard	z	Pr > |z|	Borne inf. (95%)	Borne sup. (95%)
Nbre amis	-0,02	0,12	-0,20	0,58	-0,27	0,22
Min. ethn.-0	0,000					
Min. ethn.-1	0,53	0,09	5,54	<0,0001	0,34	0,72

Coefficients normalisés

	VP*	ES**	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds Ratio
Nbre amis	-0,51	0,15	11,40	0,001	-0,80	-0,21	-0,51
Min. ethn.-0	0,00	0,00					0,00
Min. ethn.-1	0,65	0,14	19,92	<0,0001	0,36	0,94	0,65

* VP : Valeur des paramètres
** ES : Erreur standard

2.4. Interprétation des résultats

Le tableau des paramètres de régression montre que le nombre d'amis et l'appartenance à une minorité ethnique ont chacun un effet significatif (p-values inférieures à 0,05) sur la probabilité d'éprouver un sentiment de harcèlement. Plus le nombre d'amis est faible, plus la probabilité de se sentir harcelé(e) est élevée (signe négatif du coefficient). Le tableau des effets marginaux indique que quand le nombre d'amis diminue d'une unité, la probabilité de se sentir harcelé(e) augmente de 0,027 points de pourcentage. Il montre aussi que la probabilité de se sentir harcelé(e) est plus forte de 0,532 points de pourcentage chez les élèves qui appartiennent à une minorité ethnique que

chez les autres élèves. Le rapport des cotes (*odds ratio*) est de 10,965 pour les élèves issus de minorités ethniques, ce qui signifie qu'ils ont 996,5% plus de risques relatifs[193] que les autres de se sentir harcelés (10,965 – 1 = 9,965 = 996,5%). Enfin, le tableau des coefficients normalisés montre que, des deux facteurs présents dans le modèle, l'appartenance à une minorité ethnique est celui qui joue le rôle le plus déterminant (coefficient normalisé de 0,65 contre 0,51 en valeur absolue pour le nombre d'amis) dans la probabilité du sentiment de harcèlement.

EXEMPLE 21.2

On dispose de données sur les choix de carrière de 202 étudiants de premier cycle universitaire (Tableau 21.2). Les données renseignent sur :

(a) la préférence – ou non – de l'étudiant pour des activités bénéficiant à la communauté (C) ;

(b) la préférence – ou non – de l'étudiant pour un métier intellectuellement stimulant (I) ;

(c) l'opinion de l'étudiant sur le niveau minimum de prestige social que doit présenter une profession pour être choisie (P, sur une échelle de 1 à 100) ;

[193] Probabilité de se sentir harcelé(e) divisée par probabilité de ne pas sentir harcelé(e).

(d) le niveau minimum de rémunération jugé acceptable par l'étudiant (R, en EUR, en moyenne sur une période de référence) ; et

(e) la profession que vise l'étudiant.

On s'intéresse plus particulièrement au choix d'une carrière dans l'enseignement (E), et on cherche à en identifier les déterminants.

Tableau 21.2.
Choix de carrière et motivations chez 202 étudiants de premier cycle universitaire (Codage des colonnes 2, 3 et 6 : 0=Non, 1=Oui)

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
1	1	1	34	1639	1
2	0	0	27	1815	0
3	0	0	43	2037	0
4	1	1	30	1196	1
5	0	0	64	2935	0
6	1	1	27	1722	1
7	1	1	35	1966	1
8	0	0	33	1907	0
9	0	0	35	1969	0
10	0	0	36	1494	0
11	1	1	38	1598	1
12	0	0	65	2312	0

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
13	0	0	39	1016	1
14	0	0	80	2525	0
15	1	1	22	1505	1
16	1	1	34	1387	1
17	1	0	37	1544	1
18	0	0	32	1787	1
19	0	0	53	3101	0
20	0	0	53	2292	0
21	0	1	35	1461	1
22	0	0	50	2534	0
23	0	1	58	2607	0
24	0	1	24	1736	1
25	0	1	59	2200	0
26	0	0	72	2565	0
27	0	1	66	2045	0
28	0	0	54	2743	0
29	1	1	18	1789	1
30	0	1	56	2022	0
31	0	1	69	2839	0
32	0	1	43	2271	0
33	0	0	44	3320	0
34	1	1	30	1880	1
35	1	0	36	2611	1
36	1	1	47	3401	0
37	1	1	37	2996	1
38	0	1	58	2850	0

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
39	0	0	53	3028	0
40	0	1	57	2301	0
41	1	0	49	2327	0
42	0	1	43	2444	0
43	0	1	54	2873	0
44	1	0	73	2895	0
45	0	1	48	2465	0
46	1	0	46	2497	0
47	0	0	58	2530	0
48	1	1	60	2196	0
49	0	0	55	2238	0
50	1	1	42	2508	0
51	1	1	25	2097	1
52	1	0	60	2006	0
53	1	1	32	1518	1
54	0	0	44	2032	0
55	0	0	54	2130	0
56	0	1	51	2287	0
57	1	1	35	1944	1
58	1	1	39	1439	1
59	0	0	63	2263	0
60	0	0	80	2433	0
61	0	0	56	2464	0
62	1	1	31	1238	1
63	0	0	49	2983	0
64	0	0	52	2166	0

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
65	0	0	64	1992	0
66	0	0	43	2459	0
67	0	1	73	2407	0
68	0	1	32	1130	1
69	0	0	47	2440	0
70	0	0	50	3054	0
71	0	0	44	3214	0
72	1	1	29	1239	1
73	1	1	21	1574	1
74	1	0	41	2985	0
75	1	0	40	1563	1
76	0	0	81	3496	0
77	0	0	57	3153	0
78	0	1	57	2145	0
79	1	1	29	1901	1
80	0	0	69	2605	0
81	1	0	44	2877	0
82	0	0	55	2701	0
83	1	1	38	1945	1
84	0	0	51	2616	0
85	1	1	7	442	1
86	1	1	16	1306	1
87	0	1	55	3125	0
88	0	0	66	2375	0
89	1	1	16	1417	1
90	0	0	49	2648	0

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
91	1	1	36	1574	1
92	1	1	15	1221	1
93	0	0	60	2721	0
94	0	0	68	2465	0
95	0	0	49	2472	0
96	0	0	42	2024	0
97	0	0	66	2383	0
98	0	1	52	2609	0
99	0	0	42	2566	0
100	0	0	70	2705	0
101	1	1	49	2841	0
102	1	1	26	1595	1
103	0	0	53	2029	0
104	1	1	31	1885	1
105	0	1	69	2208	0
106	1	1	38	1903	1
107	0	0	43	2321	0
108	1	1	36	1548	1
109	0	1	55	2422	0
110	1	1	40	1891	1
111	1	1	65	2117	0
112	1	1	29	1567	1
113	1	0	63	2741	0
114	0	0	60	2190	0
115	0	0	54	2079	0
116	0	0	56	2205	0

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
117	1	1	55	1699	1
118	1	0	51	2766	0
119	0	0	58	3388	0
120	1	1	59	2687	0
121	1	0	61	2170	0
122	1	1	47	2050	0
123	1	1	49	445	1
124	1	0	50	2239	0
125	0	0	44	2898	0
126	1	0	48	2045	0
127	0	0	56	2284	0
128	0	0	57	2173	0
129	0	0	62	2657	0
130	1	1	44	2928	0
131	1	1	60	1991	1
132	0	0	58	3282	0
133	1	1	27	1443	1
134	0	1	26	1316	1
135	1	0	58	2083	0
136	0	1	56	2443	0
137	0	1	44	3422	0
138	0	1	48	2532	0
139	1	1	28	1425	1
140	0	1	26	1640	1
141	0	0	58	3110	0
142	1	0	75	2325	0

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
143	0	0	74	2159	0
144	0	0	66	2387	0
145	1	1	34	1987	1
146	0	0	69	2412	0
147	0	1	56	2282	0
148	1	0	58	2327	0
149	1	1	37	1923	1
150	0	1	41	3497	0
151	0	1	56	2171	0
152	0	1	64	2703	0
153	1	1	36	1064	1
154	0	0	82	3424	0
155	1	0	44	2273	0
156	0	0	46	2892	0
157	1	1	20	1976	1
158	0	0	48	2368	0
159	1	0	31	1583	1
160	1	1	8	1847	1
161	1	0	32	1719	1
162	0	0	48	2708	0
163	0	0	69	2933	0
164	1	1	39	1862	1
165	1	1	59	2057	0
166	1	1	34	1672	1
167	1	1	36	1885	1
168	0	0	44	2358	0

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
169	0	0	72	3115	0
170	1	1	31	1901	1
171	1	0	53	2295	0
172	0	0	54	2565	0
173	1	1	53	2626	0
174	1	1	36	1340	1
175	1	0	45	2458	0
176	1	0	77	2879	0
177	0	1	58	3461	0
178	1	1	14	1467	1
179	0	1	54	2063	0
180	0	1	24	1791	1
181	1	0	42	2366	0
182	0	0	54	2813	0
183	0	1	56	2195	0
184	1	1	18	1071	1
185	0	0	55	2653	0
186	1	1	32	1447	1
187	0	1	56	2735	0
188	1	1	29	1703	1
189	1	0	40	1738	1
190	0	0	25	1965	1
191	1	1	40	2440	0
192	0	0	42	2428	0
193	0	1	44	2526	0
194	0	1	37	1844	1

Identifiant (1)	C (2)	I (3)	P (4)	R (5)	E (6)
195	1	1	75	2072	0
196	1	1	19	1938	1
197	1	1	33	1299	1
198	1	1	36	1187	1
199	1	1	41	2522	0
200	1	1	52	2257	0
201	1	0	36	1908	1
202	0	0	60	1938	0

1. Vérification des conditions préalables

1.1. Condition d'absence de valeur aberrante parmi les variables indépendantes quantitatives

1.1.1. On vérifie l'absence de valeur aberrante univariée au moyen du test de Grubbs. Le test nécessite que la distribution soit gaussienne.

- Pour la variable Prestige social

  - Les tests de normalité sont concluants :

p			
Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
0,46	0,31	0,04	0,50

- Le test de Grubbs montre qu'il n'y a pas de valeur aberrante : l'hypothèse nulle du test est qu'il n'y a pas de valeur aberrante, l'hypothèse alternative que le maximum ou le minimum est une valeur aberrante. La p-value s'établit à 0,20 c'est-à-dire au-dessus du seuil de significativité de 5%.

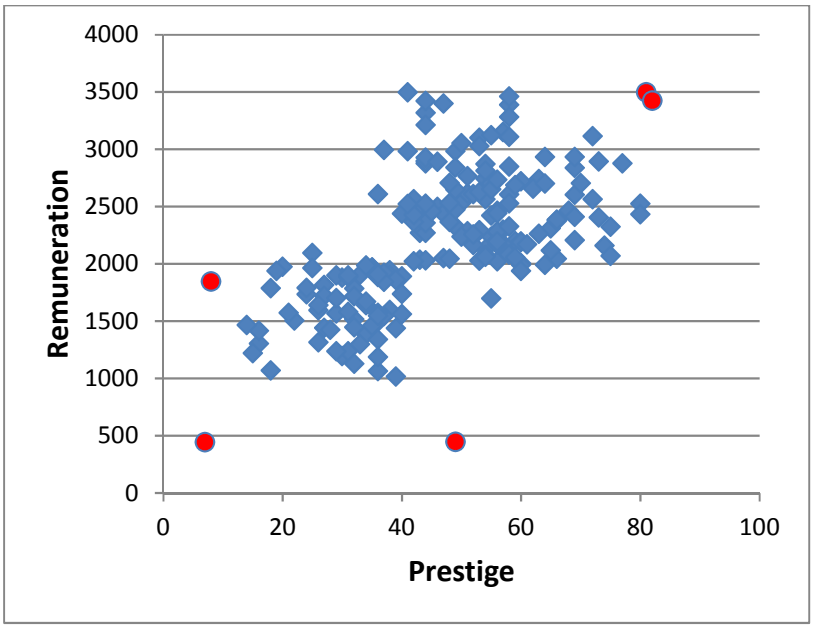
- Pour la variable Rémunération

- Les tests de normalité sont également concluants :

p			
Shapiro-Wilk	Anderson-Darling	Lilliefors	Jarque-Bera
0,49	0,94	0,87	0,76

- Le test de Grubbs montre qu'il n'y a pas de valeur aberrante. Comme précédemment, l'hypothèse nulle du test est qu'il n'y a pas de valeur aberrante, l'hypothèse alternative que le maximum ou le minimum est une valeur aberrante. Le seuil de significativité est de 5%. La p-value s'établit à 0,44.

1.1.2. On vérifie l'absence de valeur aberrante multivariée au moyen du nuage de points :



Cinq points apparaissent particulièrement excentrés, correspondant aux étudiant-e-s 76, 85, 123, 154 et 160. Le test de la distance de Mahalanobis montre qu'au seuil de significativité de 0,001, seul le point 123 constitue une valeur aberrante (distance = 15,33 avec p-value = 0,0004). Ce point est donc retiré de l'échantillon pour la suite de l'analyse.

1.2. L'échantillon comporte donc 201 individus pour quatre variables explicatives, et est donc de taille suffisante.

1.3. On vérifie l'absence de multicolinéarité entre variables indépendantes quantitatives en calculant le facteur d'inflation de variance :

	Prestige	Rémunération
Tolérance	0,61	0,61
VIF	1,64	1,64

La valeur du VIF s'établit autour de 1. Donc il n'y a pas de multicolinéarité.

1.4. On vérifie les effectifs par case des tableaux de contingence observé et théorique des variables indépendantes qualitatives :

Tableau de contingence observé
(Communauté / Intellectuel)

	0	1	Total
0	73	37	110
1	26	65	91
Total	99	102	201

Tableau de contingence théorique
(Communauté / Intellectuel)

	0	1	Total
0	54,18	55,82	110
1	44,82	46,18	91
Total	99	102	201

On constate qu'aucun effectif observé n'est nul et que tous les effectifs théoriques sont supérieurs à 5. La condition relative aux effectifs par case des tableaux de contingence observé et théorique est donc remplie.

2. Mise en œuvre de la régression logistique dichotomique

2.1. Estimation initiale du modèle logit

Paramètres du modèle							
	VP*	ES**	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds Ratio
Cte	17,54	4,224	17,26	<0,0001	9,27	25,82	
P	-0,26	0,076	12,01	0,0005	-0,41	-0,11	0,76
R	-0,005	0,001	12,65	0,0003	-0,008	-0,002	0,99
C-0	0,00	0,00					
C-1	4,32	1,56	7,66	0,0056	1,26	7,38	75,47
I-0	0,00	0,00					
I-1	2,30	1,08	4,49	0,0340	0,17	4,42	9,97

* VP : Valeur des paramètres
** ES : Erreur standard

On constate que tous les paramètres sont significatifs. Avant toute interprétation, il importe de s'assurer de la validité de ce résultat en vérifiant que la condition de linéarité est remplie.

2.2. Vérification de la condition de linéarité entre les variables indépendantes quantitatives et le logit de la variable réponse

On introduit les variables indépendantes quantitatives construites. Les observations à modéliser deviennent :

Id.	C	I	P	P*ln(P)	R	R*ln(R)	E
1	1	1	34	119,89	1639	12131,61	1
2	0	0	27	88,98	1815	13619,47	0
3	0	0	43	161,73	2037	15520,37	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
201	1	0	36	129,00	1908	14412,67	1
202	0	0	60	245,66	1938	14669,52	0

On estime à nouveau le modèle, qui comprend à présent les variables construites. Le tableau des paramètres de régression indique[194] :

[194] Il est possible que la corrélation entre chaque facteur quantitatif et sa version construite soit trop élevée, ce qui peut empêcher le calcul des paramètres. Sous XLSTAT, on autorise le calcul pour des variables corrélées dans la fenêtre *Options*, en faisant passer la *tolérance* de son niveau standard 0,001 à un niveau plus faible, par exemple 0,00001.

Source	VP*	ES**	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds ratio
Cte	83,36	34,74	5,75	0,01	15,25	151,46	
P	-3,43	1,88	3,32	0,06	-7,13	0,25	0,03
P * ln(P)	0,66	0,38	2,95	0,08	-0,09	1,41	1,93
R	-0,15	0,09	2,57	0,10	-0,34	0,03	0,85
R * ln(R)	0,01	0,01	2,43	0,11	-0,004	0,03	1,01
C-0	0,00	0,00					
C-1	5,76	2,14	7,22	0,00	1,56	9,97	320,28
I-0	0,00	0,00					
I-1	2,50	1,34	3,48	0,06	-0,12	5,13	12,24

* VP : Valeur des paramètres
** ES : Erreur standard

Les coefficients des variables construites sont non-significatifs, donc la condition de linéarité est remplie[195]. Les paramètres de l'estimation initiale sont donc valides, et peuvent à présent être interprétés.

[195] Si les coefficients des variables construites avaient été significatifs, par exemple si le coefficient de la variable Prestige construite (0,66) avait été significatif, il aurait fallu remplacer chaque valeur x de la variable Prestige par $x^{-1,5}$ dans le modèle initial ($1 + \frac{0,66}{-0,264} = -1,5$).

