paramètres

Robert R. Haccoun • Denis Cousineau

Statistiques

Concepts et applications

DEUXIÈME ÉDITION REVUE ET AUGMENTÉE

Les Presses de l'Université de Montréal

STATISTIQUES

Page laissée blanche

Robert R. Haccoun
et
Denis Cousineau

STATISTIQUES

Concepts et applications

Deuxième édition revue et augmentée

Les Presses de l'Université de Montréal

Catalogage avant publication de Bibliothèque et Archives nationales du Québec et Bibliothèque et Archives Canada

Haccoun, Robert R.

Statistiques : concepts et applications

2e éd.

Publ. à l'origine dans la coll. : Paramètres. 2007.

ISBN 978-2-7606-2113-8

eISBN 978-2-7606-2584-6

1. Statistique mathématique. 2. Probabilités. 3. Statistique. I. Cousineau, Denis, 1967- . II. Titre.

QA276.H185 2010 519.5 C2009-942664-1

Dépôt légal : 1er trimestre 2010

Bibliothèque et Archives nationales du Québec

Les Presses de l'Université de Montréal reconnaissent l'aide financière du gouvernement du Canada par l'entremise du Programme d'aide au développement de l'industrie de l'édition (PADIÉ) pour leurs activités d'édition.

Les Presses de l'Université de Montréal remercient de leur soutien financier le Conseil des arts du Canada et la Société de développement des entreprises culturelles du Québec (SODEC).

Imprimé au Canada en janvier 2010

À la « grande fille » de son Pappy, Orli Haya Abramson.
Robert Haccoun

À Élysabeth Aguila et Richard Shifrin, pour leur patience ∞
Denis Cousineau

Page laissée blanche

TABLE DES MATIÈRES

Page laissée blanche

AVANT-PROPOS

Si la plupart des programmes de premier cycle exigent que les étudiants suivent un cours de base en méthodes statistiques, c'est que cette formation est essentielle pour maîtriser les aspects scientifiques d'une discipline, notamment en sciences sociales.

Ce manuel est d'abord destiné aux étudiants qui suivront peut-être un seul cours de statistiques dans leur formation, mais il pourra également servir d'entrée en matière à ceux qui suivront des cours plus avancés. De la construction d'une distribution d'effectifs jusqu'à l'analyse de variance factorielle, il explique les fondements logiques, les résultats et les interprétations que les techniques statistiques permettent et celles qu'elles ne permettent pas. Nous avons délibérément laissé de côté les méthodes plus avancées qui permettent d'analyser des données expérimentales ou corrélationnelles complexes.

L'étude des statistiques suscite souvent des appréhensions qui ne facilitent pas l'apprentissage ; les formules mathématiques d'apparence complexe peuvent en rebuter plus d'un. Nous avons donc voulu proposer une approche intuitive et graduelle qui soit rassurante. Bien entendu, les statistiques s'expriment par des formules, présentées dans ce volume, mais il ne faut pas perdre de vue que ces formules servent d'abord à rendre des concepts plus concrets. C'est pourquoi ce sont les concepts et non leur expression mathématique qui sont au cœur de notre approche. *Nous préférons utiliser le concept pour expliquer la formule, plutôt que d'utiliser la formule pour expliquer le concept.*

Nous croyons que la logique statistique peut être plus facilement comprise lorsqu'elle fait appel au raisonnement de l'étudiant et qu'elle s'appuie

sur des exemples capables de susciter son intérêt. C'est dans cet esprit que l'humour est parfois mis à contribution.

Cette deuxième édition maintient l'approche et l'esprit de l'édition originale, mais elle contient des changements importants : non seulement la présentation graphique et les textes ont été entièrement révisés, mais plusieurs sections ont été refondues pour les rendre plus claires, notamment celles décrivant les aspects plus complexes portant sur l'inférence statistique. On y trouve aussi de nouveaux contenus, dont un chapitre additionnel sur l'*analyse non paramétrique*.

Comme dans la première édition, chaque chapitre est ponctué de « quiz rapides » qui permettent aux étudiants de vérifier leur niveau de maîtrise des concepts et se termine par des questions à choix multiples. On y trouve évidemment les réponses aux uns et aux autres.

Le site Internet (www.pum.umontreal.ca/ca/fiches/978-2-7606-2113-8.html) reste actif. On y trouve, pour chaque chapitre du livre, des banques de données, des exercices et les commandes d'analyse pour le logiciel SPSS ; on y trouve également des discussions sur l'interprétation des résultats produits par le logiciel.

Nous tenons à exprimer nos remerciements aux personnes qui nous ont aidés à préparer cette nouvelle édition. Merci d'abord à Corinne Zacharyas pour sa lecture aussi attentive que généreuse, et à M. Élie Haccoun pour ses précieux conseils. Merci tout spécialement à nos familles et amis qui ont eu à supporter les écarts d'humeur qu'implique une entreprise de ce genre. Enfin, nous exprimons notre reconnaissance aux Presses de l'Université de Montréal, à son directeur, M. Antoine Del Busso, et à notre éditrice, Mme Natacha Monnier, pour leur soutien indéfectible.

Robert R. Haccoun
Denis Cousineau

COMMENT UTILISER CET OUVRAGE

Cet ouvrage explique les concepts et la logique statistiques ; il est principalement destiné aux étudiants du premier cycle universitaire qui étudient la statistique. La présentation des concepts et techniques se fait de façon progressive : les premiers chapitres s'attachent à des aspects élémentaires (la construction d'une distribution, la nature des mesures de tendances centrales, les indices de dispersion, etc.), les chapitres suivants passent à des aspects plus élaborés (la logique de l'inférence, le test t, l'analyse de variance, etc.).

Afin de favoriser le développement graduel d'une compréhension intégrée de la statistique, chaque chapitre débute par un exemple qui fait le lien avec les connaissances acquises dans le chapitre antérieur. Chaque nouvel élément ou concept abordé dans le chapitre est illustré par des exemples concrets et simples, la plupart étant extraits de la vie quotidienne, certains étant même plutôt humoristiques. Il s'agit donc d'une approche volontairement « conviviale » de la statistique qui vise à susciter l'intérêt des étudiants pour lesquels cette matière semble trop souvent rébarbative.

Cependant, cet ouvrage n'est pas un livre de mathématiques ! Les présentations et les explications n'exigent généralement pas une formation poussée en mathématiques. Si on y présente, comme il se doit, des formules statistiques de façon formelle, les explications qui les accompagnent décrivent leurs logiques plutôt que leurs dérivations algébriques. Et lorsque certaines preuves mathématiques sont requises, elles sont isolées du texte et placées dans des encadrés. Dans certains encadrés, le lecteur trouvera de brèves biographies, des anecdotes et d'autres exposés qui illustrent ou montrent l'origine des concepts qui sont abordés dans les chapitres.

Après avoir lu chaque chapitre, le lecteur devra normalement être en mesure de comprendre l'utilisation de la technique décrite, sa logique et les

interprétations qui peuvent ou non être faites à partir des résultats qu'elle génère.

Dans tous les chapitres, nous avons présenté des « quiz rapides ». Ces courts exercices permettent au lecteur de tester ses connaissances au fur et à mesure de la lecture de l'ouvrage. Les réponses à ces exercices se trouvent à la fin du volume.

À la fin de chaque chapitre, une série de questions à choix multiple est présentée et les réponses figurent à leur suite. Ces questions permettent aux lecteurs d'évaluer leur connaissance des concepts dans une forme qui s'apparente à celle des examens universitaires. Les réponses à ces questions exigent peu de calculs arithmétiques, voire n'en exigent aucun. L'accent est mis sur la compréhension des concepts et des techniques plutôt que sur la computation mécanique des formules pertinentes.

Comme accompagnement à ce manuel, des fichiers contenant plusieurs banques de données, des explications du logiciel d'analyse statistique SPSS et des exercices pouvant être analysés avec ce logiciel sont disponibles sur le site Internet des Presses de l'Université de Montréal à l'adresse suivante : www.pum.umontreal.ca/ca/fiches/978-2-7606-2113-8.html. Les principales règles de syntaxe et les fonctionnalités du logiciel SPSS qui se rattachent à chacun des 13 chapitres du volume sont expliquées et exemplifiées dans cette partie du site Internet. Les professeurs qui ne prévoient pas de sessions de pratique pour les étudiants pourront simplement ignorer cet ajout au site Internet.

Ainsi, en lisant chacun des chapitres de cet ouvrage, en faisant les quiz rapides, en répondant aux questions de fin de chapitre et en exécutant les exercices du site Internet, le lecteur pourra mettre immédiatement en pratique ses connaissances.

CHAPITRE 1
LA DESCRIPTION DES DONNÉES

Page laissée blanche

CHAPITRE 1

LA DESCRIPTION DES DONNÉES

Les statistiques sont un inventaire de techniques et de procédures qui permettent d'organiser et de faire le sommaire d'une masse d'informations afin d'en dégager des conclusions utiles à la compréhension d'un phénomène.