2.3. Interprétation des résultats

Paramètres du modèle							
	VP*	ES**	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds Ratio
Cte	17,54	4,224	17,26	<0,0001	9,27	25,82	
P	-0,26	0,076	12,01	0,0005	-0,41	-0,11	0,76
R	-0,005	0,001	12,65	0,0003	-0,008	-0,002	0,99
C-0	0,00	0,00					
C-1	4,32	1,56	7,66	0,0056	1,26	7,38	75,47
I-0	0,00	0,00					
I-1	2,30	1,08	4,49	0,0340	0,17	4,42	9,97

* VP : Valeur des paramètres
** ES : Erreur standard

Effets marginaux à la moyenne						
	Effet marginal	ES*	z	Pr > |z|	Borne inf. (95%)	Borne sup. (95%)
P	-0,0075	0,058	-0,130	0,552	-0,121	0,106
R	-0,0001	0,053	-0,003	0,501	-0,104	0,104
C-0	0,000					
C-1	0,2402	0,117	2,051	0,020	0,011	0,470
I-0	0,000					
I-1	0,0767	0,072	1,070	0,142	-0,064	0,218

* ES : Erreur standard

Coefficients normalisés						
	Valeur	ES*	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)
P	-2,204	0,63	12,01	0,0005	-3,450	-0,95
R	-1,705	0,47	12,65	0,0003	-2,643	-0,76
C-0	0,000	0,00				
C-1	1,187	0,42	7,664	0,0056	0,347	2,027
I-0	0,000	0,00				
I-1	0,634	0,29	4,490	0,0340	0,048	1,220

* ES : Erreur standard

Il apparait que chacun des quatre facteurs sur lesquels des données ont été collectées a un effet significatif sur la probabilité d'opter pour une carrière dans l'enseignement.

La préférence ou non pour des activités bénéficiant à la communauté, tout d'abord, différencie significativement ceux qui optent pour l'enseignement et les autres. Le tableau des effets marginaux indique que la probabilité d'opter pour l'enseignement est plus forte de 0,24 points de pourcentage chez les étudiants qui expriment une telle préférence que chez les autres. Le rapport des cotes (*odds ratio*) est de 75,47 chez les étudiants préférant des activités bénéficiant a la communauté, ce qui signifie qu'il y a 7447% plus de chances relatives[196] d'opter pour l'enseignement (75,47 – 1 = 74,47 = 7447%) chez eux que chez les autres.

La préférence ou non pour un métier intellectuellement stimulant exerce elle aussi un effet discriminant. Ceux qui préfèrent un métier intellectuellement stimulant ont une probabilité d'opter pour l'enseignement plus forte de 0,07. Le rapport des cotes (*odds ratio*) est de 9,975 chez les étudiants préférant un métier

[196] Probabilité d'opter pour l'enseignement divisée par probabilité de ne pas opter pour l'enseignement.

intellectuellement stimulant, ce qui signifie qu'il y a 897,5% plus de chances relatives[197] d'opter pour l'enseignement (9,975 – 1 = 8,975 = 897,5%) chez eux que chez les autres.

L'opinion de l'étudiant sur le niveau minimum de prestige social que doit présenter une profession pour être choisie exerce aussi un effet significatif sur la probabilité d'opter pour l'enseignement. Mais l'effet ici est négatif : plus la probabilité d'opter pour l'enseignement augmente, plus faible est le niveau minimum de prestige que l'étudiant exige d'une profession (sur une échelle de 1 à 100, quand le niveau de prestige attendu augmente de 1, la probabilité d'opter pour l'enseignement diminue de 0,0075 point de pourcentage).

De même, l'opinion de l'étudiant sur le niveau de rémunération acceptable exerce aussi un effet significatif négatif sur la probabilité d'opter pour l'enseignement : plus la probabilité d'opter pour l'enseignement augmente, plus faible est le niveau de rémunération que l'étudiant est disposé à accepter (quand le niveau de rémunération acceptable augmente d'une unité, la probabilité d'opter pour l'enseignement diminue de 0,0001 point de pourcentage).

---

[197] Probabilité d'opter pour l'enseignement divisée par probabilité de ne pas opter pour l'enseignement.

Enfin, le tableau des coefficients normalisés montre que, des quatre facteurs présents dans le modèle, le prestige attendu d'une profession est le facteur qui joue le rôle le plus déterminant (coefficient normalisé de 2,204 en valeur absolue) dans la probabilité d'opter pour l'enseignement. Viennent ensuite, par ordre d'importance décroissante, la rémunération attendue, la préférence ou non pour des activités bénéficiant à la communauté, et enfin la préférence ou non pour un métier intellectuellement stimulant.

21.3. RÉGRESSION LOGISTIQUE MULTINOMIALE

Dans la régression logistique multinomiale, la variable qualitative dépendante est nominale et comporte plus de deux modalités. Les conditions de validité de la régression multinomiale sont les conditions générales de la régression logistique[198]. L'interprétation des

[198] Cependant, si c'est le modèle logit qui est utilisé, une condition supplémentaire doit être remplie : la condition de *non-dépendance aux alternatives non-pertinentes – NDANP (Independence from irrelevant alternatives)*. La condition de NDANP exige que la probabilité pour un individu de préférer une modalité A par rapport à une modalité B soit indépendante de toute autre modalité C. Suivant un exemple classique de la littérature du domaine, si l'individu a connaissance des options de transport « voiture personnelle » et « bus publics – ligne

résultats s'effectue à partir des tableaux des paramètres de régression.

EXEMPLE 21.3

On dispose de données sur les caractéristiques de candidats à la validation des acquis de l'expérience (VAE) et sur leur choix de canal de validation, c'est-à-dire d'organisme certificateur (Tableau 21.3). Les caractéristiques des candidats couvrent :

---

rouge », ses préférences personnelles l'amènent à choisir l'une des deux options avec une probabilité relative $p = \frac{probabilite\ de\ choisir\ la\ voiture}{probabilite\ de\ choisir\ le\ bus\ rouge}$ donnée. Si on introduit une option supplémentaire, par exemple « bus publics – ligne bleue », la condition de NDANP est remplie si la probabilité relative de choisir l'option voiture ou l'option bus rouge reste inchangée. Si au contraire l'individu n'est pas indifférent à la nouvelle option plus récemment introduite, et modifie sa probabilité de choisir la voiture ou le bus rouge, alors la probabilité relative initiale p change. Dans ce cas, la condition de NDANP n'est pas remplie. Différentes méthodes de vérification de la condition de NDANP existent, cependant leur mise en œuvre est techniquement exigeante, et les logiciels statistiques ne les proposent pas forcément. Lorsque la condition de NDANP n'est pas remplie, la validité des paramètres de la régression logit multinomiale devient douteuse. Pour obtenir des paramètres plus fiables, d'autres modèles de régression logistique multinomiale doivent alors être utilisés, par exemple le modèle probit multinomial ou le modèle logit mixte, qui ne nécessitent pas que la condition de NDANP soit remplie (mais que tous les logiciels ne proposent pas).

(a) le genre ;

(b) l'âge (20-39 ans ; 40-59 ans) ;

(c) la position au regard du marché du travail (emploi ; chômage/inactivité), et le niveau de titre visé (Niveau II – Licence / Cadre ; Niveau I – Master / Ingénieur).

Trois canaux de validation sont accessibles à cette population : Université ; Université de Technologie ; et Conservatoire national des arts et métiers (CNAM).

Tableau 21.3.
Caractéristiques de candidats à la VAE et choix du canal de validation

Codage :

Genre (G)	1 : Femme 2 : Homme
Âge (A)	1 : 20-39 ans 2 : 40-59 ans
Position sur le marché du travail (P)	1 : Emploi 2 : Chômage ou Inactivité
Niveau de titre visé (N)	1 : Niveau II, Licence / Cadre 2 : Niveau I, Master / Ingénieur

Canaux de validation (C)
1 : Université
2 : Université de Technologie
3 : Conservatoire national des arts et métiers (CNAM)

ID.	Genre	Age	(In)activité	Niveau	Canal
1	2	2	2	1	1
2	2	2	1	2	3
3	2	2	1	2	3
4	2	1	2	2	2
5	1	1	2	2	2
6	1	1	2	1	1
7	2	1	2	2	1
8	1	1	2	2	2
9	1	1	1	2	2
10	2	1	2	2	2
11	1	2	2	2	2
12	1	1	2	2	2
13	2	1	1	1	1
14	1	1	2	2	2
15	1	2	2	1	2
16	1	1	2	1	1
17	1	2	1	2	3
18	2	1	2	2	2
19	2	2	1	2	3
20	1	2	2	2	2
21	1	1	2	2	3
22	1	2	2	2	3

ID.	Genre	Age	(In)activité	Niveau	Canal
23	2	1	2	1	1
24	1	1	2	2	2
25	2	2	1	2	3
26	2	2	2	2	3
27	1	1	2	1	1
28	2	2	1	2	3
29	1	2	1	1	3
30	2	2	2	1	1
31	2	1	1	2	3
32	2	1	2	1	1
33	2	1	2	2	2
34	1	1	2	2	2
35	1	1	2	1	1
36	1	1	2	2	2
37	2	2	1	2	3
38	1	1	2	1	1
39	1	1	2	1	1
40	1	2	1	1	3
41	2	1	2	2	1
42	1	2	1	2	3
43	1	2	1	2	3
44	1	2	1	2	3
45	1	1	1	1	1
46	1	1	2	1	1

1. Vérification des conditions générales de validité

Aucune variable indépendante n'est quantitative ou qualitative ordinale, donc la question de la présence de valeurs aberrantes, la question de la multicolinéarité et celle de la linéarité ne se posent pas. Le modèle comporte quatre variables indépendantes qualitatives, donc, avec 46 individus sans observation manquante, l'échantillon est de taille suffisante.

On vérifie la condition relative aux effectifs des cases des tableaux de contingence observés et théoriques pour chaque paire de variables indépendantes qualitatives :

Tableau de contingence observé Genre / Age

	1	2	Total
1	17	10	27
2	10	9	19
Total	27	19	46

Tableau de contingence théorique Genre / Age

	1	2	Total
1	15,84	11,15	27
2	11,15	7,84	19
Total	27	19	46

Tableau de contingence observé Genre / (In)activité

	1	2	Total
1	8	19	27
2	8	11	19
Total	16	30	46

Tableau de contingence théorique Genre / (In)activité

	1	2	Total
1	9,39	17,60	27
2	6,60	12,39	19
Total	16	30	46

Tableau de contingence observé Genre / Niveau

	1	2	Total
1	11	16	27
2	5	14	19
Total	16	30	46

Tableau de contingence théorique Genre / Niveau

	1	2	Total
1	9,39	17,60	27
2	6,60	12,39	19
Total	16	30	46

Tableau de contingence observé Age / (In)activité

	1	2	Total
1	4	23	27
2	12	7	19
Total	16	30	46

Tableau de contingence théorique Age / (In)activité

	1	2	Total
1	9,39	17,60	27
2	6,60	12,39	19
Total	16	30	46

Tableau de contingence observé Age / Niveau

	1	2	Total
1	11	16	27
2	5	14	19
Total	16	30	46

Tableau de contingence théorique Age / Niveau

	1	2	Total
1	9,39	17,60	27
2	6,60	12,39	19
Total	16	30	46

Tableau de contingence observé (In)activité / Niveau

	1	2	Total
1	4	12	16
2	12	18	30
Total	16	30	46

Tableau de contingence théorique (In)activité / Niveau

	1	2	Total
1	5,56	10,43	16
2	10,43	19,56	30
Total	16	30	46

Aucune des cases des tableaux de contingence observés n'a un effectif nul. Aucune des cases des tableaux de contingence théoriques n'a un effectif inférieur à 5. Les conditions générales de validité sont donc remplies[199].

2. Mise en œuvre, résultats et interprétation

La régression multinomiale s'effectue et s'interprète toujours par rapport à l'une des modalités de la variable dépendante, qui sert de référence. On commence donc par choisir la modalité de la variable dépendante qui servira de référence, par exemple ici la modalité 1

[199] La condition de non-dépendance aux alternatives non-pertinentes, vérifiée ici au moyen du *suest-based Hausman test* du logiciel STATA, est également remplie.

(Université). Le tableau des paramètres de régression se présente comme suit :

Paramètres de la régression par rapport à la modalité 1 de la variable dépendante

Modalité	Source	Valeur	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds ratio
2	Cte	-6,77	0,07	0,78	-55,16	41,60	
	G-1	0,00					
	G-2	-14,44	0,13	0,71	-91,56	62,67	0
	A-1	0,00					
	A-2	16,86	0,17	0,67	-61,26	94,99	21116103
	P-1	0,00					
	P-2	-3,88	0,03	0,85	-46,32	38,55	0.021
	N-1	0,00					
	N-2	25,69	0,26	0,60	-72,85	124,23	1,44E+11
3	Cte	-7,94	0,10	0,74	-56,44	40,55	
	G-1	0,00					
	G-2	-12,98	0,10	0,74	-90,11	64,13	0
	A-1	0,00					
	A-2	20,27	0,25	0,61	-57,90	98,44	636503330
	P-1	0,00					
	P-2	-7,56	0,12	0,72	-49,98	34,86	0.001
	N-1	0,00					
	N-2	26,95	0,28	0,59	-71,58	125,50	5,106E+11

Le premier groupe de lignes (« Modalité 2 ») fournit les résultats des estimations pour la deuxième modalité de la variable dépendante (Université de Technologie) *par rapport* à la première modalité (Université). La significativité des paramètres est fournie dans la colonne « Pr > Khi² ». On constate qu'aucun des coefficients n'est significatif au seuil usuel de 5%. On en conclut qu'aucune des variables indépendantes n'affecte significativement la probabilité de choisir l'Université de Technologie plutôt que l'Université. En d'autres termes, les variables indépendantes présentes dans le modèle n'entrainent pas de préférence pour l'un ou l'autre des deux canaux que sont l'Université et l'Université de Technologie.

Le deuxième groupe de lignes (« Modalité 3 ») fournit les résultats des estimations pour la troisième modalité de la variable dépendante (CNAM) par rapport à la première modalité (Université). Là non plus, aucun des coefficients n'est significatif au seuil de 5%. Donc aucune des variables indépendantes n'affecte significativement la probabilité de choisir le CNAM plutôt que l'Université.

Reste à savoir si les variables indépendantes affectent le choix entre Université de Technologie et CNAM. Il faut donc relancer l'estimation, mais en adoptant cette fois-ci la modalité 2 (Université de Technologie)

comme modalité de référence. Le tableau des paramètres de régression indique à présent :

Paramètres de la régression par rapport à la modalité 2 de la variable dépendante

Modalité	Source	Valeur	Khi² de Wald	Pr > Khi²	Borne inf. (95%)	Borne sup. (95%)	Odds ratio
1	Cte	6,77	0,07	0,78	-41,6	55,16	
	G-1	0,00					
	G-2	14,44	0,13	0,71	-62,6	91,56	1883533,6
	A-1	0,00					
	A-2	-16,8	0,17	0,67	-94,9	61,26	0,00
	P-1	0,00					
	P-2	3,88	0,03	0,85	-38,5	46,32	48,78
	N-1	0,00					
	N-2	-25,6	0,26	0,60	-124,2	72,8	0,00
3	Cte	-1,16	0,25	0,61	-5,68	3,35	
	G-1	0,00					
	G-2	1,46	0,87	0,35	-1,60	4,52	4,31
	A-1	0,00					
	A-2	3,40	4,79	0,02	0,35	6,45	30,14
	P-1	0,00					
	P-2	-3,67	5,98	0,01	-6,61	-0,73	0,025
	N-1	0,00					
	N-2	1,26	0,36	0,54	-2,82	5,35	3,54

Le point d'intérêt ici (puisque la comparaison entre modalités 1 et 2 a déjà été effectuée à l'étape

précédente) est le deuxième groupe de lignes, qui compare la modalité 3 (CNAM) à la modalité de référence 2 (Université de Technologie). On constate que deux paramètres sont significatifs.

C'est le cas tout d'abord du coefficient de la modalité 2 (40-59 ans) de la variable Age : coefficient = 3,40 ; p=0,02. On note que le coefficient est positif. On peut donc conclure que les 40-59 ans sont significativement plus enclins (ont plus tendance) que les 20-39 ans (catégorie de référence de la variable Age) à préférer le CNAM à l'Université de Technologie. L'odds ratio permet de quantifier cette préférence : les chances relatives de préférer le CNAM à l'Université de Technologie sont 30,14 fois plus élevées chez les 40-59 ans que chez les 20-39 ans.

C'est le cas ensuite du coefficient de la modalité 2 (Chômage/Inactivité) de la variable *Position sur le marché du travail* : coefficient = -3,67 ; p=0,01. On note que le coefficient est négatif. On en conclut que les répondants qui sont en emploi ont significativement plus tendance (sont plus enclins) que ceux qui sont inactifs ou au chômage à préférer le CNAM à l'Université de Technologie. L'odds ratio indique que les chances relatives de préférer le CNAM à l'Université de Technologie sont de 97,5% plus faibles chez les répondants inactifs ou au chômage que chez les répondants en emploi.

Donc, en résumé, les variables indépendantes du modèle influencent significativement le choix du canal de validation : les répondants de plus de 40 ans et les répondants en emploi ont significativement plus tendance que les autres à préférer le CNAM à l'Université de Technologie.

21.4. RÉGRESSION LOGISTIQUE ORDINALE

La régression logistique ordinale s'applique au cas dans lequel la variable réponse est qualitative ordinale polytomique, tandis que la ou les variables indépendantes sont qualitatives et/ou quantitatives.