LA DESCRIPTION ET L'INFÉRENCE EN STATISTIQUE

Les statistiques se divisent en deux branches, complémentaires et interreliées : celles qui permettent une description des informations ; et celles qui permettent, à partir de ces descriptions, de faire des inférences. Les *statistiques descriptives* font le sommaire et simplifient l'information dans le but de la clarifier et de révéler ses tendances lourdes. L'*inférence statistique* est une série de procédures qui se servent de ces descriptions pour tirer des conclusions plus générales sur le phénomène à l'étude.

Tous les phénomènes mesurés peuvent être analysés statistiquement, à condition que l'information soit exprimée numériquement. C'est donc dire que les statistiques ne doivent utiliser que des informations quantitatives.

Données quantitatives

Nous avons tous l'habitude de mesurer les choses en nous servant de chiffres. Notre âge, notre poids, le montant de nos dettes, le nombre d'enfants dans notre famille ou notre température corporelle peuvent tous être mesurés quantitativement. Les attitudes, les opinions, les croyances, la personnalité et les comportements peuvent aussi être mesurés quantitativement. En principe, on se sert d'un questionnaire comme instrument de mesure. Chaque réponse possible à une question est décrite par un chiffre. Par exemple, dans un questionnaire qui mesure la satisfaction au travail, on pourrait demander aux employés d'une compagnie d'indiquer leur degré d'accord ou de désaccord avec des phrases telles que : « Aller au travail m'est très désagréable », « Si je le pouvais, je donnerais ma démission aujourd'hui », etc. Les réponses possibles sont : « Totalement en accord » (indexé par le chiffre 1), « Plutôt en accord » (chiffre 2), « Ni en accord ni en désaccord » (chiffre 3), « Plutôt en désaccord » (chiffre 4), et « Totalement en désaccord » (chiffre 5). Ainsi, les personnes ayant plus de satisfaction au travail auraient tendance à être en désaccord avec ces énoncés et, par conséquent, fourniraient des réponses plus près de 5, alors que les réponses des personnes ayant peu de satisfaction au travail seraient concentrées autour de 1. Ces procédures permettent de « quantifier » les attitudes, de les exprimer numériquement. Ce faisant, il devient possible d'en faire une analyse statistique.

Les statistiques que nous allons étudier dans ce livre sont mises à profit pour faciliter la compréhension de phénomènes aussi diversifiés que la croissance économique d'une société, les comportements sociaux, l'efficacité d'une technique chirurgicale, ou même les réactions chimiques. Les cognitivistes utilisent les statistiques pour déduire l'organisation du cerveau et ses liens avec la pensée. Les psychologues font appel aux statistiques afin de mieux comprendre les caractéristiques individuelles comme la personnalité, l'intelligence ou le comportement déviant à l'école. Les sociologues s'en servent pour mieux comprendre la violence sociale ou la relation entre les idéologies et l'éducation. Les experts en marketing y recourent afin d'analyser et d'améliorer les stratégies de mise en marché. Dans le monde des affaires, on s'en sert pour planifier les inventaires ou pour établir les marges de profit. Et ce sont les statistiques qui déterminent, en grande partie, les décisions des gouvernements.

Nous lisons tous les jours dans les journaux des résultats de sondages. Ces sondages guident, dans une certaine mesure, les décisions concernant les activités des institutions, publiques ou privées, l'impact de ces activités se répercutant sur presque chacun de nous : étudiants, consommateurs, travailleurs. Tous, nous sommes personnellement affectés par les statistiques

et un grand nombre de décisions qui nous touchent trouvent leur origine dans le résultat d'une analyse statistique.

En analyse statistique, les informations (quantitatives) sont recueillies, organisées et soumises à des procédures arithmétiques. Le résultat final de ces procédés est une simplification de l'information qui permet de dégager des tendances afin de mieux comprendre le phénomène étudié et d'en tirer des conclusions utiles. Les statistiques nous permettent de voir la forêt malgré les arbres !

Les analyses statistiques offrent la possibilité de mieux comprendre les caractéristiques des individus (l'intelligence, la sociabilité), des groupes (la performance des équipes ou la compétitivité des entreprises), ou des communautés plus larges (le degré de pauvreté dans différents pays, le coût des logements dans différentes villes). La source des données (les individus, les équipes, les entreprises, les villes, etc.) se nomme le *sujet d'analyse* ou l'*unité d'analyse, ou encore l'observation*. Le sujet d'analyse définit donc l'origine de l'information. Les conclusions, par conséquent, s'appliqueront exclusivement à cette source. Ainsi, lorsque nous mesurons la densité des populations dans les villes, le sujet d'analyse est la ville et les conclusions s'appliquent aux villes. Lorsque les informations sont recueillies auprès des individus, le sujet d'analyse est l'individu. Si nous mesurons le comportement des chiens, le sujet d'analyse est le chien. Chaque sujet d'analyse fournit une ou des observations. Ainsi, lorsque nous analysons l'intention de vote de 1 000 citoyens, nous avons 1 000 observations.

L'ORGANISATION D'UNE BANQUE DE DONNÉES POUR L'ANALYSE STATISTIQUE[1]

Les informations fournies par les sujets sont généralement organisées sous la forme d'un tableau comprenant des colonnes et des lignes (rangées). Chaque sujet d'analyse (chaque répondant à un sondage par exemple) occupe une ligne du tableau. Les variables (chaque question du sondage) occupent les colonnes. À l'intersection de chaque colonne et de chaque

1. Le site Internet du livre (www.pum.umontreal.ca/ca/fiches/978-2-7606-2113-8.html) contient plusieurs banques de données, y compris celle portant sur les salaires des joueurs de hockey. Un extrait de ce dossier est reproduit au Tableau 1.1.

rangée, nous trouvons une cellule. C'est dans cette cellule que sera placée la valeur numérique qui sera analysée.

Chaque colonne contient une seule variable. Si nous demandons à 10 personnes de fournir une réponse à trois questions, la banque de données comprendra 10 lignes et 3 colonnes. En lisant horizontalement, sur une ligne, nous observons la valeur obtenue par un sujet d'analyse pour chaque variable. À l'inverse, avec une lecture verticale, le long d'une colonne, nous obtenons la réponse de tous les sujets sur une variable. Par exemple, le Tableau 1.1 présente une banque de données sur les salaires annuels des joueurs de hockey. Chaque ligne du tableau identifie un joueur de hockey spécifique. Chaque colonne définit une variable différente. À l'intersection de chaque rangée et de chaque colonne, nous trouvons une cellule qui indique la valeur de la variable X pour le joueur Y. Ainsi, en lisant les colonnes consacrées à l'athlète décrit dans la rangée 1, nous voyons son nom (variable « joueur »), son équipe (variable « équipe »), la position qu'il occupe au jeu (variable « position ») et son salaire annuel (variable « salaire »).

Tableau 1.1
Banque de données organisées pour une analyse statistique

Joueurs	*Équipes*	*Position*	*Salaire ($US)*
Joueur 1	Flames de Calgary	G	400 000
Joueur 2	Canadiens de Montréal	C	3 300 000
Joueur 3	Predators de Nashville	G	900 000
Joueur 4	Sénateurs d'Ottawa	G	1 500 000
Joueur 5	Canucks de Vancouver	G	425 000
Joueur 6	Avalanches du Colorado	C	11 000 000
Joueur 7	Blues de Saint-Louis	G	165 000

Il est relativement facile d'organiser des banques de données en se servant de logiciels tels que Word ou Excel, ou de programmes d'analyse statistique spécialisés tels que SPSS ou SAS. L'important est de se souvenir de cette règle : *chaque rangée définit un sujet différent et chaque colonne, une variable diffé-*

rente. Une fois la banque de données construite, en se servant d'un logiciel de traitement de texte (comme le Bloc-notes) ou d'un chiffrier (tel qu'Excel), il est facile de l'importer dans un logiciel d'analyse statistique (tel que SPSS).

LES VARIABLES

Une variable est une caractéristique que l'on mesure et qui sera soumise à des analyses. On l'appelle *variable* parce que les sujets d'analyse peuvent lui attribuer des valeurs différentes. L'âge, le sexe, le quotient intellectuel (QI) et la condition sociale, entre autres, sont des variables. Le QI est une variable parce qu'il peut être différent selon les personnes. L'anxiété est une variable puisque certains peuvent être très anxieux, d'autres très calmes et d'autres encore peuvent se situer quelque part entre ces deux extrêmes. Le genre – homme versus femme – est lui aussi une variable.

Lorsque toutes ces valeurs sont identiques, la variable prend le nom de *constante*. Si la variable ne peut prendre qu'une valeur unique, cette variable devient une constante. Par exemple, lorsque nous mesurons le degré de sociabilité des femmes, le sexe, qui est habituellement une variable, devient une constante (toutes les personnes mesurées étant des femmes).