Les conditions de validité sont les conditions générales auxquelles s'ajoute la condition spécifique de *proportionnalité des cotes*[200]. La condition de proportionnalité des cotes signifie que la corrélation entre variable(s) indépendante(s) et variable dépendante est la même quelle que soit la modalité de variable dépendante que l'on considère[201]. Différents

[200] Ou « condition de régression parallèle », ou encore « hypothèse de lignes parallèles ».

[201] En effet, dans la procédure de régression logistique ordinale, les modalités de la variable dépendante sont, par définition, ordonnées (modalité 1 < modalité 2 < modalité 3 <…), et la

tests de proportionnalité des cotes existent[202], mais tous les logiciels ne les proposent pas. Lorsque la condition n'est pas remplie, la fiabilité des paramètres de régression est douteuse. Une alternative consiste alors à établir plusieurs versions binaires de la variable

---

procédure vise à déterminer les chances relatives d'une modalité *et de toutes celles qui la précèdent* par rapport aux modalités qui lui sont supérieures. La procédure est, en ce sens, dite « cumulative ». Par exemple si la variable dépendante a quatre modalités, la régression ordinale vise à déterminer les chances relatives de la modalité 1 par rapport à l'ensemble des modalités 2, 3 et 4 ; les chances relatives l'ensemble des modalités 1 et 2 par rapport à l'ensemble des modalités 3 et 4 ; et les chances relatives de l'ensemble des modalités 1, 2 et 3 par rapport à la modalité 4. Les chances relatives (*odds*) s'écrivent :

Pour la modalité 1 par rapport aux modalités 2, 3 et 4 :	$\frac{p_1}{p_2 + p_3 + p_4}$
Pour l'ensemble des modalités 1 et 2 par rapport à l'ensemble des modalités 3 et 4 :	$\frac{p_1 + p_2}{p_3 + p_4}$
Pour l'ensemble des modalités 1, 2 et 3 par rapport à la modalité 4 :	$\frac{p_1 + p_2 + p_3}{p_4}$

où p_m est la probabilité de la modalité *m*. La condition de proportionnalité des chances relatives signifie que l'effet des variables indépendantes est identique quel que soit l'ensemble de modalités dont on examine les probabilités relatives.

[202] Basés notamment sur le rapport de vraisemblance, le Khi-carré de Wald ou le score. L'hypothèse nulle des tests est que les cotes sont proportionnelles. L'hypothèse nulle est rejetée si la p-value est inférieure au seuil de significativité.

dépendante[203] et à effectuer une régression logistique dichotomique pour chaque version. On peut aussi décider de recourir plutôt (si le logiciel utilisé s'y prête) à des procédures de régression logistique ordinale n'exigeant pas la proportionnalité des cotes (« modèles à cotes non proportionnelles » ou « modèles à proportionnalité partielle »).

L'interprétation des résultats s'effectue à partir du tableau des paramètres de régression.

EXEMPLE 21.4

On dispose de données sur l'évaluation d'un enseignement par des étudiants (Tableau 21.4). L'information disponible couvre le genre et le style d'apprentissage de l'étudiant (préférence pour le cours magistral ou au contraire pour une pédagogie active)

[203] Par exemple, si la variable dépendante comportait initialement quatre modalités, trois versions binaires sont possibles :

	Nouvelle modalité 1	Nouvelle modalité 2
Version binaire 1	Ancienne modalité 1	Anciennes modalités 2, 3, 4
Version binaire 2	Anciennes modalités 1 et 2	Anciennes modalités 3 et 4
Version binaire 3	Anciennes modalités 1, 2 et 3	Ancienne modalité 4

ainsi que les notes octroyées (de D minimum à A maximum). On cherche à savoir dans quelle mesure le genre et/ou le style d'apprentissage des étudiants déterminent les notes qu'ils décernent.

Tableau 21.4.
Genre et style d'apprentissage des étudiants et évaluation de l'enseignement

Codage
Genre – Femme : 1 ; Homme : 2
Style d'apprentissage – Cours magistral : 1 ; Pédagogie active : 2
Notes – D : 1 ; C : 2 ; B : 3 ; A : 4.

Identifiant	Genre	Style	Notes
1	2	1	2
2	1	2	1
3	1	2	1
4	1	2	3
5	2	1	2
6	1	2	3
7	1	1	1
8	2	1	2
9	2	1	2
10	2	1	2
11	1	2	3
12	1	2	3
13	1	2	3
14	2	1	4

Identifiant	Genre	Style	Notes
15	1	2	4
16	1	2	3
17	2	1	4
18	1	2	1
19	1	2	3
20	1	2	1
21	2	2	1
22	2	2	1
23	1	1	4
24	2	2	1
25	2	2	1
26	1	2	3
27	1	1	4
28	2	2	1
29	2	1	2
30	1	1	4
31	1	2	3
32	1	1	2
33	1	1	4
34	2	1	2
35	1	1	4
36	1	1	4
37	2	1	2
38	1	1	2
39	2	2	3
40	1	1	4

1. Vérification des conditions générales de validité

Aucune variable indépendante n'est quantitative ou qualitative ordinale, donc la question de la présence de valeurs aberrantes, la question de la multicolinéarité et celle de la linéarité ne se posent pas. Le modèle comporte deux variables indépendantes qualitatives, donc, avec 40 individus sans observation manquante, l'échantillon est de taille suffisante.

On vérifie la condition relative aux effectifs des cases des tableaux de contingence observé et théorique :

Tableau de contingence observé Genre / Style

	1	2	Total
1	10	14	24
2	10	6	16
Total	20	20	40

Tableau de contingence théorique Genre / Style

	1	2	Total
1	12	12	24
2	8	8	16
Total	20	20	40

Aucune des cases du tableau de contingence observé n'a un effectif nul. Aucune des cases du tableau de

contingence théorique n'a un effectif inférieur à 5. Les conditions générales de validité sont donc remplies[204].

2. Résultats et interprétation

Le tableau des paramètres de régression se présente comme suit :

---

[204] Cependant, la condition de proportionnalité des cotes, vérifiée via le test de la commande *omodel* du logiciel STATA, n'est pas remplie (p-value = 0,0000) :

omodel logit note genre style
Ordered logit estimates
Number of obs = 40
LR chi2(2) = 16.17
Prob > chi2 = 0.0003
Log likelihood = -47.364559
Pseudo R2 = 0.1458

note	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
genre	-2.455237	.7656388	-3.21	0.001	-3.955862	-.9546126
style	-2.137105	.7086526	-3.02	0.003	-3.526038	-.7481713
_cut1	-8.115739	2.030708	(Ancillary parameters)			
_cut2	-6.732294	1.927591				
_cut3	-5.290037	1.822735				

Approximate likelihood-ratio test of proportionality of odds across response categories:
chi2(4) = 36.58
Prob > chi2 = 0.0000

Source	Valeur	ES*	Khi² de Wald	Pr > Khi²	Wald Borne inf. (95%)	Wald Borne sup. (95%)
Constante1	-3,52	0,81	18,75	<0,0001	-5,11	-1,92
Constante2	-2,14	0,70	9,32	0,002	-3,51	-0,76
Constante3	-0,69	0,64	1,18	0,277	-1,95	0,56
Genre-1	0,00	0,00				
Genre-2	2,45	0,76	10,28	0,001	0,95	3,95
Style-1	0,00	0,00				
Style-2	2,13	0,70	9,09	0,003	0,74	3,52

* ES : Erreur standard

On voit que les paramètres des deux variables indépendantes sont significatifs.

De façon générale, le paramètre d'une variable indépendante représente l'effet qu'exerce une variation de cette variable indépendante sur le logarithme de la cote des modalités de la variable dépendante. Les modalités de la variable dépendante étant 1, 2, 3, etc., leurs cotes s'écrivent :

- $\frac{p_1}{p_2+p_3+\cdots}$ pour la cote de la modalité 1 par rapport aux modalités 2, 3, etc. (où p_m est la probabilité de la modalité m) ;

- $\frac{p_1+p_2}{p_3+\cdots}$ pour la cote des modalités 1 et 2 par rapport aux modalités 3, etc. ;
- et ainsi de suite.

Un paramètre positif signifie donc qu'une augmentation d'une unité de la variable indépendante (si la variable indépendante est quantitative) ou le passage de la catégorie de référence à une autre catégorie de la variable indépendante (si la variable indépendante est qualitative) entraine une augmentation de la probabilité des modalités placées au numérateur de la cote. Ainsi, dans l'exemple présenté ici, les hommes (Genre-2) ont, par rapport aux femmes (Genre-1, catégorie de référence), une plus forte probabilité d'attribuer une note D (modalité 1) plutôt qu'une note C, B ou A (modalités 2, 3 et 4).

La condition de proportionnalité des cotes n'est pas remplie dans cet exemple[205]. Si elle l'avait été, cela aurait signifié que l'interprétation est la même pour toutes les cotes : par rapport aux femmes, les hommes ont aussi une plus forte probabilité d'attribuer une note D ou C plutôt qu'une note B ou A ; et ils ont aussi une

[205] Donc, pour poursuivre l'analyse, il conviendrait d'adopter une autre approche, par exemple établir plusieurs versions binaires de la variable dépendante (comme expliqué plus haut en note 203) puis effectuer une régression logistique dichotomique pour chacune de ces versions.

plus forte probabilité d'attribuer une note D, C ou B plutôt qu'une note A.

Il importe tout particulièrement de s'assurer de la façon dont le logiciel utilisé définit les cotes des modalités lors de la régression ordinale. Il y a en effet deux méthodes possibles :

Cotes auxquelles renvoient les paramètres de régression Exemple avec quatre modalités de variable dépendante (p_m est la probabilité de la modalité *m*)	
Méthode 1	Méthode 2
$\frac{p_1}{p_2 + p_3 + p_4}$	$\frac{p_4}{p_3 + p_2 + p_1}$
$\frac{p_1 + p_2}{p_3 + p_4}$	$\frac{p_4 + p_3}{p_2 + p_1}$
$\frac{p_1 + p_2 + p_3}{p_4}$	$\frac{p_4 + p_3 + p_2}{p_1}$

Dans XLSTAT (utilisé pour le présent exemple), c'est la première méthode qui est appliquée. Un coefficient positif pour une variable indépendante signifie donc qu'une hausse dans la variable indépendante entraine une hausse de la probabilité des modalités *inférieures*

et une baisse de la probabilité des modalités supérieures de la variable dépendante. Un coefficient négatif signifierait au contraire une hausse de la probabilité des modalités supérieures par rapport aux modalités inférieures.

D'autres logiciels, au contraire utilisent la deuxième méthode, de sorte qu'un coefficient positif signifie une hausse de la probabilité des modalités supérieures ; et un coefficient négatif, une hausse de la probabilité des modalités inférieures.

Enfin, pour quantifier l'effet d'une variation dans la variable indépendante, on utilise le rapport des cotes (*odds ratio*) en calculant (lorsqu'il n'est pas déjà fourni par le logiciel) l'exponentielle du coefficient de la variable. Dans l'exemple, le paramètre de la modalité Homme est 2,455. On peut calculer que l'exponentielle de 2,455 est 11,64. Donc les probabilités relatives (cotes) d'attribuer une note inférieure sont chez les hommes 11,64 fois ce qu'elles sont chez les femmes (autrement dit les hommes ont 1064% plus tendance[206] que les femmes à attribuer une note inférieure). De même, les étudiants ayant une préférence pour la pédagogie active ont 747,39% plus tendance que les autres à attribuer une note inférieure.

[206] C'est-à-dire 1064% plus de « probabilités relatives », « chances relatives », « risques relatifs ».

Chapitre 22. Modèles d'équations structurelles : mesurer l'effet de construits abstraits

Il arrive parfois que le facteur dont on cherche à évaluer l'incidence soit de nature abstraite. Par exemple, si l'on s'intéresse aux déterminants la réussite académique des élèves, le nombre d'heures d'enseignement ou le niveau de diplôme des enseignants sont de possibles facteurs concrets, tandis que l'origine sociale des élèves ou le climat scolaire sont des facteurs abstraits. Bien évidemment, on peut essayer de représenter un facteur de nature abstraite par des *indicateurs* concrets. Par exemple, on peut représenter l'origine sociale de l'élève par la catégorie socioprofessionnelle à laquelle appartient le chef de famille, ou encore le climat scolaire par le sentiment de sécurité et de soutien ressenti (ou non) par les élèves. Mais ces indicateurs concrets ne sont au mieux que des manifestations du phénomène, pas le phénomène lui-même, et ce que l'on mesure finalement, c'est l'effet des indicateurs, pas au sens strict l'effet du phénomène. C'est précisément l'objectif des modèles d'équations structurelles (MES) que de chercher à mesurer l'effet exercé par ou sur le phénomène lui-même. Dans la terminologie des MES, le phénomène sous-jacent non-directement observable est dénommé *variable latente*, par opposition à ses indicateurs observables dénommés *variables manifestes*.

La problématique de la modélisation par équations structurelles rappelle quelque peu celle de l'analyse factorielle exploratoire[207], du fait de l'intérêt, dans les deux méthodes, pour l'analyse de facteurs sous-jacents. La différence entre ces deux méthodes est cependant que l'analyse factorielle exploratoire ne postulait pas les relations entre les variables et les facteurs sous-jacents. Dans l'analyse factorielle, on cherchait à déceler l'éventuelle présence de facteurs sous-jacents. Dans la modélisation par équations structurelles, les possibles facteurs sous-jacents sont préalablement identifiés et postulés, et l'analyse vise à confirmer le modèle théorique postulé. La modélisation par équations structurelles est, en ce sens, parfois appelée « analyse factorielle confirmatoire » (*confirmatory factor analysis – CFA*).

22.1. DÉMARCHE GÉNÉRALE D'UTILISATION DES MODÈLES D'ÉQUATIONS STRUCTURELLES

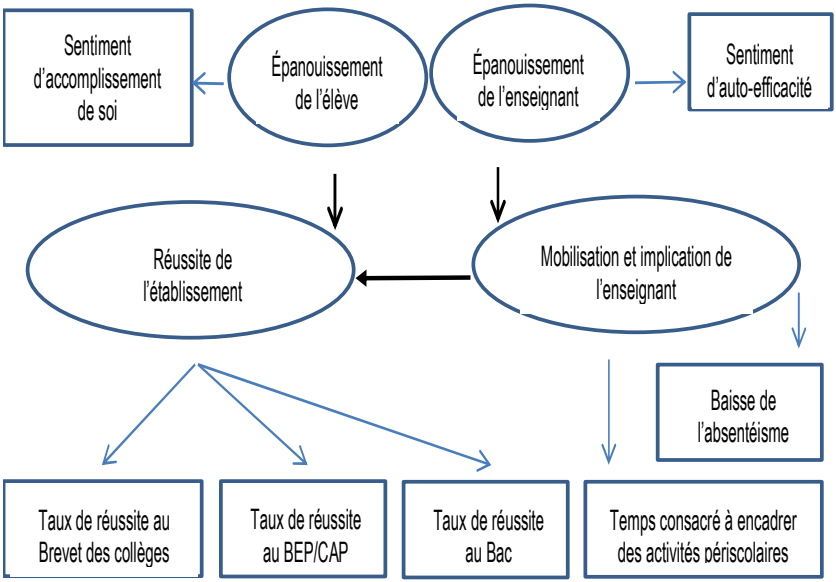
L'objectif de l'utilisation[208] des modèles d'équations structurelles est d'identifier l'effet exercé par ou sur des variables latentes. Une variable latente peut être *exogène* (c'est-à-dire indépendante) ou *endogène*

[207] Voir chapitre 14.

[208] Pour une présentation des fondements et méthodes des modèles d'équations structurelles, voir par exemple Jakobowicz (2007, 2013).

(c'est-à-dire dépendante). Les MES permettent ainsi d'estimer l'effet d'une variable latente exogène sur une variable latente endogène (par exemple l'effet de la qualité de l'enseignement sur l'image d'un établissement). Cela étant, les MES peuvent être utilisés aussi pour analyser l'effet de variables observées sur des variables latentes dont elles ne sont pas les variables manifestes, ou inversement l'effet de variables latentes sur des variables observées autres que leurs propres indicateurs.

La procédure d'utilisation des MES s'organise en trois principales étapes. La première consiste à concevoir et traduire en termes graphiques le modèle théorique dans le cadre duquel s'inscrivent les effets à mesurer. Dans le modèle, chaque variable latente (inobservable par définition) est représentée avec sa ou ses variables manifestes, c'est-à-dire avec l'indicateur ou les indicateurs qui en sont une traduction partielle quantifiée. Imaginons par exemple qu'on cherche à identifier l'effet de l'épanouissement des enseignants sur la réussite de l'établissement dans le cadre d'un modèle dont la représentation graphique s'établit comme suit :



Les variables latentes (en ovales) sont accompagnées de leurs variables manifestes (en carrés). Un MES peut comprendre plusieurs variables latentes indépendantes et plusieurs variables latentes dépendantes. Les flèches en gras indiquent le sens des effets. Les effets peuvent être directs et/ou indirects. Il n'est pas nécessaire qu'une variable latente soit reliée à plus d'une autre variable latente. Une variable latente peut avoir plusieurs variables manifestes. Toutes les variables manifestes doivent être quantitatives ou qualitatives ordinales. La cohérence interne des variables manifestes d'une variable latente, vérifiable au moyen

de l'alpha de Cronbach[209], doit être assurée. On appelle « modèle structurel » l'ensemble des relations entre variables latentes, et « modèle de mesure » l'ensemble comprenant les relations entre variables latentes et variables manifestes ainsi que les relations entre variables manifestes.

La deuxième étape consiste à saisir les données numériques nécessaires à l'estimation. Par hypothèse, il n'existe de données que pour les variables manifestes. La procédure de saisie dépend évidemment du logiciel utilisé. Dans ce chapitre, c'est le logiciel SPSS AMOS d'IBM qui est utilisé[210].

La troisième étape consiste à lancer l'estimation, puis à interpréter les résultats. Les résultats de l'estimation

---

[209] Voir section 2.1 (paragraphe 2.1.1.2.3).

[210] Téléchargeable sur le site d'IBM : https://www.ibm.com/fr-fr/marketplace/structural-equation-modeling-sem. Une version d'évaluation peut être testée pendant quinze jours. En cas d'acquisition, l'accès au logiciel n'est cependant pas garanti au-delà d'un an. Le guide de l'utilisateur (voir tout particulièrement la première partie puis le chapitre 5 de la deuxième partie) est accessible à :

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/25.0/en/amos/Manuals/IBM_SPSS_Amos_User_Guide.pdf

L'importation de données stockées en fichier Excel 1997-2003 (.xls) est possible. L'importation s'effectue via la commande *File / Data Files*.

permettent d'évaluer (a) la qualité de l'ajustement du modèle ; et (b) la significativité statistique :
- des liens entre les variables latentes et leurs variables manifestes ; et
- des effets des variables latentes.

La qualité de l'ajustement s'interprète à partir d'indices de qualité d'ajustement. Dans AMOS, plusieurs indices de qualité d'ajustement sont proposés. D'usage commode est la probabilité associée au Khi-carré compte-tenu du nombre de degrés de libertés lors du test du rapport de vraisemblance (*Likelihood Ratio test*). Si le modèle est bien ajusté, la probabilité est d'au moins 5%. Si elle est inférieure à 5%, la qualité de l'ajustement est mauvaise, et le modèle doit être rejeté.

La significativité des coefficients de régression s'apprécie de la même façon que dans une régression linéaire. Une p-value inférieure à 5% indique une relation significative.

Les MES sont gourmands en données : on estime généralement à 200 minimum la taille d'échantillon requise pour obtenir des estimations stables.

Deux importantes conditions de validité des MES sont l'absence de valeurs aberrantes multivariées[211] et la multinormalité des données[212] :

- AMOS fournit, avec les résultats d'estimation, la distance de Mahalanobis calculée pour chaque observation (*Mahalanobis d-squared*). La liste de ces distances est classée par ordre décroissant. On considère comme valeurs aberrantes les observations pour lesquelles p1<0,001. Les valeurs aberrantes doivent être retirées de l'échantillon ;
- AMOS fournit également des indicateurs de normalité et de multinormalité. La multinormalité se vérifie ici au moyen d'un « ratio critique » (*critical ratio* ou *c.r.*) associé au coefficient d'aplatissement de Mardia (*multivariate kurtosis*). La condition de multinormalité peut être considérée comme remplie tant que le ratio critique n'est pas supérieur à 1,96. Il y a non-multinormalité au-dessus de 1,96. Lorsque la condition de multinormalité n'est pas remplie, une alternative consiste à effectuer les estimations au moyen d'une procédure ADF (*asymptotically distribution-free*)[213] , non sensible à la non-normalité, en lieu et place de l'estimation par maximum de vraisemblance (*maximum likelihood*).

---

[211] Voir section 15.3.1.

[212] Voir section 15.3.1.

[213] Option disponible dans AMOS via la commande *View / Analysis properties / Estimation*.

22.2. EXEMPLE

On cherche à vérifier une théorie selon laquelle l'exposition à la diversité culturelle à l'école (DC) favorise le développement de la personnalité de l'élève (DP). On dispose de données sur un échantillon de 45 élèves (Tableau 22.1), indiquant pour chaque élève le nombre de nationalités représentées dans l'établissement fréquenté (*nbnat*), le nombre de langues (régionales ou étrangères) enseignées dans l'établissement (*nblang*), et les scores obtenus aux items Sociabilité (*sociab*) et Confiance en soi (*conf*) de questionnaires de personnalité.

Tableau 22.1.
Développement de la personnalité d'élèves en fonction de la diversité culturelle dans l'établissement fréquenté

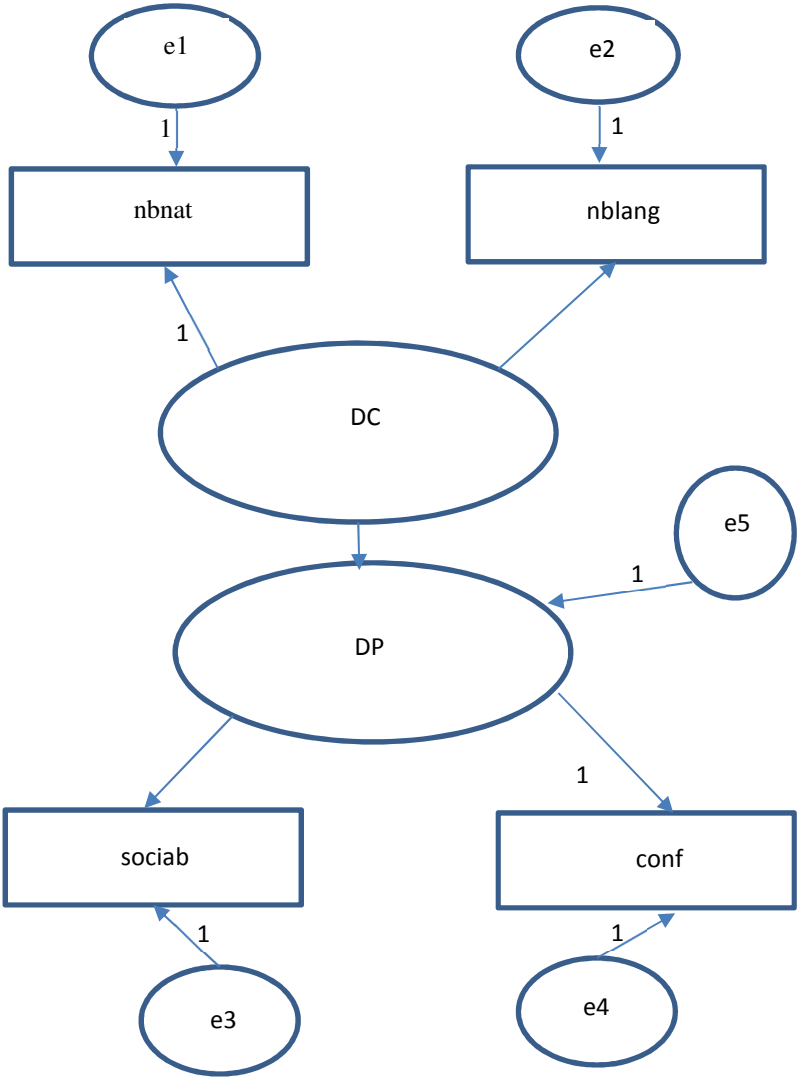
Identifiant de l'élève	nbnat	nblang	sociab	conf
1	4	8	67	50
2	3	8	79	40
3	8	8	25	26
4	2	10	42	54
5	7	6	44	21
6	20	10	80	60
7	3	7	19	17
8	3	10	18	38
9	4	7	41	33
10	1	6	61	56

Identifiant de l'élève	nbnat	nblang	sociab	conf
11	7	6	50	18
12	16	9	42	41
13	5	10	41	20
14	3	1	80	17
15	6	10	19	46
16	8	3	80	54
17	18	9	63	54
18	3	8	53	4
19	7	6	66	17
20	10	7	56	57
21	16	8	61	49
22	19	10	74	57
23	15	5	24	2
24	19	9	72	55
25	7	3	37	20
26	19	9	68	55
27	12	2	75	57
28	14	7	46	53
29	4	3	73	19
30	20	10	79	60
31	11	6	17	12
32	12	1	17	33
33	7	8	46	46
34	3	2	74	6
35	8	6	21	49
36	9	3	59	12

Identifiant de l'élève	nbnat	nblang	sociab	conf
37	19	9	64	55
38	10	5	50	28
39	19	9	73	55
40	11	3	22	29
41	15	9	39	46
42	1	10	73	23
43	16	8	63	50
44	19	10	78	59
45	15	7	44	26

Représentation graphique du modèle

On commence par établir le diagramme du modèle :



Les variables non directement observables sont représentées par des ovales ou des cercles, les variables

observées par des rectangles. Diversité culturelle et développement de la personnalité sont les variables latentes, chacune reliée à ses variables manifestes.

Chaque variable dépendante, qu'elle soit latente ou manifeste, est associée à un terme d'erreur (e1 à e5). Le terme d'erreur représente tout à la fois les éventuelles erreurs de mesure et les autres déterminants possibles de la variable dépendante non présents dans le modèle. Par exemple, le nombre de langues enseignées dans un établissement résulte sans doute aussi d'autres facteurs que la diversité culturelle de l'établissement et n'est sans doute pas un parfait reflet de cette dernière.

En outre, pour les besoins de l'estimation, à chaque variable inobservée doit être associée une « *contrainte d'identification* », c'est-à-dire un paramètre permettant d'étalonner les calculs. La contrainte d'identification apparait sous forme d'un chiffre figurant sur une des flèches partant de chaque variable inobservée. La valeur de la contrainte est fixée arbitrairement puisque son rôle est d'étalonner des mesures dont, par définition, on ne connait pas l'unité de référence (variables latentes et erreurs sont inobservées). En pratique, la valeur des contraintes est habituellement fixée à 1.

Vérification des conditions de validité

La cohérence interne entre variables manifestes[214], tout d'abord, n'est pas établie :

	Diversité culturelle	Développement personnel
Alpha de Cronbach	0,46	0,52

On peut donc s'attendre à des liens non significatifs entre chaque variable latente et ses indicatrices.

Les distances de Mahalanobis (D), ensuite, s'établissent comme suit :

Identifiant	D	p1	p2
32	10,44	0,03	0,79
23	8,97	0,06	0,78
42	8,33	0,08	0,71
27	8,23	0,08	0,52
14	7,78	0,10	0,47
10	7,16	0,13	0,52
34	7,12	0,13	0,37
16	6,68	0,15	0,39
4	6,65	0,16	0,26

[214] Vérification non effectuée par le logiciel Amos.

Identifiant	D	p1	p2
18	6,53	0,16	0,19
8	6,28	0,18	0,17
15	5,67	0,23	0,30
35	5,29	0,26	0,38
40	5,13	0,27	0,34
31	4,90	0,30	0,35
2	4,49	0,34	0,49
13	4,46	0,35	0,39
29	4,26	0,37	0,40
6	4,05	0,40	0,43
30	3,95	0,41	0,39
36	3,76	0,44	0,41
7	3,71	0,45	0,34
44	3,51	0,48	0,38
1	3,42	0,49	0,34
22	3,18	0,53	0,42
39	2,67	0,61	0,75
19	2,66	0,62	0,65
24	2,61	0,63	0,59
25	2,57	0,63	0,50
26	2,40	0,66	0,55
37	2,28	0,69	0,55
3	2,25	0,69	0,45
45	2,19	0,70	0,39
20	1,99	0,74	0,47
17	1,87	0,76	0,47

Identifiant	D	p1	p2
12	1,84	0,77	0,37
41	1,70	0,79	0,38
28	1,59	0,81	0,37
9	1,36	0,85	0,48
33	1,30	0,86	0,39
11	1,29	0,86	0,24
43	1,03	0,91	0,37
21	0,97	0,91	0,25
5	0,86	0,93	0,17
38	0,65	0,96	0,14

On voit qu'aucune observation n'a une distance dont la probabilité p1 est inférieure à 0,001. Il n'y a donc pas de valeur aberrante multivariée dans l'échantillon.

Les statistiques de normalité, ensuite, s'établissent comme suit :

Variable	min	max	skew	c.r.	kurtosis	c.r.
conf	2	60	-0.36	-1.00	-1.28	-1.75
sociab	17	80	-0.36	-0.99	-1.14	-1.56
nblang	1	10	-0.70	-1.92	-0.64	-0.88
nbnat	1	20	0.20	0.55	-1.34	-1.84
Multivariate					-0.83	-0.40

La valeur absolue du ratio critique du coefficient de Mardia (le *multivariate kurtosis*) s'établit à 0,40. Donc il n'y a pas de non-normalité significative.

Qualité de l'ajustement

Les résultats du test du rapport de vraisemblance (*Likelihood Ratio test*) indiquent :

Chi-square = .841
Degrees of freedom = 1
Probability level = .359

On voit donc que la probabilité associée au Khi-carré compte-tenu du nombre de degrés de liberté est supérieure à 5%. Par conséquent, le modèle est acceptable.

Significativité des relations entre variables

Le tableau des paramètres de régression indique :

Regression Weights: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	p-value	Label
DP	<---	DC	4.592	2.054	2.236	.025	par_2
nbnat	<---	DC	1.000				
nblang	<---	DC	.413	.153	2.696	.007	par_1
sociab	<---	DP	.346	.273	1.266	.206	par_3
conf	<---	DP	1.000				

On voit que le coefficient de régression 4,592 associé à la variable *Diversité culturelle* est significatif au seuil de 5% : la p-value est égale à 0,025. Ce résultat confirme l'hypothèse théorique qui sous-tend le modèle étudié : la diversité culturelle exerce un effet positif significatif sur le développement de la personnalité de l'élève[215].

On voit aussi que la relation entre la variable latente *Diversité culturelle* et sa variable manifeste *Nombre de langues* est significative. En revanche, la relation entre la variable latente *Développement de la personnalité* et sa variable manifeste *Sociabilité* ne l'est pas.

Les paramètres des variables *Nombre de nationalités* et *Confiance en soi* ne peuvent être calculés car affectés par les contraintes d'identification[216]. Pour les calculer,

[215] Rappelons si besoin est qu'il ne s'agit ici que de résultats fictifs basés sur des données fictives construites à des fins pédagogiques pour illustrer la mise en œuvre de la méthode des modèles d'équations structurelles.

[216] Des paramètres standardisés sont cependant fournis pour toutes les variables :

il faut déplacer les contraintes d'identification vers les variables *Nombre de langues* et *Sociabilité*. Ce sont alors désormais les paramètres de ces deux variables qui sont affectés. On obtient les p-values 0,007 et 0,206 pour *nbnat* et *conf* respectivement.

La non-significativité des paramètres des variables *Sociabilité* et *Confiance en soi* suggère que le modèle pourrait être amélioré en remplaçant ces variables par d'autres mieux liées au développement de la personnalité. Ce que laissait prévoir, du reste, la faiblesse des alphas de Cronbach.

Standardized Regression Weights: (Group number 1 - Default model)

			Estimate
DP	<---	DC	.821
nbnat	<---	DC	.563
nblang	<---	DC	.525
sociab	<---	DP	.323
conf	<---	DP	1.082

Conclusion générale

Cette *Introduction pratique* devrait aider le lecteur à se doter de solides compétences en matière d'analyse quantitative, et à identifier avec aisance à quelle méthode recourir, quand, comment, pourquoi, avec quelles attentes mais aussi quelles limites. Maîtriser ces compétences constitue un atout majeur pour le chercheur en éducation et formation, non seulement au moment de les mettre en œuvre concrètement lors d'investigations sur le terrain, mais aussi en amont, déjà au moment de concevoir une recherche. Car alors, dès la phase de conception, le chercheur dispose d'un champ de vision plus large et d'une plus grande liberté de choix quant aux approches possibles, qualitatives, quantitatives, ou mixtes, pour tendre au mieux vers les qualités d'envergure et de portée, de pertinence et de rigueur, de finesse et de solidité, qui distinguent une bonne recherche.

Le champ couvert dans cet ouvrage n'est évidemment pas exhaustif. Les méthodes efficaces et d'usage plus aisé pour des utilisateurs non familiers ont été privilégiées, au détriment d'autres, parfois plus traditionnelles et populaires, mais éventuellement aussi plus délicates du point de vue de la procédure de mise en œuvre et/ou de l'interprétation des résultats. D'autres méthodes encore, bien que d'intérêt pour les analystes et chercheurs en éducation et formation, sont

absentes également. C'est le cas par exemple de la modélisation multi-niveaux, de la régression linéaire multivariée, ou encore de l'analyse de covariance multivariée. La raison ici est l'absence de fonctions prenant en charge une part suffisante des procédures nécessaires dans la plupart des logiciels statistiques accessibles. Sur ce plan, l'évolution des logiciels à l'avenir pourrait donc conduire à étendre davantage le champ couvert lors de futures éditions de cet ouvrage. Mais d'ici-là, le bagage acquis ici devrait déjà permettre au lecteur d'aller plus loin de façon autonome, pour explorer de nouvelles approches.

C'est toute l'ambition de cet ouvrage.

Index

T

U

V

Y

Références

Laurence Bardin (2013), *L'analyse de contenu*, Paris : Presses Universitaires de France (Collection Quadrige Manuels, 2ème édition).

Justin Bem, Martin Mba, Ludovic Subran (2008), Calcul de précision et plan de sondage : application aux enquêtes camerounaises auprès des ménages (ECAM 2 et ECAM 3), *STATECO*, No 102, 31-43 : https://www.insee.fr/fr/statistiques/fichier/2120926/stec102d.pdf.

Philippe Cibois (1993), Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence, *Bulletin de méthodologie sociologique*, 40, 43-63.