Quiz rapide 1.1
Le site Internet du livre contient une banque de données sur les joueurs de la Ligue nationale de hockey (les sujets de cette analyse). Prenez seulement la partie des données qui porte sur les Canucks de Vancouver et identifiez les variables et les constantes. Si vous n'utilisez pas le site WEB, répondez au quiz en vous référant au Tableau 1.1.

Les variables indépendantes et les variables dépendantes

Les chercheurs ou intervenants qui font de l'expérimentation distinguent les *variables indépendantes* des *variables dépendantes*. L'expérimentateur contrôle ou choisit la façon dont la variable indépendante varie afin de déterminer le degré d'impact que cette variable indépendante a sur la variable dépendante. La variable dépendante est la réaction du sujet d'analyse à la variable indépendante. Par exemple, une professeure cherche à déterminer si les étudiants réussissent mieux leurs examens lorsqu'elle y convertit

des questions formulées de façon habituelle en questions humoristiques. Elle prépare alors deux examens, l'un commence par 10 questions humoristiques, et l'autre, par 10 questions neutres. La forme d'examen (avec ou sans questions humoristiques) est la variable indépendante (on peut remarquer qu'il s'agit d'une variable parce que nous avons deux valeurs possibles pour l'examen : humoristique ou non). La note obtenue à l'examen devient la variable dépendante (c'est une variable, parce que les étudiants peuvent obtenir différentes notes, et elle est dépendante, parce que nous posons l'hypothèse que la note obtenue dépend du type d'examen).

LES ÉCHELLES DE MESURE

Il faut mesurer une variable afin d'en faire l'analyse. Cette mesure consiste à fournir une valeur numérique qui indique la position de l'observation sur la ou les variables. Par exemple, pour mesurer le poids d'une personne, on se sert d'un pèse-personne qui indique une valeur numérique, et cette valeur décrit son poids (Jeanne pèse 50 kg et Harold 70 kg : la variable « poids » contient les valeurs « 50 » et « 70 »). Pour les résultats d'une course de chevaux, la mesure numérique est définie par l'ordre d'arrivée : on attribue la valeur 1 au cheval le plus rapide, 2 au suivant, etc. Dans ce cas, la variable « course » est composée d'observations qui indiquent l'ordre d'arrivée. La signification des valeurs numériques que nous attribuons aux différents types de variables n'est pas toujours la même : obtenir 1 % à un « examen » n'est pas la même chose que d'être 1er de classe à la variable « résultat », même si le même code numérique (« 1 ») est attribué aux deux valeurs !

Les variables peuvent contenir différents types d'informations. Nous appelons le type d'informations l'« échelle » de mesure. Il existe quatre types d'échelles de mesure : *nominale, ordinale, à intervalles* et *de rapport.* Il est important de reconnaître l'échelle de mesure de chaque variable, car les procédures statistiques utilisables en dépendent.

Les variables (échelles) nominales

Certaines variables ne peuvent qu'indiquer (nommer) la catégorie à laquelle chaque observation appartient. Ces variables s'appellent ainsi des *variables*

nominales (ou variables catégorielles). Le « prénom » est une variable nominale qui est elle-même composée d'un grand nombre de catégories, chacune décrivant un nom différent. La couleur des yeux est une autre variable nominale. Puisqu'il n'existe qu'un nombre limité de couleurs, cette variable nominale sera composée d'un nombre moindre de catégories que la variable nominale « prénom ». L'origine ethnique, la ville de naissance ou le champ d'études sont d'autres variables qui se mesurent sur des échelles nominales.

Chaque observation d'une variable nominale n'appartient obligatoirement qu'à une seule catégorie : par exemple, pour la mesure du genre, une variable nominale *dichotomique* (ayant deux catégories), chaque observation ne peut prendre que l'une ou l'autre de deux valeurs : « femme » ou « homme », mais pas les deux. Par contre, la religion est une variable nominale *multichotomique* (ayant plusieurs catégories), car elle peut contenir beaucoup plus de catégories : on peut être catholique, juif, protestant, musulman, athée, etc.

Il est souvent pratique d'identifier les catégories d'une variable nominale par des codes numériques (yeux bleus = 1, yeux verts = 2, etc.). La variable nominale servant exclusivement à identifier la catégorie à laquelle chaque observation appartient, ses différentes valeurs ne représentent que des étiquettes, des codes numériques. Le chiffre qui code chaque valeur de la variable nominale est arbitraire – nous pourrions inscrire « 17 » et « 145 » pour catégoriser les personnes aux yeux bleus et celles aux yeux verts pour la variable « couleur des yeux ». Par conséquent, les informations contenues dans les variables nominales n'ont aucune propriété mathématique. Ces valeurs ne peuvent être ni soustraites ni additionnées et, bien sûr, nous ne pouvons pas calculer leur moyenne. Puisque les valeurs d'une variable n'ont pas de signification mathématique particulière, nous ne pouvons que compter le nombre de répondants qui se situent dans chacune des catégories. Par exemple, la variable nominale « intention de vote à la prochaine élection » pourrait contenir quatre catégories : les partis politiques Rouge, Vert, Jaune et Bleu. Règle générale, l'analyse statistique pour cette variable consistera exclusivement à compter le nombre (ou la proportion) de répondants qui entendent voter pour chaque parti.

Puisque la variable nominale identifie les catégories, il importe, lorsque nous codons les valeurs d'une variable nominale, d'associer les observa-

tions à la bonne catégorie. Pour ce faire, il faut respecter les deux règles suivantes : a) la même valeur numérique est attribuée à toutes les observations qui appartiennent à la même catégorie nominale (les fumeurs reçoivent le code « 1 » et les non-fumeurs « 2 ») ; b) une observation qui appartient à une catégorie de la variable ne peut appartenir à une autre catégorie (une personne qui fume occasionnellement n'appartient ni à la catégorie 1 ni à la catégorie 2 ; pour l'analyser, il nous faudra la définir par une autre étiquette, par exemple la valeur « 3 »).

Quiz rapide 1.2
Vous devez coder la couleur des yeux de 1 000 personnes. Vous établissez les catégories « bleus = 1 », « bruns = 2 » et « verts = 3 ». Une personne a un œil bleu et l'autre vert. Comment allez-vous coder les yeux de cet individu ?

Les variables (échelles) ordinales

Les *variables ordinales* permettent de mesurer la position de chaque observation par rapport aux autres observations sur une variable. Cette position se nomme le *rang*. Le résultat obtenu à une course de chevaux est mesuré sur une échelle ordinale, car ce qui importe est l'ordre d'arrivée des chevaux, leurs rangs respectifs. Ainsi, la valeur « 1 » est attribuée au cheval qui traverse le premier la ligne d'arrivée, la valeur « 2 » au suivant, etc. Dans une course comprenant 8 chevaux, le dernier cheval obtient la valeur « 8 » sur la mesure indiquant sa position (par rapport à celles des autres chevaux) au fil d'arrivée. Contrairement aux variables nominales, le chiffre numérique attribué à chaque observation n'est pas arbitraire, mais a une signification. Cette signification représente la position de chaque observation relative aux autres observations. Ainsi, aux Jeux olympiques, nous savons que le nageur qui gagne la médaille d'or a nagé plus vite que celui qui a obtenu la médaille d'argent, et que le médaillé de bronze est moins rapide que les deux autres. Les codes numériques que nous assignons (1, 2 et 3) représentent une différence réelle : l'athlète qui obtient la valeur 1 à la variable « résultat » a nagé plus vite que tous ses compétiteurs.

Les variables ordinales ne sont cependant pas en mesure de déterminer l'ampleur des différences entre les observations. Ainsi, nous ne savons pas

si le médaillé d'or a gagné la course avec une longue ou une très courte avance sur les autres médaillés. Techniquement, nous disons que les *variables ordinales indiquent le rang, mais elles n'indiquent pas la magnitude des différences entre les rangs*. Par conséquent, avec une mesure ordinale, la différence entre le rang 1 et le rang 2 n'est pas nécessairement égale à la différence entre le rang 2 et le rang 3.

Il existe de nombreuses situations où l'utilisation de variables ordinales est nécessaire. Quand ils sélectionnent des candidats, les employeurs les mettent en rang : celui que l'on considère le plus apte à remplir le poste reçoit le rang 1, le suivant le rang 2, etc. On procède de la même façon dans les universités lorsqu'il faut sélectionner les étudiants, particulièrement pour les programmes d'études de deuxième et troisième cycles, qui sont très contingentés. Les Nations Unies produisent un rapport annuel décrivant la qualité de vie dans différents pays. On mesure un ensemble de caractéristiques dans chaque pays, comme l'espérance de vie, le revenu moyen et le niveau de chômage, afin de produire une valeur globale indexant la qualité de vie pour chaque pays. Le pays qui obtient la valeur la plus forte obtient le rang 1, ce qui indique que ce pays offre la meilleure qualité de vie. Naturellement, comme il s'agit d'une mesure ordinale, lorsque le Canada obtient le premier rang, on ne sait pas si la qualité de vie au Canada est légèrement ou fortement supérieure aux pays qui obtiennent les rangs 2, 3 ou 20 !