Anna B. Costello, Jason W. Osborne (2005), Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis, *Practical Assessment, Research & Evaluation*, 10(7), 4: http://pareonline.net/pdf/v10n7.pdf.

F. Daumas (1982), Méthodes de normalisation des données, *Revue de Statistique Appliquée*, 30(4), 23-38 : http://www.numdam.org/article/RSA_1982__30_4_23_0.pdf.

M. A. Fligner, T. J. Killeen, (1976), Distribution-free two-sample tests for scale, *Journal of the American Statistical Association*, 71(353), 210-213.

Debbie L. Hahs-Vaughn (2016), *Applied Multivariate Statistical Concepts*, Routledge.

Alexandra Marie Henchy (2013), Review and Evaluation of Reliability Generalization Research, *Theses and Dissertations – Educational, School, and Counseling Psychology*, 5, 11-15: https://uknowledge.uky.edu/edp_etds/5

Glenn D. Israel (2015), *Sampling the Evidence of Extension Program Impact*, PEOD5, The Agricultural Education and Communication Department, Institute of Food and Agricultural Sciences, University of Florida: http://edis.ifas.ufl.edu/pdffiles/PD/PD00500.pdf.

Glenn D. Israel (2012), *Determining Sample Size*, PEOD6, The Agricultural Education and Communication Department, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida https://www.psycholosphere.com/Determining%20sample%20size%20by%20Glen%20Israel.pdf.

Emmanuel Jakobowicz (2007), *Contribution aux modèles d'équations structurelles à variables latentes*, thèse : https://tel.archives-ouvertes.fr/file/index/docid/207990/filename/these_modeles_struct_EJakobowicz.pdf.

Emmanuel Jakobowicz (2013), *Les modèles d'équations structurelles à variables latentes*, cours de statistique multivariée approfondie au CNAM : http://cedric.cnam.fr/~saporta/STA201%20-%20Equations%20structurelles%20-%201%20-%20Jakobowicz.pdf.

V. Ménoni, N. Lucas, J. F. Leforestier, J. Dimet, F. Doz, G. Chatellier, *et al*. (2010), The Readability of Information and Consent Forms in Clinical Research in France, PLoS ONE 5(5): e10576: https://doi.org/10.1371/journal.pone.0010576.

V. Planchon (2005), Traitement des valeurs aberrantes : concepts actuels et tendances générales, *Biotechnol. Agron. Soc. Environ.*, 9 (1), 19-34 : http://www.pressesagro.be/base/text/v9n1/19.pdf.

Noëlle Sorin (1996), De la lisibilité linguistique à une lisibilité sémiotique, *Revue québécoise de linguistique*, 25(1), 61-98.

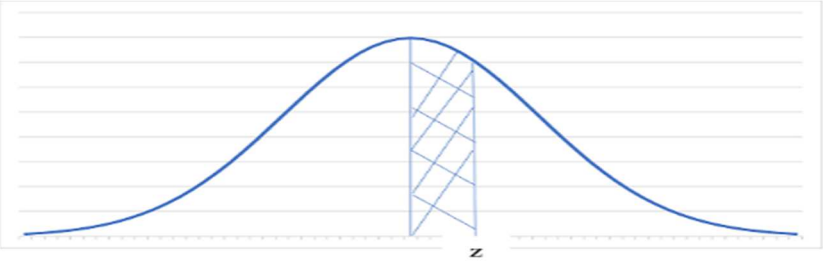
B. G. Tabachnick, L. S. Fidell (2007), *Using Multivariate Statistics* (5th Ed.), Boston: Pearson.

Yamane, Taro (1967), *Statistics: An Introductory Analysis*, 2nd Ed., New York: Harper and Row.

Annexes

ANNEXE 1. TABLE DE $Z_{\alpha/2}$

La table de la loi normale centrée réduite (table de Z) indique pour chaque coefficient de confiance (en abscisse de la courbe), le degré de confiance α correspondant (mesuré par la surface sous la courbe). La version de la table de Z présentée ici permet de lire directement le coefficient de confiance (z) associe à $\frac{\alpha}{2}$ (la surface hachurée située sous la partie droite de la courbe) :



Z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,0199	0,0239	0,0279	0,0319	0,0359
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,1	0,0398	0,0438	0,0478	0,0517	0,0557
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,1	0,0596	0,0636	0,0675	0,0714	0,0753

	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,2	0,0793	0,0832	0,0871	0,0910	0,0948
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,2	0,0987	0,1026	0,1064	0,1103	0,1141
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,3	0,1179	0,1217	0,1255	0,1293	0,1331
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,3	0,1368	0,1406	0,1443	0,1480	0,1517
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,4	0,1554	0,1591	0,1628	0,1664	0,1700
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,4	0,1736	0,1772	0,1808	0,1844	0,1879
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,5	0,1915	0,1950	0,1985	0,2019	0,2054
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,5	0,2088	0,2123	0,2157	0,2190	0,2224
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,6	0,2257	0,2291	0,2324	0,2357	0,2389
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,6	0,2422	0,2454	0,2486	0,2517	0,2549
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,7	0,2580	0,2611	0,2642	0,2673	0,2704
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,7	0,2734	0,2764	0,2794	0,2823	0,2852

	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,8	0,2881	0,2910	0,2939	0,2967	0,2995
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,8	0,3023	0,3051	0,3078	0,3106	0,3133
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
0,9	0,3159	0,3186	0,3212	0,3238	0,3264
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
0,9	0,3289	0,3315	0,3340	0,3365	0,3389
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,0	0,3413	0,3438	0,3461	0,3485	0,3508
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,0	0,3531	0,3554	0,3577	0,3599	0,3621
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,1	0,3643	0,3665	0,3686	0,3708	0,3729
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,1	0,3749	0,3770	0,3790	0,3810	0,3830
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,2	0,3849	0,3869	0,3888	0,3907	0,3925
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,2	0,3944	0,3962	0,3980	0,3997	0,4015
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,3	0,4032	0,4049	0,4066	0,4082	0,4099

	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,3	0,4115	0,4131	0,4147	0,4162	0,4177
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,4	0,4192	0,4207	0,4222	0,4236	0,4251
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,4	0,4265	0,4279	0,4292	0,4306	0,4319
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,5	0,4332	0,4345	0,4357	0,4370	0,4382
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,5	0,4394	0,4406	0,4418	0,4429	0,4441
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,6	0,4452	0,4463	0,4474	0,4484	0,4495
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,6	0,4505	0,4515	0,4525	0,4535	0,4545
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,7	0,4554	0,4564	0,4573	0,4582	0,4591
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,7	0,4599	0,4608	0,4616	0,4625	0,4633
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,8	0,4641	0,4649	0,4656	0,4664	0,4671
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,8	0,4678	0,4686	0,4693	0,4699	0,4706

	0,00	**0,01**	**0,02**	**0,03**	**0,04**
1,9	0,4713	0,4719	0,4726	0,4732	0,4738
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
1,9	0,4744	0,4750	0,4756	0,4761	0,4767
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,0	0,4772	0,4778	0,4783	0,4788	0,4793
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,0	0,4798	0,4803	0,4808	0,4812	0,4817
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,1	0,4821	0,4826	0,4830	0,4834	0,4838
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,1	0,4842	0,4846	0,4850	0,4854	0,4857
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,2	0,4861	0,4864	0,4868	0,4871	0,4875
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,2	0,4878	0,4881	0,4884	0,4887	0,4890
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,3	0,4893	0,4896	0,4898	0,4901	0,4904
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,3	0,4906	0,4909	0,4911	0,4913	0,4916
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,4	0,4918	0,4920	0,4922	0,4925	0,4927
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,4	0,4929	0,4931	0,4932	0,4934	0,4936

	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,5	0,4938	0,4940	0,4941	0,4943	0,4945
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,5	0,4946	0,4948	0,4949	0,4951	0,4952
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,6	0,4953	0,4955	0,4956	0,4957	0,4959
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,6	0,4960	0,4961	0,4962	0,4963	0,4964
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,7	0,4965	0,4966	0,4967	0,4968	0,4969
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,7	0,4970	0,4971	0,4972	0,4973	0,4974
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,8	0,4974	0,4975	0,4976	0,4977	0,4977
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,8	0,4978	0,4979	0,4979	0,4980	0,4981
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
2,9	0,4981	0,4982	0,4982	0,4983	0,4984
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
2,9	0,4984	0,4985	0,4985	0,4986	0,4986
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,0	0,4987	0,4987	0,4987	0,4988	0,4988

	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,0	0,4989	0,4989	0,4989	0,4990	0,4990
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,1	0,4990	0,4991	0,4991	0,4991	0,4992
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,1	0,4992	0,4992	0,4992	0,4993	0,4993
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,2	0,4993	0,4993	0,4994	0,4994	0,4994
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,2	0,4994	0,4994	0,4995	0,4995	0,4995
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,3	0,4995	0,4995	0,4995	0,4996	0,4996
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,3	0,4996	0,4996	0,4996	0,4996	0,4997
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,4	0,4997	0,4997	0,4997	0,4997	0,4997
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,4	0,4997	0,4997	0,4997	0,4997	0,4998
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,5	0,4998	0,4998	0,4998	0,4998	0,4998
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,5	0,4998	0,4998	0,4998	0,4998	0,4998

	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,6	0,4998	0,4998	0,4999	0,4999	0,4999
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,6	0,4999	0,4999	0,4999	0,4999	0,4999
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,7	0,4999	0,4999	0,4999	0,4999	0,4999
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,7	0,4999	0,4999	0,4999	0,4999	0,4999
	0,00	**0,01**	**0,02**	**0,03**	**0,04**
3,8	0,4999	0,4999	0,4999	0,4999	0,4999
	0,05	**0,06**	**0,07**	**0,08**	**0,09**
3,8	0,4999	0,4999	0,4999	0,4999	0,4999

ANNEXE 2. TABLE DE FISHER-SNEDECOR AU SEUIL DE SIGNIFICATIVITÉ DE 5%

DDL1 : degrés de libertés du numérateur
DDL2 : degrés de liberté du dénominateur

Lecture : au seuil de significativité de 5%, si DDL1 =3 et DDL2 = 5, l'hypothèse nulle peut être rejetée dès lors que la valeur calculée de F est supérieure à 5,41.

	DDL1				
DDL2	1	2	3	4	5
1	161,45	199,50	215,71	224,58	230,16
2	18,51	19,00	19,16	19,25	19,30
3	10,13	9,55	9,28	9,12	9,01
4	7,71	6,94	6,59	6,39	6,26
5	6,61	5,79	5,41	5,19	5,05
6	5,99	5,14	4,76	4,53	4,39
7	5,59	4,74	4,35	4,12	3,97
8	5,32	4,46	4,07	3,84	3,69
9	5,12	4,26	3,86	3,63	3,48
10	4,96	4,10	3,71	3,48	3,33
11	4,84	3,98	3,59	3,36	3,20
12	4,75	3,89	3,49	3,26	3,11
13	4,67	3,81	3,41	3,18	3,03
14	4,60	3,74	3,34	3,11	2,96
15	4,54	3,68	3,29	3,06	2,90
16	4,49	3,63	3,24	3,01	2,85
17	4,45	3,59	3,20	2,96	2,81
18	4,41	3,55	3,16	2,93	2,77
19	4,38	3,52	3,13	2,90	2,74
20	4,35	3,49	3,10	2,87	2,71
21	4,32	3,47	3,07	2,84	2,68
22	4,30	3,44	3,05	2,82	2,66
23	4,28	3,42	3,03	2,80	2,64
24	4,26	3,40	3,01	2,78	2,62
25	4,24	3,39	2,99	2,76	2,60

DDL2	DDL1 6	7	8	9	10
1	233,99	236,77	238,88	240,54	241,88
2	19,33	19,35	19,37	19,38	19,40
3	8,94	8,89	8,85	8,81	8,79
4	6,16	6,09	6,04	6,00	5,96
5	4,95	4,88	4,82	4,77	4,74
6	4,28	4,21	4,15	4,10	4,06
7	3,87	3,79	3,73	3,68	3,64
8	3,58	3,50	3,44	3,39	3,35
9	3,37	3,29	3,23	3,18	3,14
10	3,22	3,14	3,07	3,02	2,98
11	3,09	3,01	2,95	2,90	2,85
12	3,00	2,91	2,85	2,80	2,75
13	2,92	2,83	2,77	2,71	2,67
14	2,85	2,76	2,70	2,65	2,60
15	2,79	2,71	2,64	2,59	2,54
16	2,74	2,66	2,59	2,54	2,49
17	2,70	2,61	2,55	2,49	2,45
18	2,66	2,58	2,51	2,46	2,41
19	2,63	2,54	2,48	2,42	2,38
20	2,60	2,51	2,45	2,39	2,35
21	2,57	2,49	2,42	2,37	2,32
22	2,55	2,46	2,40	2,34	2,30
23	2,53	2,44	2,37	2,32	2,27
24	2,51	2,42	2,36	2,30	2,25
25	2,49	2,40	2,34	2,28	2,24

	DDL1				
	11	12	13	14	15
DDL2					
1	242,98	243,90	244,69	245,36	245,95
2	19,40	19,41	19,42	19,42	19,43
3	8,76	8,74	8,73	8,71	8,70
4	5,94	5,91	5,89	5,87	5,86
5	4,70	4,68	4,66	4,64	4,62
6	4,03	4,00	3,98	3,96	3,94
7	3,60	3,57	3,55	3,53	3,51
8	3,31	3,28	3,26	3,24	3,22
9	3,10	3,07	3,05	3,03	3,01
10	2,94	2,91	2,89	2,86	2,85
11	2,82	2,79	2,76	2,74	2,72
12	2,72	2,69	2,66	2,64	2,62
13	2,63	2,60	2,58	2,55	2,53
14	2,57	2,53	2,51	2,48	2,46
15	2,51	2,48	2,45	2,42	2,40
16	2,46	2,42	2,40	2,37	2,35
17	2,41	2,38	2,35	2,33	2,31
18	2,37	2,34	2,31	2,29	2,27
19	2,34	2,31	2,28	2,26	2,23
20	2,31	2,28	2,25	2,22	2,20
21	2,28	2,25	2,22	2,20	2,18
22	2,26	2,23	2,20	2,17	2,15
23	2,24	2,20	2,18	2,15	2,13
24	2,22	2,18	2,15	2,13	2,11
25	2,20	2,16	2,14	2,11	2,09

	DDL1				
	16	17	18	19	20
DDL2					
1	246,47	246,92	247,32	247,69	248,02
2	19,43	19,44	19,44	19,44	19,45
3	8,69	8,68	8,67	8,67	8,66
4	5,84	5,83	5,82	5,81	5,80
5	4,60	4,59	4,58	4,57	4,56
6	3,92	3,91	3,90	3,88	3,87
7	3,49	3,48	3,47	3,46	3,44
8	3,20	3,19	3,17	3,16	3,15
9	2,99	2,97	2,96	2,95	2,94
10	2,83	2,81	2,80	2,79	2,77
11	2,70	2,69	2,67	2,66	2,65
12	2,60	2,58	2,57	2,56	2,54
13	2,51	2,50	2,48	2,47	2,46
14	2,44	2,43	2,41	2,40	2,39
15	2,38	2,37	2,35	2,34	2,33
16	2,33	2,32	2,30	2,29	2,28
17	2,29	2,27	2,26	2,24	2,23
18	2,25	2,23	2,22	2,20	2,19
19	2,21	2,20	2,18	2,17	2,16
20	2,18	2,17	2,15	2,14	2,12
21	2,16	2,14	2,12	2,11	2,10
22	2,13	2,11	2,10	2,08	2,07
23	2,11	2,09	2,08	2,06	2,05
24	2,09	2,07	2,05	2,04	2,03
25	2,07	2,05	2,04	2,02	2,01

	DDL1				
	22	24	26	28	30
DDL2					
1	248,58	249,05	249,45	249,80	250,10
2	19,45	19,45	19,46	19,46	19,46
3	8,65	8,64	8,63	8,62	8,62
4	5,79	5,77	5,76	5,75	5,75
5	4,54	4,53	4,52	4,50	4,50
6	3,86	3,84	3,83	3,82	3,81
7	3,43	3,41	3,40	3,39	3,38
8	3,13	3,12	3,10	3,09	3,08
9	2,92	2,90	2,89	2,87	2,86
10	2,75	2,74	2,72	2,71	2,70
11	2,63	2,61	2,59	2,58	2,57
12	2,52	2,51	2,49	2,48	2,47
13	2,44	2,42	2,41	2,39	2,38
14	2,37	2,35	2,33	2,32	2,31
15	2,31	2,29	2,27	2,26	2,25
16	2,25	2,24	2,22	2,21	2,19
17	2,21	2,19	2,17	2,16	2,15
18	2,17	2,15	2,13	2,12	2,11
19	2,13	2,11	2,10	2,08	2,07
20	2,10	2,08	2,07	2,05	2,04
21	2,07	2,05	2,04	2,02	2,01
22	2,05	2,03	2,01	2,00	1,98
23	2,02	2,01	1,99	1,97	1,96
24	2,00	1,98	1,97	1,95	1,94
25	1,98	1,96	1,95	1,93	1,92

DDL2	DDL1 35	40	45	50	60
1	250,69	251,14	251,49	251,77	252,20
2	19,47	19,47	19,47	19,48	19,48
3	8,60	8,59	8,59	8,58	8,57
4	5,73	5,72	5,71	5,70	5,69
5	4,48	4,46	4,45	4,44	4,43
6	3,79	3,77	3,76	3,75	3,74
7	3,36	3,34	3,33	3,32	3,30
8	3,06	3,04	3,03	3,02	3,01
9	2,84	2,83	2,81	2,80	2,79
10	2,68	2,66	2,65	2,64	2,62
11	2,55	2,53	2,52	2,51	2,49
12	2,44	2,43	2,41	2,40	2,38
13	2,36	2,34	2,33	2,31	2,30
14	2,28	2,27	2,25	2,24	2,22
15	2,22	2,20	2,19	2,18	2,16
16	2,17	2,15	2,14	2,12	2,11
17	2,12	2,10	2,09	2,08	2,06
18	2,08	2,06	2,05	2,04	2,02
19	2,05	2,03	2,01	2,00	1,98
20	2,01	1,99	1,98	1,97	1,95
21	1,98	1,96	1,95	1,94	1,92
22	1,96	1,94	1,92	1,91	1,89
23	1,93	1,91	1,90	1,88	1,86
24	1,91	1,89	1,88	1,86	1,84
25	1,89	1,87	1,86	1,84	1,82

	DDL1				
	80	100	200	500	1000
DDL2					
1	252,72	253,04	253,68	254,06	254,19
2	19,48	19,49	19,49	19,49	19,49
3	8,56	8,55	8,54	8,53	8,53
4	5,67	5,66	5,65	5,64	5,63
5	4,41	4,41	4,39	4,37	4,37
6	3,72	3,71	3,69	3,68	3,67
7	3,29	3,27	3,25	3,24	3,23
8	2,99	2,97	2,95	2,94	2,93
9	2,77	2,76	2,73	2,72	2,71
10	2,60	2,59	2,56	2,55	2,54
11	2,47	2,46	2,43	2,42	2,41
12	2,36	2,35	2,32	2,31	2,30
13	2,27	2,26	2,23	2,22	2,21
14	2,20	2,19	2,16	2,14	2,14
15	2,14	2,12	2,10	2,08	2,07
16	2,08	2,07	2,04	2,02	2,02
17	2,03	2,02	1,99	1,97	1,97
18	1,99	1,98	1,95	1,93	1,92
19	1,96	1,94	1,91	1,89	1,88
20	1,92	1,91	1,88	1,86	1,85
21	1,89	1,88	1,84	1,83	1,82
22	1,86	1,85	1,82	1,80	1,79
23	1,84	1,82	1,79	1,77	1,76
24	1,82	1,80	1,77	1,75	1,74
25	1,80	1,78	1,75	1,73	1,72

	DDL1				
	1	2	3	4	5
DDL2					
26	4,23	3,37	2,98	2,74	2,59
27	4,21	3,35	2,96	2,73	2,57
28	4,20	3,34	2,95	2,71	2,56
29	4,18	3,33	2,93	2,70	2,55
30	4,17	3,32	2,92	2,69	2,53
32	4,15	3,29	2,90	2,67	2,51
34	4,13	3,28	2,88	2,65	2,49
36	4,11	3,26	2,87	2,63	2,48
38	4,10	3,24	2,85	2,62	2,46
40	4,08	3,23	2,84	2,61	2,45
42	4,07	3,22	2,83	2,59	2,44
44	4,06	3,21	2,82	2,58	2,43
46	4,05	3,20	2,81	2,57	2,42
48	4,04	3,19	2,80	2,57	2,41
50	4,03	3,18	2,79	2,56	2,40
55	4,02	3,16	2,77	2,54	2,38
60	4,00	3,15	2,76	2,53	2,37
65	3,99	3,14	2,75	2,51	2,36
70	3,98	3,13	2,74	2,50	2,35
75	3,97	3,12	2,73	2,49	2,34
80	3,96	3,11	2,72	2,49	2,33
85	3,95	3,10	2,71	2,48	2,32
90	3,95	3,10	2,71	2,47	2,32
95	3,94	3,09	2,70	2,47	2,31
100	3,94	3,09	2,70	2,46	2,31

	DDL1				
	6	7	8	9	10
DDL2					
26	2,47	2,39	2,32	2,27	2,22
27	2,46	2,37	2,31	2,25	2,20
28	2,45	2,36	2,29	2,24	2,19
29	2,43	2,35	2,28	2,22	2,18
30	2,42	2,33	2,27	2,21	2,16
32	2,40	2,31	2,24	2,19	2,14
34	2,38	2,29	2,23	2,17	2,12
36	2,36	2,28	2,21	2,15	2,11
38	2,35	2,26	2,19	2,14	2,09
40	2,34	2,25	2,18	2,12	2,08
42	2,32	2,24	2,17	2,11	2,06
44	2,31	2,23	2,16	2,10	2,05
46	2,30	2,22	2,15	2,09	2,04
48	2,29	2,21	2,14	2,08	2,03
50	2,29	2,20	2,13	2,07	2,03
55	2,27	2,18	2,11	2,06	2,01
60	2,25	2,17	2,10	2,04	1,99
65	2,24	2,15	2,08	2,03	1,98
70	2,23	2,14	2,07	2,02	1,97
75	2,22	2,13	2,06	2,01	1,96
80	2,21	2,13	2,06	2,00	1,95
85	2,21	2,12	2,05	1,99	1,94
90	2,20	2,11	2,04	1,99	1,94
95	2,20	2,11	2,04	1,98	1,93
100	2,19	2,10	2,03	1,97	1,93

	DDL1				
	11	12	13	14	15
DDL2					
26	2,18	2,15	2,12	2,09	2,07
27	2,17	2,13	2,10	2,08	2,06
28	2,15	2,12	2,09	2,06	2,04
29	2,14	2,10	2,08	2,05	2,03
30	2,13	2,09	2,06	2,04	2,01
32	2,10	2,07	2,04	2,01	1,99
34	2,08	2,05	2,02	1,99	1,97
36	2,07	2,03	2,00	1,98	1,95
38	2,05	2,02	1,99	1,96	1,94
40	2,04	2,00	1,97	1,95	1,92
42	2,03	1,99	1,96	1,94	1,91
44	2,01	1,98	1,95	1,92	1,90
46	2,00	1,97	1,94	1,91	1,89
48	1,99	1,96	1,93	1,90	1,88
50	1,99	1,95	1,92	1,89	1,87
55	1,97	1,93	1,90	1,88	1,85
60	1,95	1,92	1,89	1,86	1,84
65	1,94	1,90	1,87	1,85	1,82
70	1,93	1,89	1,86	1,84	1,81
75	1,92	1,88	1,85	1,83	1,80
80	1,91	1,88	1,84	1,82	1,79
85	1,90	1,87	1,84	1,81	1,79
90	1,90	1,86	1,83	1,80	1,78
95	1,89	1,86	1,82	1,80	1,77
100	1,89	1,85	1,82	1,79	1,77

	DDL1				
	16	17	18	19	20
DDL2					
26	2,05	2,03	2,02	2,00	1,99
27	2,04	2,02	2,00	1,99	1,97
28	2,02	2,00	1,99	1,97	1,96
29	2,01	1,99	1,97	1,96	1,94
30	1,99	1,98	1,96	1,95	1,93
32	1,97	1,95	1,94	1,92	1,91
34	1,95	1,93	1,92	1,90	1,89
36	1,93	1,92	1,90	1,88	1,87
38	1,92	1,90	1,88	1,87	1,85
40	1,90	1,89	1,87	1,85	1,84
42	1,89	1,87	1,86	1,84	1,83
44	1,88	1,86	1,84	1,83	1,81
46	1,87	1,85	1,83	1,82	1,80
48	1,86	1,84	1,82	1,81	1,79
50	1,85	1,83	1,81	1,80	1,78
55	1,83	1,81	1,79	1,78	1,76
60	1,82	1,80	1,78	1,76	1,75
65	1,80	1,78	1,76	1,75	1,73
70	1,79	1,77	1,75	1,74	1,72
75	1,78	1,76	1,74	1,73	1,71
80	1,77	1,75	1,73	1,72	1,70
85	1,76	1,74	1,73	1,71	1,70
90	1,76	1,74	1,72	1,70	1,69
95	1,75	1,73	1,71	1,70	1,68
100	1,75	1,73	1,71	1,69	1,68

	DDL1				
	22	24	26	28	30
DDL2					
26	1,97	1,95	1,93	1,91	1,90
27	1,95	1,93	1,91	1,90	1,88
28	1,93	1,91	1,90	1,88	1,87
29	1,92	1,90	1,88	1,87	1,85
30	1,91	1,89	1,87	1,85	1,84
32	1,88	1,86	1,85	1,83	1,82
34	1,86	1,84	1,82	1,81	1,80
36	1,85	1,82	1,81	1,79	1,78
38	1,83	1,81	1,79	1,77	1,76
40	1,81	1,79	1,77	1,76	1,74
42	1,80	1,78	1,76	1,75	1,73
44	1,79	1,77	1,75	1,73	1,72
46	1,78	1,76	1,74	1,72	1,71
48	1,77	1,75	1,73	1,71	1,70
50	1,76	1,74	1,72	1,70	1,69
55	1,74	1,72	1,70	1,68	1,67
60	1,72	1,70	1,68	1,66	1,65
65	1,71	1,69	1,67	1,65	1,63
70	1,70	1,67	1,65	1,64	1,62
75	1,69	1,66	1,64	1,63	1,61
80	1,68	1,65	1,63	1,62	1,60
85	1,67	1,65	1,63	1,61	1,59
90	1,66	1,64	1,62	1,60	1,59
95	1,66	1,63	1,61	1,59	1,58
100	1,65	1,63	1,61	1,59	1,57

	DDL1				
	35	40	45	50	60
DDL2					
26	1,87	1,85	1,84	1,82	1,80
27	1,86	1,84	1,82	1,81	1,79
28	1,84	1,82	1,80	1,79	1,77
29	1,83	1,81	1,79	1,77	1,75
30	1,81	1,79	1,77	1,76	1,74
32	1,79	1,77	1,75	1,74	1,71
34	1,77	1,75	1,73	1,71	1,69
36	1,75	1,73	1,71	1,69	1,67
38	1,73	1,71	1,69	1,68	1,65
40	1,72	1,69	1,67	1,66	1,64
42	1,70	1,68	1,66	1,65	1,62
44	1,69	1,67	1,65	1,63	1,61
46	1,68	1,65	1,64	1,62	1,60
48	1,67	1,64	1,62	1,61	1,59
50	1,66	1,63	1,61	1,60	1,58
55	1,64	1,61	1,59	1,58	1,55
60	1,62	1,59	1,57	1,56	1,53
65	1,60	1,58	1,56	1,54	1,52
70	1,59	1,57	1,55	1,53	1,50
75	1,58	1,55	1,53	1,52	1,49
80	1,57	1,54	1,52	1,51	1,48
85	1,56	1,54	1,52	1,50	1,47
90	1,55	1,53	1,51	1,49	1,46
95	1,55	1,52	1,50	1,48	1,46
100	1,54	1,52	1,49	1,48	1,45

	DDL1				
	80	100	200	500	1000
DDL2					
26	1,78	1,76	1,73	1,71	1,70
27	1,76	1,74	1,71	1,69	1,68
28	1,74	1,73	1,69	1,67	1,66
29	1,73	1,71	1,67	1,65	1,65
30	1,71	1,70	1,66	1,64	1,63
32	1,69	1,67	1,63	1,61	1,60
34	1,66	1,65	1,61	1,59	1,58
36	1,64	1,62	1,59	1,56	1,56
38	1,62	1,61	1,57	1,54	1,54
40	1,61	1,59	1,55	1,53	1,52
42	1,59	1,57	1,53	1,51	1,50
44	1,58	1,56	1,52	1,49	1,49
46	1,57	1,55	1,51	1,48	1,47
48	1,56	1,54	1,49	1,47	1,46
50	1,54	1,52	1,48	1,46	1,45
55	1,52	1,50	1,46	1,43	1,42
60	1,50	1,48	1,44	1,41	1,40
65	1,49	1,46	1,42	1,39	1,38
70	1,47	1,45	1,40	1,37	1,36
75	1,46	1,44	1,39	1,36	1,35
80	1,45	1,43	1,38	1,35	1,34
85	1,44	1,42	1,37	1,34	1,32
90	1,43	1,41	1,36	1,33	1,31
95	1,42	1,40	1,35	1,32	1,30
100	1,41	1,39	1,34	1,31	1,30

	DDL1				
	1	2	3	4	5
DDL2					
125	3,92	3,07	2,68	2,44	2,29
150	3,90	3,06	2,66	2,43	2,27
200	3,89	3,04	2,65	2,42	2,26
300	3,87	3,03	2,63	2,40	2,24
500	3,86	3,01	2,62	2,39	2,23
1000	3,85	3,00	2,61	2,38	2,22
2000	3,85	3,00	2,61	2,38	2,22

	DDL1				
	6	7	8	9	10
DDL2					
125	2,17	2,08	2,01	1,96	1,91
150	2,16	2,07	2,00	1,94	1,89
200	2,14	2,06	1,98	1,93	1,88
300	2,13	2,04	1,97	1,91	1,86
500	2,12	2,03	1,96	1,90	1,85
1000	2,11	2,02	1,95	1,89	1,84
2000	2,10	2,01	1,94	1,88	1,84

	DDL1				
	11	12	13	14	15
DDL2					
125	1,87	1,83	1,80	1,77	1,75
150	1,85	1,82	1,79	1,76	1,73
200	1,84	1,80	1,77	1,74	1,72
300	1,82	1,78	1,75	1,72	1,70
500	1,81	1,77	1,74	1,71	1,69
1000	1,80	1,76	1,73	1,70	1,68
2000	1,79	1,76	1,73	1,70	1,67

	DDL1				
	16	17	18	19	20
DDL2					
125	1,73	1,71	1,69	1,67	1,66
150	1,71	1,69	1,67	1,66	1,64
200	1,69	1,67	1,66	1,64	1,62
300	1,68	1,66	1,64	1,62	1,61
500	1,66	1,64	1,62	1,61	1,59
1000	1,65	1,63	1,61	1,60	1,58
2000	1,65	1,63	1,61	1,59	1,58

	DDL1				
	22	24	26	28	30
DDL2					
125	1,63	1,60	1,58	1,57	1,55
150	1,61	1,59	1,57	1,55	1,54
200	1,60	1,57	1,55	1,53	1,52
300	1,58	1,55	1,53	1,51	1,50
500	1,56	1,54	1,52	1,50	1,48
1000	1,55	1,53	1,51	1,49	1,47
2000	1,55	1,52	1,50	1,48	1,46

	DDL1				
	35	40	45	50	60
DDL2					
125	1,52	1,49	1,47	1,45	1,42
150	1,50	1,48	1,45	1,44	1,41
200	1,48	1,46	1,43	1,41	1,39
300	1,46	1,43	1,41	1,39	1,36
500	1,45	1,42	1,40	1,38	1,35
1000	1,43	1,41	1,38	1,36	1,33
2000	1,43	1,40	1,38	1,36	1,32

	DDL1				
	80	100	200	500	1000
DDL2					
125	1,39	1,36	1,31	1,27	1,26
150	1,37	1,34	1,29	1,25	1,24
200	1,35	1,32	1,26	1,22	1,21
300	1,32	1,30	1,23	1,19	1,17
500	1,30	1,28	1,21	1,16	1,14
1000	1,29	1,26	1,19	1,13	1,11
2000	1,28	1,25	1,18	1,12	1,09

ANNEXE 3. TABLE DE STUDENT POUR TEST BILATÉRAL POUR 1 À 1000 DEGRÉS DE LIBERTÉ

1. Seuils de significativité de 1‰ à 5%

	0,05	**0,02**	**0,01**	**0,001**
DDL				
1	12,706	31,821	63,657	636,62
2	4,3027	6,9646	9,9248	31,599
3	3,1824	4,5407	5,8409	12,924
4	2,7764	3,7469	4,6041	8,6103
5	2,5706	3,3649	4,0321	6,8688
6	2,4469	3,1427	3,7074	5,9588
7	2,3646	2,998	3,4995	5,4079
8	2,306	2,8965	3,3554	5,0413
9	2,2622	2,8214	3,2498	4,7809
10	2,2281	2,7638	3,1693	4,5869
11	2,201	2,7181	3,1058	4,437
12	2,1788	2,681	3,0545	4,3178
13	2,1604	2,6503	3,0123	4,2208
14	2,1448	2,6245	2,9768	4,1405
15	2,1314	2,6025	2,9467	4,0728
16	2,1199	2,5835	2,9208	4,015
17	2,1098	2,5669	2,8982	3,9651

	0,05	**0,02**	**0,01**	**0,001**
DDL				
18	2,1009	2,5524	2,8784	3,9216
19	2,093	2,5395	2,8609	3,8834
20	2,086	2,528	2,8453	3,8495
21	2,0796	2,5176	2,8314	3,8193
22	2,0739	2,5083	2,8188	3,7921
23	2,0687	2,4999	2,8073	3,7676
24	2,0639	2,4922	2,7969	3,7454
25	2,0595	2,4851	2,7874	3,7251
26	2,0555	2,4786	2,7787	3,7066
27	2,0518	2,4727	2,7707	3,6896
28	2,0484	2,4671	2,7633	3,6739
29	2,0452	2,462	2,7564	3,6594
30	2,0423	2,4573	2,75	3,646
40	2,0211	2,4233	2,7045	3,551
60	2,0003	2,3901	2,6603	3,4602
80	1,9901	2,3739	2,6387	3,4163
120	1,9799	2,3578	2,6174	3,3735
1000	1,962	2,33	2,581	3,3