Quiz rapide 1.3
Trois étudiants obtiennent les résultats suivants à leur examen de statistique : Paul = 50 %, Marie = 80 %, Julie = 80,4 %. Indiquez la performance de chacun sur une échelle ordinale.

Les variables (échelles) à intervalles

Les *variables à intervalles* (ou variables relatives) sont souvent utilisées pour mesurer des phénomènes en sciences humaines. Le psychopédagogue qui mesure le niveau d'intelligence (le QI) des élèves, le psychologue qui mesure la personnalité, l'entreprise qui mesure le degré de satisfaction de la clientèle ou le psychiatre qui étudie le stress se servent de variables à intervalles. Les variables à intervalles mesurent non seulement la position relative de chaque

observation, mais indiquent aussi l'ampleur des différences entre elles. Ainsi, les QI de Peter, Paul et Marie sont respectivement de 95, 100 et 120. Bien sûr, Marie occupe le rang 1, Paul le rang 2 et Peter le rang 3, mais parce que le QI est une mesure à intervalles, nous mesurons aussi la magnitude des différences entre ces rangs. Ainsi, nous pouvons conclure que l'écart entre le QI de Marie et celui de Paul (20) est plus grand que celui entre Peter et Paul (5). Les valeurs d'une mesure à intervalles contiennent plus d'informations que ne le font les valeurs des échelles ordinales et des échelles nominales. La grande majorité des variables psychologiques sont des variables à intervalles : un psychologue est en mesure d'indiquer non seulement si quelqu'un est moins anxieux que sa mère, mais s'il l'est beaucoup ou légèrement moins.

Les variables à intervalles souffrent néanmoins d'une limite importante : elles n'ont pas de point zéro. La valeur « 0 » existe lorsque l'absence totale de la caractéristique mesurée est possible. L'absence d'un zéro absolu pour certaines mesures apparaît lorsque cette valeur est impossible. Par exemple, avec la mesure de la personnalité (intervalle), il n'existe pas de valeur « 0 », car l'absence totale de personnalité est inconcevable. Similairement, l'absence totale d'intelligence n'existe pas (même si vous avez cru la constater chez certains politiciens !). Comme nous n'avons pas de point zéro, il n'est, par conséquent, pas possible de calculer des ratios entre deux valeurs. Ainsi, il est impossible de conclure qu'une personne ayant un QI de 120 est deux fois plus intelligente qu'une personne ayant un QI de 60 (bien qu'arithmétiquement 120 soit dans un ratio de 2 pour 1 par rapport à 60).

La mesure de la chaleur en degrés Celsius (°C) ou Fahrenheit (°F) est une mesure à intervalles. S'il fait 10 °C lundi, 15 °C mardi et 30 °C mercredi, nous pouvons conclure que la température a davantage augmenté de mardi à mercredi qu'elle ne l'a fait entre lundi et mardi. Il serait faux de conclure qu'il fait deux fois plus chaud mercredi que mardi, car une température de zéro ne signifie pas une absence totale de chaleur (sinon les températures de –10 °C ou –20 °F n'existeraient pas). L'échelle de température Kelvin, en revanche, n'est pas une mesure à intervalles, car elle inclut une valeur « 0 » qui indique une absence absolue de chaleur. Lorsqu'une variable contient un vrai point zéro, celui-ci indiquant l'absence totale de la caractéristique, l'information qu'elle contient est mesurée sur une échelle de rapport (dont nous discutons plus loin).

Les mesures psychologiques sont souvent prises avec des échelles linéaires. L'encadré au début du chapitre en donne un exemple. Lorsque nous demandons au répondant d'indiquer son degré d'accord ou de désaccord avec une phrase déclarative, nous nommons ce type d'échelle « échelle de Likert ». Il existe différentes variantes de cette échelle. Par exemple, nous pourrions poser la question suivante : « Jusqu'à quel point êtes-vous satisfait de votre cours ? » Le répondant choisit la réponse qui correspond le mieux à son opinion : 1 = totalement satisfait ; 2 = satisfait ; 3 = ni satisfait ni insatisfait ; 4 = insatisfait ; 5 = totalement insatisfait. À strictement parler, ces échelles sont des échelles ordinales. Mais les psychologues, entre autres, traitent ces réponses comme si elles étaient collectées sur des échelles à intervalles. La raison en est qu'ils présument que la caractéristique mesurée (dans ce cas, il s'agit de la satisfaction par rapport à un cours) est une mesure continue où il est possible d'avoir des degrés de satisfaction et non seulement un ordre. Il devient donc possible de dire que Monsieur X est beaucoup plus satisfait que ne l'est Madame Y ; mais parce qu'il s'agit d'une variable à intervalles, il n'est pas possible de dire que Monsieur X est deux fois plus satisfait que Madame Y.

Les variables (échelles) de rapport

Les *variables de rapport* (ou échelles absolues) ont toutes les propriétés des échelles à intervalles, mais, en plus, elles ont un point zéro absolu. La plupart des caractéristiques physiques sont des échelles de rapport : la taille et le montant d'argent en banque sont des échelles de rapport car il est concevable d'avoir une absence totale de taille ou d'argent. Les échelles de rapport nous permettent de dire que quelque chose est deux fois plus grand ou plus petit que quelque chose d'autre. Ainsi, si nous avons 1 000 $ et que notre frère en a 2 000 $, il a deux fois plus d'argent que nous. De manière similaire, si votre équipe a gagné 30 parties l'année dernière et 45 parties cette année, elle a gagné 50 % plus de parties. Enfin, si vous avez obtenu 90 % à votre examen de chimie et que votre copine a obtenu 45 %, vous avez obtenu le double de ses points. Ces conclusions sont valides, car il est possible de ne pas avoir d'argent, de n'avoir gagné aucune partie ou de n'avoir répondu correctement à aucune question à un examen.

Quiz rapide 1.4
Deux étudiants obtiennent les résultats suivants à l'examen de statistique : Paul = 40 %, Marie = 80 %. Est-ce que vous pouvez conclure que Marie a réussi son examen deux fois mieux que Paul ?

Les relations entre les diverses échelles de mesure

Les échelles de mesure fournissent de l'information au sujet des observations et les quatre types d'échelles sont organisés de manière hiérarchique. Ainsi, l'échelle nominale nous indique exclusivement la catégorie à laquelle chaque observation appartient (A appartient à la catégorie 1, B à la catégorie 2) ; l'échelle ordinale nous indique l'ordre entre les observations (A est plus grand que B) aussi bien que la catégorie (A est premier, les autres ne le sont pas) ; l'échelle à intervalles nous donne la différence relative entre les observations (la différence entre A et B est plus grande que la différence entre B et C) en plus de la catégorie et de l'ordre ; et enfin, l'échelle de rapport nous indique, en plus des trois autres niveaux d'information, la différence absolue entre les mesures (A est deux fois plus grand que B). Le Tableau 1.2 décrit les relations entre les informations fournies par les quatre types d'échelles.

Les mesures nominales et ordinales prennent parfois le nom d'échelles de type I alors que les mesures à intervalles et de rapport sont parfois appelées échelles de type II.

Tableau 1.2
Comparaison des échelles de mesure

Échelle de mesure		*Catégorie*	*Ordre*	*Différence relative*	*Différence absolue*
Type I	nominale	✓			
	ordinale	✓	✓		
Type II	à intervalles	✓	✓	✓	
	de rapport	✓	✓	✓	✓

Quiz rapide 1.5
Voici les résultats obtenus à un examen de statistique par trois étudiants : Marie = 90 %, Paul = 71 %, Julie = 70 %. Tirez les conclusions nominales, ordinales, à intervalles et de rapport pour ces trois observations.

SOMMAIRE DU CHAPITRE

Les statistiques aident à tirer des conclusions au sujet d'informations quantitatives qui sont organisées en banque de données. Une information quantitative est une information numérique, et une banque de données est un tableau à double entrée. La banque de données contient les informations que les sujets d'analyse fournissent pour une ou plusieurs variables. Les variables sont les caractéristiques qui sont mesurées et pour lesquelles plusieurs réponses sont possibles. Les réponses peuvent être nominales (elles indiquent si le sujet d'analyse détient ou ne détient pas la caractéristique mesurée), ordinales (elles indiquent le rang, la position relative, de chaque observation), à intervalles (elles indiquent la différence relative entre les observations) ou de rapport (elles indiquent la différence absolue entre les observations). Les techniques d'analyse statistique utilisables ne sont pas les mêmes pour les différents types d'échelles de mesure. La plupart des techniques statistiques, y compris celles décrites dans ce livre, exigent que les variables soient à intervalles ou de rapport. Mais il est aussi possible de faire une analyse statistique valide lorsque les données sont de type I. Dans ce cas, il faudra faire appel aux procédures « non paramétriques » qui, elles, sont décrites au chapitre 13 de ce livre.