2. Seuils de significativité de 10% à 50%

	0,50	**0,30**	**0,20**	**0,10**
DDL				
1	1	1,9626	3,0777	6,3138
2	0,8165	1,3862	1,8856	2,92
3	0,7649	1,2498	1,6377	2,3534
4	0,7407	1,1896	1,5332	2,1318
5	0,7267	1,1558	1,4759	2,015
6	0,7176	1,1342	1,4398	1,9432
7	0,7111	1,1192	1,4149	1,8946
8	0,7064	1,1081	1,3968	1,8595
9	0,7027	1,0997	1,383	1,8331
10	0,6998	1,0931	1,3722	1,8125
11	0,6974	1,0877	1,3634	1,7959
12	0,6955	1,0832	1,3562	1,7823
13	0,6938	1,0795	1,3502	1,7709
14	0,6924	1,0763	1,345	1,7613
15	0,6912	1,0735	1,3406	1,7531
16	0,6901	1,0711	1,3368	1,7459
17	0,6892	1,069	1,3334	1,7396
18	0,6884	1,0672	1,3304	1,7341
19	0,6876	1,0655	1,3277	1,7291
20	0,687	1,064	1,3253	1,7247
21	0,6864	1,0627	1,3232	1,7207

	0,50	**0,30**	**0,20**	**0,10**
DDL				
22	0,6858	1,0614	1,3212	1,7171
23	0,6853	1,0603	1,3195	1,7139
24	0,6848	1,0593	1,3178	1,7109
25	0,6844	1,0584	1,3163	1,7081
26	0,684	1,0575	1,315	1,7056
27	0,6837	1,0567	1,3137	1,7033
28	0,6834	1,056	1,3125	1,7011
29	0,683	1,0553	1,3114	1,6991
30	0,6828	1,0547	1,3104	1,6973
40	0,6807	1,05	1,3031	1,6839
60	0,6786	1,0455	1,2958	1,6706
80	0,6776	1,0432	1,2922	1,6641
120	0,6765	1,0409	1,2886	1,6577
1000	0,675	1,037	1,282	1,646

ANNEXE 4. TABLE DU KHI-CARRÉ AUX SEUILS DE SIGNIFICATIVITÉ DE 10%, 5%, 1% ET 1‰, POUR 1 À 1000 DEGRÉS DE LIBERTÉ

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
1	2,706	3,841	6,635	10,828
2	4,605	5,991	9,21	13,816
3	6,251	7,815	11,345	16,266
4	7,779	9,488	13,277	18,467
5	9,236	11,07	15,086	20,515
6	10,645	12,592	16,812	22,458
7	12,017	14,067	18,475	24,322
8	13,362	15,507	20,09	26,124
9	14,684	16,919	21,666	27,877
10	15,987	18,307	23,209	29,588
11	17,275	19,675	24,725	31,264
12	18,549	21,026	26,217	32,909
13	19,812	22,362	27,688	34,528
14	21,064	23,685	29,141	36,123
15	22,307	24,996	30,578	37,697
16	23,542	26,296	32	39,252
17	24,769	27,587	33,409	40,79
18	25,989	28,869	34,805	42,312
19	27,204	30,144	36,191	43,82

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
20	28,412	31,41	37,566	45,315
21	29,615	32,671	38,932	46,797
22	30,813	33,924	40,289	48,268
23	32,007	35,172	41,638	49,728
24	33,196	36,415	42,98	51,179
25	34,382	37,652	44,314	52,62
26	35,563	38,885	45,642	54,052
27	36,741	40,113	46,963	55,476
28	37,916	41,337	48,278	56,892
29	39,087	42,557	49,588	58,301
30	40,256	43,773	50,892	59,703
31	41,422	44,985	52,191	61,098
32	42,585	46,194	53,486	62,487
33	43,745	47,4	54,776	63,87
34	44,903	48,602	56,061	65,247
35	46,059	49,802	57,342	66,619
36	47,212	50,998	58,619	67,985
37	48,363	52,192	59,893	69,346
38	49,513	53,384	61,162	70,703
39	50,66	54,572	62,428	72,055
40	51,805	55,758	63,691	73,402
41	52,949	56,942	64,95	74,745

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
42	54,09	58,124	66,206	76,084
43	55,23	59,304	67,459	77,419
44	56,369	60,481	68,71	78,75
45	57,505	61,656	69,957	80,077
46	58,641	62,83	71,201	81,4
47	59,774	64,001	72,443	82,72
48	60,907	65,171	73,683	84,037
49	62,038	66,339	74,919	85,351
50	63,167	67,505	76,154	86,661
51	64,295	68,669	77,386	87,968
52	65,422	69,832	78,616	89,272
53	66,548	70,993	79,843	90,573
54	67,673	72,153	81,069	91,872
55	68,796	73,311	82,292	93,168
56	69,919	74,468	83,513	94,461
57	71,04	75,624	84,733	95,751
58	72,16	76,778	85,95	97,039
59	73,279	77,931	87,166	98,324
60	74,397	79,082	88,379	99,607
61	75,514	80,232	89,591	100,888
62	76,63	81,381	90,802	102,166
63	77,745	82,529	92,01	103,442

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
64	78,86	83,675	93,217	104,716
65	79,973	84,821	94,422	105,988
66	81,085	85,965	95,626	107,258
67	82,197	87,108	96,828	108,526
68	83,308	88,25	98,028	109,791
69	84,418	89,391	99,228	111,055
70	85,527	90,531	100,425	112,317
71	86,635	91,67	101,621	113,577
72	87,743	92,808	102,816	114,835
73	88,85	93,945	104,01	116,092
74	89,956	95,081	105,202	117,346
75	91,061	96,217	106,393	118,599
76	92,166	97,351	107,583	119,85
77	93,27	98,484	108,771	121,1
78	94,374	99,617	109,958	122,348
79	95,476	100,749	111,144	123,594
80	96,578	101,879	112,329	124,839
81	97,68	103,01	113,512	126,083
82	98,78	104,139	114,695	127,324
83	99,88	105,267	115,876	128,565
84	100,98	106,395	117,057	129,804
85	102,079	107,522	118,236	131,041

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
86	103,177	108,648	119,414	132,277
87	104,275	109,773	120,591	133,512
88	105,372	110,898	121,767	134,745
89	106,469	112,022	122,942	135,978
90	107,565	113,145	124,116	137,208
91	108,661	114,268	125,289	138,438
92	109,756	115,39	126,462	139,666
93	110,85	116,511	127,633	140,893
94	111,944	117,632	128,803	142,119
95	113,038	118,752	129,973	143,344
96	114,131	119,871	131,141	144,567
97	115,223	120,99	132,309	145,789
98	116,315	122,108	133,476	147,01
99	117,407	123,225	134,642	148,23
100	118,498	124,342	135,807	149,449
101	119,589	125,458	136,971	150,667
102	120,679	126,574	138,134	151,884
103	121,769	127,689	139,297	153,099
104	122,858	128,804	140,459	154,314
105	123,947	129,918	141,62	155,528
106	125,035	131,031	142,78	156,74
107	126,123	132,144	143,94	157,952

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
108	127,211	133,257	145,099	159,162
109	128,298	134,369	146,257	160,372
110	129,385	135,48	147,414	161,581
111	130,472	136,591	148,571	162,788
112	131,558	137,701	149,727	163,995
113	132,643	138,811	150,882	165,201
114	133,729	139,921	152,037	166,406
115	134,813	141,03	153,191	167,61
116	135,898	142,138	154,344	168,813
117	136,982	143,246	155,496	170,016
118	138,066	144,354	156,648	171,217
119	139,149	145,461	157,8	172,418
120	140,233	146,567	158,95	173,617
121	141,315	147,674	160,1	174,816
122	142,398	148,779	161,25	176,014
123	143,48	149,885	162,398	177,212
124	144,562	150,989	163,546	178,408
125	145,643	152,094	164,694	179,604
126	146,724	153,198	165,841	180,799
127	147,805	154,302	166,987	181,993
128	148,885	155,405	168,133	183,186
129	149,965	156,508	169,278	184,379

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
130	151,045	157,61	170,423	185,571
131	152,125	158,712	171,567	186,762
132	153,204	159,814	172,711	187,953
133	154,283	160,915	173,854	189,142
134	155,361	162,016	174,996	190,331
135	156,44	163,116	176,138	191,52
136	157,518	164,216	177,28	192,707
137	158,595	165,316	178,421	193,894
138	159,673	166,415	179,561	195,08
139	160,75	167,514	180,701	196,266
140	161,827	168,613	181,84	197,451
141	162,904	169,711	182,979	198,635
142	163,98	170,809	184,118	199,819
143	165,056	171,907	185,256	201,002
144	166,132	173,004	186,393	202,184
145	167,207	174,101	187,53	203,366
146	168,283	175,198	188,666	204,547
147	169,358	176,294	189,802	205,727
148	170,432	177,39	190,938	206,907
149	171,507	178,485	192,073	208,086
150	172,581	179,581	193,208	209,265
151	173,655	180,676	194,342	210,443

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
152	174,729	181,77	195,476	211,62
153	175,803	182,865	196,609	212,797
154	176,876	183,959	197,742	213,973
155	177,949	185,052	198,874	215,149
156	179,022	186,146	200,006	216,324
157	180,094	187,239	201,138	217,499
158	181,167	188,332	202,269	218,673
159	182,239	189,424	203,4	219,846
160	183,311	190,516	204,53	221,019
161	184,382	191,608	205,66	222,191
162	185,454	192,7	206,79	223,363
163	186,525	193,791	207,919	224,535
164	187,596	194,883	209,047	225,705
165	188,667	195,973	210,176	226,876
166	189,737	197,064	211,304	228,045
167	190,808	198,154	212,431	229,215
168	191,878	199,244	213,558	230,383
169	192,948	200,334	214,685	231,552
170	194,017	201,423	215,812	232,719
171	195,087	202,513	216,938	233,887
172	196,156	203,602	218,063	235,053
173	197,225	204,69	219,189	236,22

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
174	198,294	205,779	220,314	237,385
175	199,363	206,867	221,438	238,551
176	200,432	207,955	222,563	239,716
177	201,5	209,042	223,687	240,88
178	202,568	210,13	224,81	242,044
179	203,636	211,217	225,933	243,207
180	204,704	212,304	227,056	244,37
181	205,771	213,391	228,179	245,533
182	206,839	214,477	229,301	246,695
183	207,906	215,563	230,423	247,857
184	208,973	216,649	231,544	249,018
185	210,04	217,735	232,665	250,179
186	211,106	218,82	233,786	251,339
187	212,173	219,906	234,907	252,499
188	213,239	220,991	236,027	253,659
189	214,305	222,076	237,147	254,818
190	215,371	223,16	238,266	255,976
191	216,437	224,245	239,386	257,135
192	217,502	225,329	240,505	258,292
193	218,568	226,413	241,623	259,45
194	219,633	227,496	242,742	260,607
195	220,698	228,58	243,86	261,763

	Seuils de significativité			
	0,1	0,05	0,01	0,001
DDL				
196	221,763	229,663	244,977	262,92
197	222,828	230,746	246,095	264,075
198	223,892	231,829	247,212	265,231
199	224,957	232,912	248,329	266,386
200	226,021	233,994	249,445	267,541
250	279,05	287,882	304,94	324,832
300	331,789	341,395	359,906	381,425
350	384,306	394,626	414,474	437,488
400	436,649	447,632	468,724	493,132
450	488,849	500,456	522,717	548,432
500	540,93	553,127	576,493	603,446
550	592,909	605,667	630,084	658,215
600	644,8	658,094	683,516	712,771
650	696,614	710,421	736,807	767,141
700	748,359	762,661	789,974	821,347
750	800,043	814,822	843,029	875,404
800	851,671	866,911	895,984	929,329
850	903,249	918,937	948,848	983,133
900	954,782	970,904	1001,63	1036,826
950	1006,272	1022,816	1054,334	1090,418
1000	1057,724	1074,679	1106,969	1143,917

ANNEXE 5. TABLE DE DURBIN-WATSON AU SEUIL DE SIGNIFICATIVITÉ DE 5% POUR MODÈLES AVEC CONSTANTE

k' = nombre de variables indépendantes (constante exclue)
n = nombre d'observations (taille d'échantillon)
$dL(Lower\ limit)$ = limite inférieure
$dU(Upper\ limit)$ = limite supérieure

1. n = 6 à 15 ; k' = 1 à 3

	k'=1		*k'=2*		*k'=3*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
6	0,61	1,4	—	—	—	—
7	0,7	1,356	0,467	1,896	—	—
8	0,763	1,332	0,559	1,777	0,367	2,287
9	0,824	1,32	0,629	1,699	0,455	2,128
10	0,879	1,32	0,697	1,641	0,525	2,016
11	0,927	1,324	0,758	1,604	0,595	1,928
12	0,971	1,331	0,812	1,579	0,658	1,864
13	1,01	1,34	0,861	1,562	0,715	1,816
14	1,045	1,35	0,905	1,551	0,767	1,779
15	1,077	1,361	0,946	1,543	0,814	1,75

2. n = 6 à 15 ; k' = 4 à 6

	k'=4		*k'=5*		*k'=6*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
6	—	—	—	—	—	—
7	—	—	—	—	—	—
8	—	—	—	—	—	—
9	0,296	2,588	—	—	—	—
10	0,376	2,414	0,243	2,822	—	—
11	0,444	2,283	0,315	2,645	0,203	3,004
12	0,512	2,177	0,38	2,506	0,268	2,832
13	0,574	2,094	0,444	2,39	0,328	2,692
14	0,632	2,03	0,505	2,296	0,389	2,572
15	0,685	1,977	0,562	2,22	0,447	2,471

3. n = 6 à 15 ; k' = 7 à 10

	k'=7		*k'=8*		*k'=9*		*k'=10*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
6	—	—	—	—	—	—	—	—
7	—	—	—	—	—	—	—	—
8	—	—	—	—	—	—	—	—
9	—	—	—	—	—	—	—	—
10	—	—	—	—	—	—	—	—
11	—	—	—	—	—	—	—	—
12	0,17	3,149	—	—	—	—	—	—
13	0,23	2,985	0,15	3,27	—	—	—	—
14	0,29	2,848	0,2	3,11	0,127	3,36	—	—
15	0,34	2,727	0,25	2,98	0,175	3,216	0,111	3,44

4. n = 16 à 200 ; k' = 1 à 3

	k'=1		*k'=2*		*k'=3*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
16	1,11	1,371	0,98	1,54	0,857	1,728
17	1,13	1,381	1,02	1,54	0,897	1,71
18	1,16	1,391	1,05	1,54	0,933	1,696
19	1,18	1,401	1,07	1,54	0,967	1,685
20	1,2	1,411	1,1	1,54	0,998	1,676

	k'=1		*k'=2*		*k'=3*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
21	1,22	1,42	1,13	1,54	1,026	1,669
22	1,24	1,429	1,15	1,54	1,053	1,664
23	1,26	1,437	1,17	1,54	1,078	1,66
24	1,27	1,446	1,19	1,55	1,101	1,656
25	1,29	1,454	1,21	1,55	1,123	1,654
26	1,3	1,461	1,22	1,55	1,143	1,652
27	1,32	1,469	1,24	1,56	1,162	1,651
28	1,33	1,476	1,26	1,56	1,181	1,65
29	1,34	1,483	1,27	1,56	1,198	1,65
30	1,35	1,489	1,28	1,57	1,214	1,65
31	1,36	1,496	1,3	1,57	1,229	1,65
32	1,37	1,502	1,31	1,57	1,244	1,65
33	1,38	1,508	1,32	1,58	1,258	1,651
34	1,39	1,514	1,33	1,58	1,271	1,652
35	1,4	1,519	1,34	1,58	1,283	1,653
36	1,41	1,525	1,35	1,59	1,295	1,654
37	1,42	1,53	1,36	1,59	1,307	1,655
38	1,43	1,535	1,37	1,59	1,318	1,656
39	1,44	1,54	1,38	1,6	1,328	1,658
40	1,44	1,544	1,39	1,6	1,338	1,659
45	1,48	1,566	1,43	1,62	1,383	1,666

	k'=1		*k'=2*		*k'=3*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
50	1,5	1,585	1,46	1,63	1,421	1,674
55	1,53	1,601	1,49	1,64	1,452	1,681
60	1,55	1,616	1,51	1,65	1,48	1,689
65	1,57	1,629	1,54	1,66	1,503	1,696
70	1,58	1,641	1,55	1,67	1,525	1,703
75	1,6	1,652	1,57	1,68	1,543	1,709
80	1,61	1,662	1,59	1,69	1,56	1,715
85	1,62	1,671	1,6	1,7	1,575	1,721
90	1,64	1,679	1,61	1,7	1,589	1,726
95	1,65	1,687	1,62	1,71	1,602	1,732
100	1,65	1,694	1,63	1,72	1,613	1,736
150	1,72	1,747	1,71	1,76	1,693	1,774
200	1,76	1,779	1,75	1,79	1,738	1,799