EXERCICES DE COMPRÉHENSION

1. Une caractéristique ou un phénomène pouvant prendre différentes valeurs est
   a) une constante
   b) une donnée brute
   c) une population
   d) une variable

2. Le but de l'inférence statistique est de tirer une conclusion _____ à partir d'une information _____.
   a) plus générale ; spécifique
   b) juste ; fausse
   c) spécifique ; générale
   d) générale ; générale

3. Déterminer le type d'échelle de ces mesures.
   a) Âge : _____
   b) Ethnie : _____
   c) Résultats d'une course à pied : _____
   d) Quotient intellectuel : _____

4. Le regroupement d'individus dans des catégories telles que « faible », « moyen » et « fort » implique quel type d'échelle ?
   a) Échelle nominale
   b) Échelle ordinale
   c) Échelle à intervalles
   d) Échelle de rapport

5. Une échelle définit la catégorie à laquelle une personne appartient. Il s'agit alors d'une échelle _____.
   a) nominale
   b) ordinale
   c) à intervalles
   d) de rapport

6. Transposer une mesure d'une échelle à une autre n'est pas possible dans le cas suivant :
   a) d'une échelle nominale à une échelle à intervalles
   b) d'une échelle à intervalles à une échelle ordinale
   c) d'une échelle de rapport à une échelle nominale
   d) d'une échelle de rapport à une échelle à intervalles

7. Lorsque l'on dit : « Mario est plus beau que Simon », quel type d'échelle utilise-t-on ?
   a) Échelle nominale
   b) Échelle ordinale
   c) Échelle à intervalles
   d) Échelle de rapport

8. Laquelle de ces mesures nous donne le plus d'informations ?
   a) L'ordre des chevaux à l'arrivée
   b) Le nombre d'hommes et de femmes inscrits en pharmacologie
   c) La température en degrés Celsius
   d) La distance entre la Terre et les planètes du système solaire

9. Dans une expérience, on augmente le salaire d'un groupe d'employés d'une compagnie alors que le salaire d'un autre groupe d'employés reste inchangé. Ensuite, on examine le degré de productivité des deux groupes d'employés afin de voir si le salaire affecte la productivité. La variable indépendante est _____ et la variable dépendante est _____.
   a) ceux qui reçoivent l'augmentation ; ceux qui ne la reçoivent pas
   b) la productivité ; le salaire
   c) le salaire ; la productivité
   d) le salaire ; la satisfaction de ceux qui ne reçoivent pas d'augmentation

Réponses

1. d
2. a
3. a. échelle de rapport ; b. échelle nominale ; c. échelle ordinale ; d. échelle à intervalles
4. c
5. a
6. a
7. b
8. d
9. c

CHAPITRE 2
LA DISTRIBUTION DES DONNÉES

Page laissée blanche

CHAPITRE 2

LA DISTRIBUTION DES DONNÉES

La statistique consiste à réduire une grande quantité d'informations à une expression plus simple, afin d'en tirer des renseignements utiles. Le point de départ de ce processus de simplification consiste à simplement recenser (compter) le nombre d'observations qui appartiennent à chaque valeur d'une variable. Par exemple, pour examiner le poids d'un groupe d'enfants de dix ans, nous pourrions compter le nombre d'enfants qui pèsent 40 kg, le nombre d'enfants qui pèsent 41 kg, etc. Cette simple procédure statistique établit l'*effectif,* c'est-à-dire la fréquence à laquelle chaque valeur de la variable apparaît dans la banque de données, indiquant ainsi la *distribution* (c'est-à-dire la répartition) de ces valeurs. Par exemple, nous pourrions dire que 20 % des enfants de dix ans pèsent 35 kg, 30 % en pèsent 40, etc.

L'établissement de l'effectif des données et leur distribution représentent le point de départ crucial de toutes les analyses statistiques abordées dans ce livre. Dans le présent chapitre, nous voyons les procédures à suivre pour établir et représenter la distribution des données, numériquement et visuellement, à l'aide de graphiques. Ces procédures, puisqu'elles servent à décrire l'information, s'appellent les *statistiques descriptives*. Ces statistiques sont le sujet des cinq premier chapitres.

Le salaire des joueurs de hockey professionnels

Un désaccord entre les athlètes et les propriétaires des équipes de hockey de la Ligue nationale de hockey (LNH) a mené à l'annulation complète de la saison de hockey en 2004-2005. Les propriétaires soutenaient que les salaires des joueurs étaient trop élevés, tandis que les joueurs, ce qui ne surprit personne, ne partageaient pas ce point de vue. Qui avait raison, les propriétaires ou les joueurs ? Le point de départ pour résoudre cette question se trouve dans la simple description des salaires des joueurs : combien gagnent-ils ?

Le fichier *NHLSalaire2002-2003* (voir le site Internet du livre : www.pum.umontreal.ca/ca/fiches/978-2-7606-2113-8.html) reproduit les salaires de chacun des 679 athlètes de la LNH en 2002-2003. La banque de données contient un très grand nombre d'informations : on y trouve 679 lignes d'informations (une ligne par joueur) et 5 colonnes (une colonne par variable : le nom du joueur, son prénom, son équipe, sa position au sein de l'équipe et son salaire en 2002-2003), ce qui correspond à un total de 3 395 observations (679×5). Il faut reconnaître qu'avec autant d'informations, décrire la situation salariale au sein de la LNH pour s'en faire une idée globale devient quasi impossible (sauf peut-être pour conclure que le salaire du commun des mortels semble, par comparaison, bien bas !). Comme pour toutes les analyses statistiques, nous commençons le processus de simplification de l'information en compilant la distribution des effectifs, c'est-à-dire le nombre de joueurs de hockey qui se situe à chaque niveau de salaire.

LA DISTRIBUTION SIMPLE DES DONNÉES

Une fréquence est simplement le décompte du nombre d'observations ayant obtenu une certaine valeur. On appelle aussi cela un effectif. Par exemple, en nous basant sur la banque de données des salaires de la LNH en 2002-2003, nous notons que le plus bas salaire qu'elle a payé est de 165 000 $US (désormais, dans le présent chapitre, le signe $ représentera des $US). Puisque aucun autre joueur ne gagne ce salaire, nous notons un effectif de 1 pour le niveau de salaire de 165 000 $. Le salaire suivant est de 280 000 $ et, là encore, une seule personne dans la ligue reçoit ce salaire. Par conséquent, l'effectif pour la valeur 280 000 $ de la variable « salaire » est de 1. Nous poursuivons cette procédure pour chaque valeur (chaque salaire) dans la banque de données. Par exemple, 9 joueurs touchent 350 000 $. L'effectif pour 350 000 $ est donc 9. Nous voyons aussi que le salaire maximal est la modique somme de 11 000 000 $ que touchent deux joueurs. La fréquence du salaire de 11 000 000 $ est, par conséquent, de 2. Nous pouvons maintenant comprendre l'avantage de la distribution des effectifs. Elle organise les

informations que contient la banque de données en regroupant ensemble celles qui sont identiques et permet ainsi d'en réduire le nombre.

L'utilisation de la distribution simple des effectifs est tout à fait appropriée aux sondages sur les intentions de vote, dont on trouve les résultats dans les journaux. Présentés sous forme de tableaux, ces résultats indiquent le nombre ou (plus généralement) le pourcentage des répondants qui se disent prêts à voter pour l'un ou l'autre des partis politiques. Puisque le nombre de partis politiques est relativement restreint, l'utilisation de la distribution simple représente une technique très efficace pour saisir rapidement le degré de popularité de chacun des partis.

Quiz rapide 2.1
À partir des données disponibles sur le site Internet (www.pum.umontreal.ca/ca/fiches/978-2-7606-2113-8.html) : Quelle est la taille de l'effectif lorsque le niveau des salaires est de 2 000 000 $? Est-il facile de trouver cette information dans la liste complète des salaires ?

LA DISTRIBUTION GROUPÉE DES DONNÉES

Souvenons-nous que les statistiques descriptives servent à réduire la masse d'informations afin de pouvoir s'en faire une idée globale. La construction d'une distribution simple des effectifs n'est pas toujours la manière la plus pratique pour faire le sommaire d'une banque de données. Lorsque les valeurs différentes sont nombreuses, la description de la variable devient très détaillée, ce qui complexifie l'interprétation que l'on peut en faire. Il est préférable de simplifier et de réduire davantage la banque de données.

La LNH nous offre un bel exemple d'une telle banque de données. Pour décrire la distribution des salaires de ses joueurs, nous avions comme point de départ une matrice contenant 679 rangées d'observations, ce qui est excessif pour se faire une idée globale de la distribution des salaires. En calculant les effectifs, nous avons réduit cette masse de données à environ une centaine de lignes, car il y a une centaine de salaires différents dans la LNH. C'est déjà mieux, mais c'est encore trop. En général, pour avoir une idée globale d'un ensemble de données, celui-ci ne doit pas contenir plus d'une vingtaine de catégories. Il nous faut donc réduire le nombre de catégories dans la variable. La *distribution groupée des données* est alors la procédure à suivre.

La distribution groupée des données consiste à regrouper dans la même catégorie les observations qui sont proches les unes des autres. Nous établissons maintenant les effectifs des observations qui se trouvent dans chacune de ces catégories de valeurs de la variable. Par exemple, pour obtenir le poids des enfants de dix ans, nous pourrions compter le nombre d'enfants qui pèsent entre 26 et 30 kg, entre 31 et 35 kg, etc. Dans une distribution groupée, chaque catégorie englobe plusieurs valeurs (similaires) de la variable. Pour ce qui est de la LNH, nous pourrions placer dans la même catégorie les joueurs ayant des salaires similaires et établir la fréquence de chaque catégorie. Par exemple, nous pourrions hypothétiquement établir les catégories suivantes : catégorie 1 = salaires se situant entre 0 et 499 999 $, catégorie 2 = entre 500 000 et 999 999 $, etc. Si 20 joueurs gagnent entre 0 et 499 999 $ et si 50 touchent entre 500 000 et 999 999 $, les effectifs groupés seraient respectivement de 20 et 50 pour ces deux catégories.

La distribution groupée des fréquences contiendra moins de catégories que la distribution simple et, ce faisant, il sera plus facile d'en faire une interprétation. On doit cependant noter qu'en utilisant la distribution groupée, nous sacrifions de l'information : chaque catégorie contient maintenant plusieurs niveaux de salaires. Ainsi, pour ce qui est des salaires situés entre 0 et 499 999 $, le joueur de hockey qui gagne 499 999 $ se retrouve dans la même catégorie que celui qui gagne 100 000 $, mais dans une catégorie de salaires différente de celui qui gagne 500 000 $. La simplification de la banque de données que permet une distribution groupée augmente certes la clarté de l'information, mais elle le fait en sacrifiant des détails.

Comment créer une distribution groupée des données

Pour construire une distribution groupée des données, nous devons établir une série de catégories, chacune étant définie par un intervalle de valeurs. Un intervalle spécifie la valeur maximale et la valeur minimale des observations qui seront incluses dans la catégorie. La limite supérieure définit la valeur la plus grande de l'intervalle et la limite inférieure, la valeur la plus petite. Par exemple, pour un intervalle regroupant tous les salaires entre 500 000 et 999 999 $, les limites inférieure et supérieure sont de 500 000 et 999 999 $ respectivement. On considère que tous les athlètes qui gagnent

entre 500 000 et 999 999 $ appartiennent à la catégorie de salaire 500 000-999 999 $. L'athlète qui gagne 499 999 $ appartient à une autre catégorie, soit entre 0 et 499 999 $. La définition de ces limites représente l'étape importante de la construction des distributions groupées des effectifs.

La façon de créer une distribution groupée des effectifs est très simple :

1. On décide d'abord du nombre de catégories que l'on veut. Généralement, entre 10 et 20 catégories. Mais cette règle n'est pas coulée dans le béton. Pour certaines applications, il est approprié d'en créer plus de 20 ou moins de 10.
2. Ensuite, on calcule la différence entre la plus petite et la plus grande valeur dans la distribution (cette différence, l'*étendue de la distribution*, est une statistique de base qui est décrite au chapitre 3).
3. Enfin, on divise cette différence par le nombre de catégories. Le résultat obtenu indique la taille de chaque intervalle.

Prenons les salaires des joueurs de la LNH et établissons une distribution groupée des effectifs pour 10 intervalles.

1. La différence entre le salaire le plus élevé et le plus bas est de 10 835 000 $ (11 000 000-165 000 $).
2. Puisque nous désirons établir les effectifs pour 10 catégories de salaires, nous divisons l'étendue des salaires (10 835 000 $) par 10, et ainsi chaque intervalle regroupera les salaires en tranches de 1 083 500 $.
3. Nous pouvons maintenant construire nos intervalles et établir la distribution groupée des données : la première catégorie compte le nombre de joueurs ayant un salaire situé entre 165 000 et 1 248 500 $ (165 000 $ + 1 083 500 $ = 1 248 500 $) et la deuxième inclut tous les salaires entre 1 248 501 et 2 332 000 $. Le dernier intervalle comprend tous les salaires entre 9 916 501 et 11 000 000 $.

Dans l'exemple des salaires des joueurs de la LNH, la taille de l'intervalle créé par cette façon de faire produit un chiffre peu usuel (1 083 500 $). Or, il est généralement préférable d'arrondir la taille des intervalles. Ainsi, au lieu d'utiliser un intervalle de 1 083 500 $, il est plus commode de choisir un intervalle de 1 100 000 $. Donc, le premier intervalle comprend les salaires se situant entre 0 et 1 100 000 $ inclusivement, le deuxième intervalle, les salaires supérieurs à 1 100 000 $ et égaux ou inférieurs à 2 200 000 $, le troisième intervalle, les salaires supérieurs à 2 200 000 $ et égaux ou inférieurs

à 3 300 000 $, etc. Le Tableau 2.1 montre les effectifs groupés pour les salaires des joueurs de la LNH. On peut y remarquer deux aspects importants :

- Chaque salaire appartient à une seule catégorie.
- Tous les salaires sont catégorisés.

Tableau 2.1
Distribution des données pour les salaires des joueurs de la LNH, 2002-2003, avec intervalle de 1 100 000 $

Catégorie de salaires (intervalle) en M $	*Fréquence*	*Pourcentage (proportion)*	*Pourcentage cumulatif*
Plus de 0 à 1,1	374	55,1 % (0,551)	55,1 %
Plus de 1,1 à 2,2	148	21,8 % (0,218)	76,9 %
Plus de 2,2 à 3,3	76	11,2 % (0,112)	88,1 %
Plus de 3,3 à 4,4	30	4,4 % (0,044)	92,5 %
Plus de 4,4 à 5,5	20	2,9 % (0,029)	95,4 %
Plus de 5,5 à 6,6	9	1,3 % (0,013)	96,8 %
Plus de 6,6 à 7,7	5	0,7 % (0,007)	97,5 %
Plus de 7,7 à 8,8	5	0,7 % (0,007)	98,2 %
Plus de 8,8 à 9,9	7	1,0 % (0,01)	99,3 %
Plus de 9,9 à 11	5	0,7 % (0,007)	100,0 %
TOTAL	**679**	**100,0 % (1,0)**	

Le Tableau 2.1 permet maintenant d'appréhender rapidement la *distribution des salaires* que ces athlètes reçoivent. Par exemple, la majorité des joueurs (374 sur 679, ou 55,1 %) a un salaire égal ou inférieur à 1 100 000 $ et seulement une minorité (5 sur 679, ou 0,7 %) touche plus de 9 900 000 $.

Quiz rapide 2.2
Un nouveau joueur arrive dans l'équipe. Il gagne 12 000 000 $. Est-ce qu'on doit refaire tout le Tableau 2.1 ou ajouter une nouvelle catégorie « Plus de 11 000 000 à 12 100 000 $ » ? Justifiez votre réponse.

La taille des catégories et leur nombre

Il est plus facile de faire une interprétation des distributions de données lorsqu'elles contiennent peu de catégories. Mais, moins il y a de catégories, plus grands sont les intervalles, et moins précise est l'interprétation qui pourra être faite de la distribution.

Le principe peut être illustré pour les salaires des hockeyeurs de la LNH. Si nous créions un seul intervalle (le nombre minimal d'intervalles possible), tous les salaires y seraient inclus et nous pourrions conclure que 100 % des salaires des joueurs se situent entre 0 et 11 000 000 $! Ce résultat ne nous aiderait pas beaucoup ! À l'inverse, nous pourrions représenter une catégorie par salaire (soit le nombre maximal de catégories possibles), ce qui produirait un tableau contenant 679 catégories, et cela ne nous avancerait pas plus. En général, nous nous efforçons de créer une distribution groupée des fréquences qui contient aussi peu de catégories que possible, tout en restant utile. Dans la plupart des cas, nous essayons d'établir entre 10 à 20 catégories bien que, dans certains cas, nous puissions en créer plus ou moins.

La distribution groupée des données : sommaire des étapes

La construction d'une distribution groupée des données exige le respect de trois règles fondamentales.

1. Les intervalles définissant les catégories doivent être établis de manière à ce que chaque observation soit classée dans une seule catégorie.
2. Les catégories doivent être de taille identique. Elles respectent toutes la même étendue de valeurs de la variable.
3. Les catégories doivent être choisies de manière à couvrir toutes les valeurs possibles.