5. n = 16 à 200 ; k' = 4 à 6

	k'=4		*k'=5*		*k'=6*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
16	0,73	1,935	0,62	2,16	0,502	2,388
17	0,78	1,9	0,66	2,1	0,554	2,318
18	0,82	1,872	0,71	2,06	0,603	2,258

	k'=4		*k'=5*		*k'=6*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
19	0,86	1,848	0,75	2,02	0,649	2,206
20	0,89	1,828	0,79	1,99	0,691	2,162
21	0,93	1,812	0,83	1,96	0,731	2,124
22	0,96	1,797	0,86	1,94	0,769	2,09
23	0,99	1,785	0,9	1,92	0,804	2,061
24	1,01	1,775	0,93	1,9	0,837	2,035
25	1,04	1,767	0,95	1,89	0,868	2,013
26	1,06	1,759	0,98	1,87	0,897	1,992
27	1,08	1,753	1	1,86	0,925	1,974
28	1,1	1,747	1,03	1,85	0,951	1,959
29	1,12	1,743	1,05	1,84	0,975	1,944
30	1,14	1,739	1,07	1,83	0,998	1,931
31	1,16	1,735	1,09	1,83	1,02	1,92
32	1,18	1,732	1,11	1,82	1,041	1,909
33	1,19	1,73	1,13	1,81	1,061	1,9
34	1,21	1,728	1,14	1,81	1,079	1,891
35	1,22	1,726	1,16	1,8	1,097	1,884
36	1,24	1,724	1,18	1,8	1,114	1,876
37	1,25	1,723	1,19	1,8	1,131	1,87
38	1,26	1,722	1,2	1,79	1,146	1,864
39	1,27	1,722	1,22	1,79	1,161	1,859

	k'=4		*k'=5*		*k'=6*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
40	1,29	1,721	1,23	1,79	1,175	1,854
45	1,34	1,72	1,29	1,78	1,238	1,835
50	1,38	1,721	1,34	1,77	1,291	1,822
55	1,41	1,724	1,37	1,77	1,334	1,814
60	1,44	1,727	1,41	1,77	1,372	1,808
65	1,47	1,731	1,44	1,77	1,404	1,805
70	1,49	1,735	1,46	1,77	1,433	1,802
75	1,52	1,739	1,49	1,77	1,458	1,801
80	1,53	1,743	1,51	1,77	1,48	1,801
85	1,55	1,747	1,53	1,77	1,5	1,801
90	1,57	1,751	1,54	1,78	1,518	1,801
95	1,58	1,755	1,56	1,78	1,535	1,802
100	1,59	1,758	1,57	1,78	1,55	1,803
150	1,68	1,788	1,67	1,8	1,651	1,817
200	1,73	1,809	1,72	1,82	1,707	1,831

6. n = 16 à 200 ; k' = 7 à 9

	k'=7		*k'=8*		*k'=9*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
16	0,4	2,624	0,3	2,86	0,222	3,09
17	0,45	2,537	0,36	2,76	0,272	2,975
18	0,5	2,461	0,41	2,67	0,321	2,873
19	0,55	2,396	0,46	2,59	0,369	2,783
20	0,6	2,339	0,5	2,52	0,416	2,704
21	0,64	2,29	0,55	2,46	0,461	2,633
22	0,68	2,246	0,59	2,41	0,504	2,571
23	0,72	2,208	0,63	2,36	0,545	2,514
24	0,75	2,174	0,67	2,32	0,584	2,464
25	0,78	2,144	0,7	2,28	0,621	2,419
26	0,82	2,117	0,74	2,25	0,657	2,379
27	0,85	2,093	0,77	2,22	0,691	2,342
28	0,87	2,071	0,8	2,19	0,723	2,309
29	0,9	2,052	0,83	2,16	0,753	2,278
30	0,93	2,034	0,85	2,14	0,782	2,251
31	0,95	2,018	0,88	2,12	0,81	2,226
32	0,97	2,004	0,9	2,1	0,836	2,203
33	0,99	1,991	0,93	2,09	0,861	2,181
34	1,02	1,978	0,95	2,07	0,885	2,162

	k'=7		*k'=8*		*k'=9*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
35	1,03	1,967	0,97	2,05	0,908	2,144
36	1,05	1,957	0,99	2,04	0,93	2,127
37	1,07	1,948	1,01	2,03	0,951	2,112
38	1,09	1,939	1,03	2,02	0,97	2,098
39	1,1	1,932	1,05	2,01	0,99	2,085
40	1,12	1,924	1,06	2	1,008	2,072
45	1,19	1,895	1,14	1,96	1,089	2,022
50	1,25	1,875	1,2	1,93	1,156	1,986
55	1,29	1,861	1,25	1,91	1,212	1,959
60	1,34	1,85	1,3	1,89	1,26	1,939
65	1,37	1,843	1,34	1,88	1,301	1,923
70	1,4	1,838	1,37	1,87	1,337	1,91
75	1,43	1,834	1,4	1,87	1,369	1,901
80	1,45	1,831	1,43	1,86	1,397	1,893
85	1,47	1,829	1,45	1,86	1,422	1,886
90	1,49	1,827	1,47	1,85	1,445	1,881
95	1,51	1,827	1,49	1,85	1,465	1,877
100	1,53	1,826	1,51	1,85	1,484	1,874
150	1,64	1,832	1,62	1,85	1,608	1,862
200	1,7	1,841	1,69	1,85	1,675	1,863

7. n = 16 à 200 ; k' = 10 à 12

	k'=10		*k'=11*		*k'=12*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
16	0,16	3,304	0,1	3,5	—	—
17	0,2	3,184	0,14	3,38	0,087	3,557
18	0,24	3,073	0,18	3,27	0,123	3,441
19	0,29	2,974	0,22	3,16	0,16	3,335
20	0,34	2,885	0,26	3,06	0,2	3,234
21	0,38	2,806	0,31	2,98	0,24	3,141
22	0,42	2,735	0,35	2,9	0,281	3,057
23	0,47	2,67	0,39	2,83	0,322	2,979
24	0,51	2,613	0,43	2,76	0,362	2,908
25	0,54	2,56	0,47	2,7	0,4	2,844
26	0,58	2,513	0,51	2,65	0,438	2,784
27	0,62	2,47	0,54	2,6	0,475	2,73
28	0,65	2,431	0,58	2,56	0,51	2,68
29	0,68	2,396	0,61	2,52	0,544	2,634
30	0,71	2,363	0,64	2,48	0,577	2,592
31	0,74	2,333	0,67	2,44	0,608	2,553
32	0,77	2,306	0,7	2,41	0,638	2,517
33	0,8	2,281	0,73	2,38	0,668	2,484
34	0,82	2,257	0,76	2,36	0,695	2,454

	k'=10		*k'=11*		*k'=12*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
35	0,85	2,236	0,78	2,33	0,722	2,425
36	0,87	2,216	0,81	2,31	0,748	2,398
37	0,89	2,197	0,83	2,29	0,772	2,374
38	0,91	2,18	0,85	2,27	0,796	2,351
39	0,93	2,164	0,88	2,25	0,819	2,329
40	0,95	2,149	0,9	2,23	0,84	2,309
45	1,04	2,088	0,99	2,16	0,938	2,225
50	1,11	2,044	1,06	2,1	1,019	2,163
55	1,17	2,01	1,13	2,06	1,087	2,116
60	1,22	1,984	1,18	2,03	1,145	2,079
65	1,27	1,964	1,23	2,01	1,195	2,049
70	1,31	1,948	1,27	1,99	1,239	2,026
75	1,34	1,935	1,31	1,97	1,277	2,006
80	1,37	1,925	1,34	1,96	1,311	1,991
85	1,4	1,916	1,37	1,95	1,342	1,977
90	1,42	1,909	1,4	1,94	1,369	1,966
95	1,44	1,903	1,42	1,93	1,394	1,956
100	1,46	1,898	1,44	1,92	1,416	1,948
150	1,59	1,877	1,58	1,89	1,564	1,908
200	1,67	1,874	1,65	1,89	1,643	1,896

8. n = 16 à 200 ; k' = 13 à 15

	k'=13		*k'=14*		*k'=15*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
16	—	—	—	—	—	—
17	—	—	—	—	—	—
18	0,08	3,603	—	—	—	—
19	0,11	3,496	0,07	3,64	—	—
20	0,15	3,395	0,1	3,54	0,063	3,676
21	0,18	3,3	0,13	3,45	0,091	3,583
22	0,22	3,211	0,17	3,36	0,12	3,495
23	0,26	3,128	0,2	3,27	0,153	3,409
24	0,3	3,053	0,24	3,19	0,186	3,327
25	0,34	2,983	0,28	3,12	0,221	3,251
26	0,37	2,919	0,31	3,05	0,256	3,179
27	0,41	2,859	0,35	2,99	0,291	3,112
28	0,45	2,805	0,38	2,93	0,325	3,05
29	0,48	2,755	0,42	2,87	0,359	2,992
30	0,51	2,708	0,45	2,82	0,392	2,937
31	0,55	2,665	0,48	2,78	0,425	2,887
32	0,58	2,625	0,52	2,73	0,457	2,84
33	0,61	2,588	0,55	2,69	0,488	2,796
34	0,63	2,554	0,58	2,65	0,518	2,754

	k'=13		*k'=14*		*k'=15*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
35	0,66	2,521	0,6	2,62	0,547	2,716
36	0,69	2,492	0,63	2,59	0,575	2,68
37	0,71	2,464	0,66	2,56	0,602	2,646
38	0,74	2,438	0,68	2,53	0,628	2,614
39	0,76	2,413	0,71	2,5	0,653	2,585
40	0,79	2,391	0,73	2,47	0,678	2,557
45	0,89	2,296	0,84	2,37	0,788	2,439
50	0,97	2,225	0,93	2,29	0,882	2,35
55	1,05	2,17	1	2,23	0,961	2,281
60	1,11	2,127	1,07	2,18	1,029	2,227
65	1,16	2,093	1,12	2,14	1,088	2,183
70	1,21	2,066	1,17	2,11	1,139	2,148
75	1,25	2,043	1,22	2,08	1,184	2,118
80	1,28	2,024	1,25	2,06	1,224	2,093
85	1,32	2,009	1,29	2,04	1,26	2,073
90	1,34	1,995	1,32	2,03	1,292	2,055
95	1,37	1,984	1,35	2,01	1,321	2,04
100	1,39	1,974	1,37	2	1,347	2,026
150	1,55	1,924	1,54	1,94	1,519	1,956
200	1,63	1,908	1,62	1,92	1,61	1,931

9. n = 16 à 200 ; k' = 16 à 18

	k'=16		*k'=17*		*k'=18*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
16	—	—	—	—	—	—
17	—	—	—	—	—	—
18	—	—	—	—	—	—
19	—	—	—	—	—	—
20	—	—	—	—	—	—
21	0,06	3,705	—	—	—	—
22	0,08	3,619	0,05	3,73	—	—
23	0,11	3,535	0,08	3,65	0,048	3,753
24	0,14	3,454	0,1	3,57	0,07	3,678
25	0,17	3,376	0,13	3,49	0,094	3,604
26	0,21	3,303	0,16	3,42	0,12	3,531
27	0,24	3,233	0,19	3,35	0,149	3,46
28	0,27	3,168	0,22	3,28	0,178	3,392
29	0,31	3,107	0,25	3,22	0,208	3,327
30	0,34	3,05	0,29	3,16	0,238	3,266
31	0,37	2,996	0,32	3,1	0,269	3,208
32	0,4	2,946	0,35	3,05	0,299	3,153
33	0,43	2,899	0,38	3	0,329	3,1
34	0,46	2,854	0,41	2,95	0,359	3,051

	k'=16		*k'=17*		*k'=18*	
n	*dL*	*dU*	*dL*	*dU*	*dL*	*dU*
35	0,49	2,813	0,44	2,91	0,388	3,005
36	0,52	2,774	0,47	2,87	0,417	2,961
37	0,55	2,738	0,5	2,83	0,445	2,92
38	0,58	2,703	0,52	2,79	0,472	2,88
39	0,6	2,671	0,55	2,76	0,499	2,843
40	0,63	2,641	0,58	2,72	0,525	2,808
45	0,74	2,512	0,69	2,59	0,644	2,659
50	0,84	2,414	0,79	2,48	0,747	2,544
55	0,92	2,338	0,88	2,4	0,836	2,454
60	0,99	2,278	0,95	2,33	0,913	2,382
65	1,05	2,229	1,02	2,28	0,98	2,323
70	1,11	2,189	1,07	2,23	1,038	2,275
75	1,15	2,156	1,12	2,2	1,09	2,235
80	1,2	2,129	1,17	2,17	1,136	2,201
85	1,23	2,105	1,21	2,14	1,177	2,172
90	1,27	2,085	1,24	2,12	1,213	2,148
95	1,3	2,068	1,27	2,1	1,247	2,126
100	1,32	2,053	1,3	2,08	1,277	2,108
150	1,5	1,972	1,49	1,99	1,474	2,006
200	1,6	1,943	1,59	1,96	1,576	1,967

10. n = 16 à 200 ; k' = 19 ou 20

	k'=19		*k'=20*	
n	*dL*	*dU*	*dL*	*dU*
16	—	—	—	—
17	—	—	—	—
18	—	—	—	—
19	—	—	—	—
20	—	—	—	—
21	—	—	—	—
22	—	—	—	—
23	—	—	—	—
24	0,04	3,773	—	—
25	0,07	3,702	0,04	3,79
26	0,09	3,632	0,06	3,72
27	0,11	3,563	0,08	3,66
28	0,14	3,495	0,1	3,59
29	0,17	3,431	0,13	3,53
30	0,2	3,368	0,16	3,47
31	0,22	3,309	0,18	3,41
32	0,25	3,252	0,21	3,35
33	0,28	3,198	0,24	3,29
34	0,31	3,147	0,27	3,24

	k'=19		*k'=20*	
n	*dL*	*dU*	*dL*	*dU*
35	0,34	3,099	0,3	3,19
36	0,37	3,053	0,32	3,14
37	0,4	3,009	0,35	3,1
38	0,42	2,968	0,38	3,05
39	0,45	2,929	0,4	3,01
40	0,48	2,829	0,43	2,97
45	0,6	2,733	0,55	2,81
50	0,7	2,61	0,66	2,68
55	0,8	2,512	0,75	2,57
60	0,87	2,434	0,84	2,49
65	0,94	2,371	0,91	2,42
70	1,01	2,318	0,97	2,36
75	1,06	2,275	1,03	2,32
80	1,11	2,238	1,08	2,28
85	1,15	2,206	1,12	2,24
90	1,19	2,179	1,16	2,21
95	1,22	2,156	1,2	2,19
100	1,25	2,135	1,23	2,16
150	1,46	2,023	1,44	2,04
200	1,57	1,979	1,55	1,99

ANNEXE 6. TABLE DU COEFFICIENT DE CORRÉLATION DE PEARSON AUX SEUILS DE SIGNIFICATIVITÉ DE 10%, 5% ET 1%, POUR 1 À 300 DEGRÉS DE LIBERTÉ

$DDL = Taille\ d'échantillon - 2$

	0,1	0,05	0,01
DDL			
1	0,988	0,997	0,999
2	0,9	0,95	0,99
3	0,805	0,878	0,959
4	0,729	0,811	0,917
5	0,669	0,754	0,875
6	0,621	0,707	0,834
7	0,584	0,666	0,798
8	0,549	0,632	0,765
9	0,521	0,602	0,735
10	0,497	0,576	0,708
11	0,476	0,553	0,684
12	0,458	0,532	0,661
13	0,441	0,514	0,641
14	0,426	0,497	0,623
15	0,412	0,482	0,606

	0,1	**0,05**	**0,01**
16	0,4	0,468	0,59
17	0,389	0,456	0,575
18	0,378	0,444	0,561
19	0,369	0,433	0,549
20	0,36	0,423	0,537
21	0,352	0,413	0,526
22	0,344	0,404	0,515
23	0,337	0,396	0,505
24	0,33	0,388	0,496
25	0,323	0,381	0,487
26	0,317	0,374	0,479
27	0,311	0,367	0,471
28	0,306	0,361	0,463
29	0,301	0,355	0,456
30	0,296	0,349	0,449
35	0,275	0,325	0,418
40	0,257	0,304	0,393
45	0,243	0,288	0,372
50	0,231	0,273	0,354
60	0,211	0,25	0,325
70	0,195	0,232	0,303
80	0,183	0,217	0,283

	0,1	**0,05**	**0,01**
90	0,173	0,205	0,267
100	0,164	0,195	0,254
150	0,134	0,159	0,208
300	0,095	0,113	0,148

L'ouvrage

Cet ouvrage est un compagnon de route de l'éducationniste dans l'exploration, la maîtrise progressive et l'usage des méthodes quantitatives pour l'analyse et la recherche en éducation et formation. Il est conçu de façon à accompagner l'utilisateur, des premiers stades de la formation jusqu'à l'autonomie. À cette fin, l'étendue du contenu présenté est des plus larges, allant des notions de base de statistique descriptive jusqu'à des méthodes avancées comme les modèles d'équations structurelles. S'adressant à un public non-familier des méthodes quantitatives, l'ouvrage privilégie une approche pratique, l'utilisation de logiciels statistiques, et des exemples illustrant les règles de procédure et d'interprétation.

L'auteur

Guy Tchibozo est Professeur des Universités et membre du Laboratoire des Sciences de l'éducation de l'Université de Strasbourg (LISEC). Ses travaux portent sur l'analyse des relations formation-emploi. Il intervient comme Expert au Centre européen pour le développement de la formation professionnelle (Cedefop), où il dirige le *Tableau de bord de la mobilité*.

30 € | France

ISBN 978-952-340-501-1

9 789523 405011

Atramenta

Lire, écrire, partager

www.atramenta.net