LA DISTRIBUTION RELATIVE DES DONNÉES

Le Tableau 2.1 est utile pour faire une représentation des salaires des joueurs de hockey. Ainsi, on peut noter que 374 joueurs sont payés 1 100 000 $ ou moins, tandis que seulement 5 gagnent 9 900 000 $ ou plus. Il va sans dire

qu'un salaire aux alentours de 1 000 000 $ est plus habituel dans la LNH qu'un salaire de 10 000 000 $.

Pour mieux comprendre ces effectifs, il est souvent pratique d'exprimer, pour chaque valeur ou catégorie de valeurs, la fréquence des observations qui s'y trouvent relativement au nombre total d'observations. Cette distribution prend un nom différent. On l'appelle *distribution relative des effectifs,* car la fréquence des observations pour chaque valeur exprime le nombre d'observations dans chaque valeur *relative* (par rapport) au nombre total d'observations. Nous pouvons exprimer ce rapport en *proportion* ou en *pourcentage*.

La proportion indique la fréquence des observations se trouvant dans chaque intervalle relatif au nombre total d'observations. Le calcul de la proportion est facile : il s'agit simplement de diviser la fréquence obtenue pour chaque intervalle (f_i) par le nombre total d'observations (N) :

$$\text{Proportion} = f_i/N \qquad \text{Formule 2.1}$$

La proportion est une valeur qui varie entre 0 et 1,0. Ainsi, dans une distribution qui contient 100 observations, si 50 d'entre elles se trouvent dans le même intervalle, nous disons que la proportion des observations qui se situent dans cet intervalle est de 0,5 (Proportion = $f_i/N = 50/100 = 0{,}5$). Si aucune observation n'existe pour un intervalle en particulier, la proportion pour cet intervalle est de 0,0.

Il est également facile, une fois que nous avons calculé la proportion des observations, de les exprimer en pourcentage. Les pourcentages varient entre 0 et 100. Lorsque nous multiplions la proportion par 100 %, nous obtenons le pourcentage :

$$\text{Pourcentage} = (f_i/N) \times 100\,\% \qquad \text{Formule 2.2}$$

Ainsi, lorsque nous obtenons une proportion de 0,50, cela indique que 50 % (donc la moitié) de toutes les observations tombent dans cet intervalle.

Le Tableau 2.1 présente (à la troisième colonne) le pourcentage (et la proportion qui est entre parenthèses) de joueurs dont le salaire se trouve dans chacune des catégories (intervalles). Calculons la proportion et le pourcentage de joueurs de la LNH dont le salaire se situe dans le premier intervalle (entre 0 et 1 100 000 $). Nous constatons qu'il y a 374 joueurs

dans ce premier intervalle de la distribution groupée des fréquences. Nous savons qu'au total, la banque de données inclut le salaire de 679 athlètes. Ainsi $f_1 = 374$ et $N = 679$. La proportion est donc représentée par $(f_1/N) = (374/679) = 0{,}5508$ ou 0,551, en arrondissant. Pour trouver le pourcentage, nous multiplions la proportion par 100 % = 0,5508 (100 % = 55,08 % ou 55,1 %, en arrondissant). Donc, nous constatons que 55,1 % (c'est-à-dire la majorité) des joueurs de hockey gagnent entre 0 et 1 100 000 $. Si nous reprenons la même démarche pour les athlètes les mieux payés (la dixième et dernière catégorie), nous notons que moins de 1 % (0,7 %) des joueurs de la LNH sont payés plus de 9 900 000 $ ($N = 679$ et $f_{10} = 5$; Proportion = $f_{10}/N = 5/679 = 0{,}0074$) ; ce qui équivaut au pourcentage $0{,}0074 \times 100\,\% = 0{,}74\,\%$, ou 0,7 %, en arrondissant.

Une première conclusion s'impose au sujet du différend entre les propriétaires et les athlètes de la LNH. Bien qu'il soit vrai que les salaires des joueurs peuvent grandement varier (la différence entre le salaire du joueur le mieux payé et celui du joueur le moins bien payé est de plus de 10 000 000 $) et que certains gagnent jusqu'à 11 000 000 $, il reste que la majorité des joueurs (55,1 %) gagne 1 000 000 $ ou moins par année. Est-ce que les joueurs de hockey gagnent des salaires exorbitants ? La distribution groupée des fréquences nous offre une réponse préliminaire : il est clair que certains athlètes sont très bien payés, mais la majorité obtient des salaires qui semblent plutôt ordinaires pour des athlètes professionnels.

La distribution cumulative : proportions et pourcentages

Il est souvent fort utile d'exprimer une distribution de fréquence relative en la transformant en distribution de proportion (ou de pourcentage) *cumulative*. L'idée consiste ici à établir *la proportion ou le pourcentage des observations qui se situent à chaque intervalle* PLUS *celles qui se trouvent dans tous les intervalles inférieurs.* On peut étudier, à titre illustratif, les deux premières lignes de la quatrième colonne du Tableau 2.1. On note (à la première rangée de la colonne 4) que la proportion des joueurs qui gagne 1 100 000 $ ou moins est de 0,551 (55,1 %), et, à la deuxième rangée, on voit qu'une proportion de 0,769 (76,9 %) des joueurs gagne moins de 2 200 000 $. Cette quantité (76,9 %)

est la somme de la fréquence de la deuxième rangée (21,8 %) plus celle de la première rangée (55,1 %).

La distribution de fréquence cumulative est informative, car même s'il est vrai que les salaires des joueurs de hockey peuvent aller jusqu'à 11 000 000 $, nous voyons maintenant que plus des trois quarts des joueurs (76,9 %) touchent une fraction de ce montant, en l'occurrence 2 200 000 $ ou moins. Si l'on pense que 2 200 000 $ n'est pas un salaire exorbitant pour un athlète professionnel, on va conclure que les trois quarts des joueurs de hockey n'ont pas un salaire exorbitant ! Si, en revanche, vous pensez que 2 200 000 $ est un salaire déraisonnable, la conclusion ne sera pas la même.

On peut construire une distribution cumulative des proportions, des pourcentages ou des fréquences en additionnant la proportion, le pourcentage ou la fréquence des observations qui se situent dans un intervalle particulier à la proportion, au pourcentage ou à la fréquence se trouvant dans tous les intervalles inférieurs. Par exemple, le pourcentage cumulatif pour l'intervalle 4 400 000 à 5 500 000 $ est 95,4 % (55,1 % + 21,8 % + 11,2 % + 4,4 % + 2,9 %). La fréquence cumulative pour ce même intervalle serait 648 (374 + 148 + 76 + 30 + 20). Nous concluons que, des 679 athlètes, 648 gagnent 5 500 000 $ ou moins.

Quiz rapide 2.3
Dans la banque de données du site Internet (www.pum.umontreal.ca/ca/fiches/978-2-7606-2113-8.html) : Quel est le pourcentage de joueurs de l'équipe de Tampa Bay qui gagnent 3 000 000 $ ou moins ?

LES REPRÉSENTATIONS GRAPHIQUES DE LA DISTRIBUTION DES DONNÉES

Il est souvent fort pratique de décrire graphiquement la distribution des données. Bien que plusieurs types de graphiques puissent être créés pour refléter la distribution, les graphiques des histogrammes et les polygones de fréquences sont ceux que nous rencontrons le plus fréquemment.

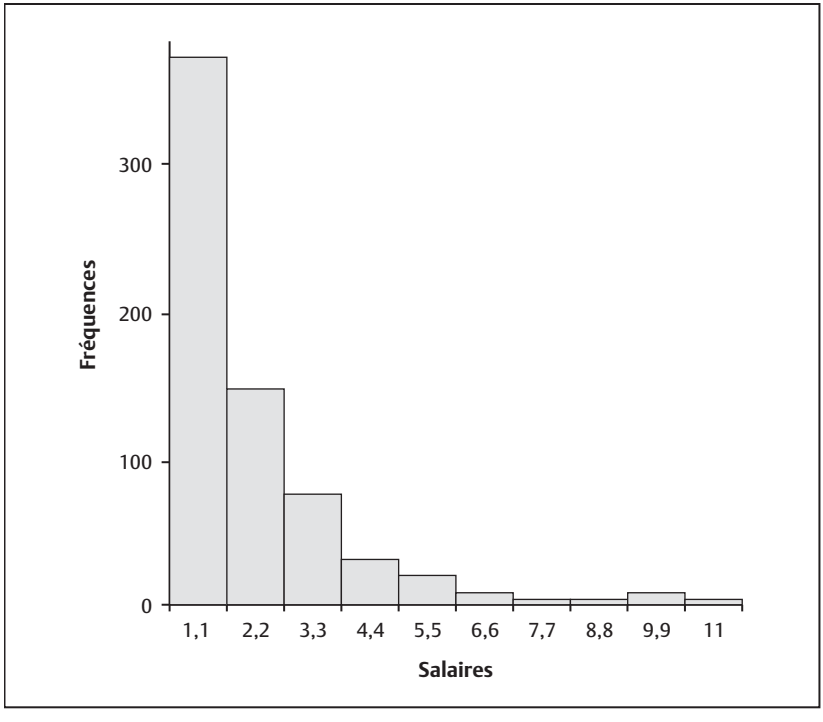
Le graphique des histogrammes

La distribution des données du Tableau 2.1 est représentée visuellement à la Figure 2.1 par un *graphique des histogrammes.* Un histogramme est une barre verticale qui représente la taille d'un effectif. Lorsque chacun des effectifs d'une distribution est identifié par un histogramme, on obtient un diagramme des histogrammes. Plus la fréquence d'une valeur ou d'une catégorie est grande, plus long est l'histogramme.

Le graphique des histogrammes contient deux axes : l'axe horizontal se nomme l'*abscisse* (généralement identifiée par le code X) tandis que l'axe vertical prend le nom d'*ordonnée* (généralement identifiée par le code Y). L'abscisse indique les catégories de réponses (les intervalles ou les catégories) et l'ordonnée indique les fréquences. La fréquence obtenue pour chaque catégorie de réponses est signalée par une barre verticale. Plus la barre est longue, plus les effectifs associés à cette catégorie sont grands. La plus petite fréquence possible étant zéro, le point où l'ordonnée coupe l'abscisse se trouve, dans ce cas, à la fréquence 0.

Pour dessiner le graphique, on commence par la première catégorie (l'intervalle 0 à 1 100 000 $) et on note sa fréquence dans le tableau de la distribution de fréquences (l'effectif de ce premier intervalle est de 374, car 374 joueurs touchent entre 0 et 1 100 000 $). Ensuite, on trouve le point, le long de l'ordonnée, qui correspond à une fréquence de 374 et l'on trace une barre qui va de l'abscisse jusqu'à ce point sur l'ordonnée. On passe alors au deuxième intervalle de salaire (1 100 000-2 200 000 $). L'effectif pour cette deuxième catégorie est de 148. On trace alors une deuxième barre qui part sur l'abscisse et qui se prolonge jusqu'à la fréquence de 148, le long de l'ordonnée. On procède ainsi pour chaque intervalle jusqu'au dernier (9 900 000-11 000 000 $, qui a un effectif de 5).

FIGURE 2.1 L'histogramme des salaires (en millions de dollars)



Quiz rapide 2.4
L'histogramme de la Figure 2.1 qui représente les salaires des hockeyeurs a une forme très particulière (le gros des salaires est dans le bas de l'échelle). Pensez-vous que l'on puisse retrouver cette même forme en ce qui concerne les salaires des joueurs de basket-ball de la National Basketball League ? ou le nombre de poissons pêchés dans une journée par des bateaux de pêche ?

En examinant la Figure 2.1, la situation des salaires des joueurs de la LNH se clarifie rapidement : la plupart d'entre eux ne touchent pas plus de 1 100 000 $ et seule une infime minorité de ces athlètes touchent plus de 5 500 000 $; dans la LNH, 10 000 000 $ ou plus est un salaire fort inhabituel. En fait, l'histogramme ne contient pas plus d'informations que la distribution de fréquences qu'il décrit, mais il les présente sous un format plus facile et rapide à saisir.

Les règles utiles pour construire des histogrammes

La construction des histogrammes exige le respect d'un certain nombre de règles.

1. Les intervalles reflétant les valeurs plus faibles de la variable se placent vers la partie gauche de l'abscisse, et les valeurs plus fortes, vers la droite. Ainsi, à la Figure 2.1, l'intervalle décrivant le salaire le plus faible (0-1 100 000 $) est à l'extrême gauche de l'abscisse, et l'intervalle du salaire le plus fort (9 900 000-11 000 000 $) est à l'extrême droite de l'abscisse.
2. Les fréquences identifiées sur l'ordonnée sont ascendantes, c'est-à-dire que la fréquence minimale (souvent zéro) est située au point où l'ordonnée et l'abscisse se coupent (s'interceptent).
3. L'étiquette qui définit chaque intervalle est inscrite sous chaque histogramme. Lorsque ces étiquettes sont trop longues, on peut alors les identifier dans une légende adjacente au graphique. Mais lorsqu'on fait ce choix, il est important d'identifier chaque barre de l'histogramme par une couleur ou une texture différente afin de pouvoir les distinguer rapidement.

Le polygone des effectifs

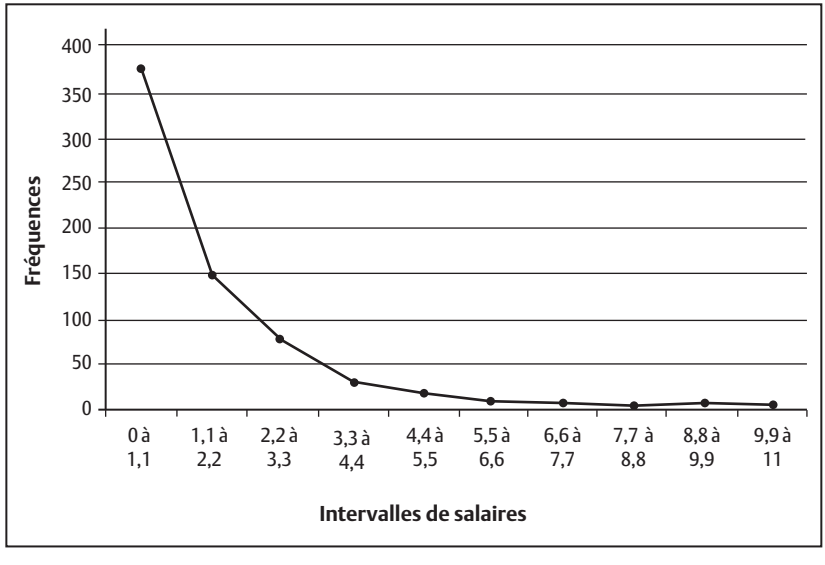
Lorsqu'on travaille avec des variables à intervalles ou de rapport, on peut aussi remplacer l'histogramme par une ligne liant les fréquences; on appelle le résultat un *graphique des polygones*, comme celui de la Figure 2.2. Les polygones des effectifs sont souvent plus lisibles que les histogrammes et, comme nous le verrons plus loin, ils sont pratiques lorsque utilisés pour décrire des distributions de fréquences relatives.

La construction d'un polygone des fréquences est très simple. Lorsqu'on travaille avec des distributions simples, il s'agit de mettre un point sur le graphique se rapportant à la fréquence de chaque valeur de la variable, et de relier ensuite chacun de ces points par une ligne. Lorsqu'on travaille avec des distributions groupées, on met le point à la valeur qui définit le centre de l'intervalle. Pour le polygone des salaires des hockeyeurs de la LNH, le

point qui décrit la première catégorie (0-1 100 000 $) est situé visuellement au centre de l'intervalle (550 000 $).

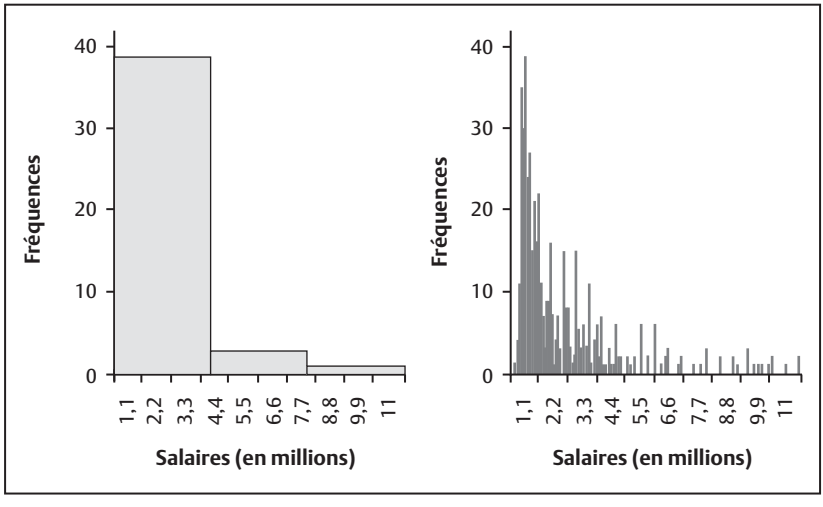
Le polygone des fréquences utilise la même information que l'histogramme, et ces formes graphiques proviennent toutes deux de la distribution. L'avantage du polygone sur l'histogramme est qu'il produit un graphique visuellement plus simple. Si on étudie la Figure 2.2, on voit très bien que la fréquence des salaires plus élevés chute de façon marquante.

FIGURE 2.2 Le polygone des salaires (en millions de dollars)



Comme pour la distribution des fréquences, il importe de construire les graphiques en se préoccupant du nombre total de catégories. L'objectif est d'accentuer la lisibilité du graphique en limitant le nombre de catégories sans pour autant le réduire trop. Par exemple, à la Figure 2.3, on a deux distributions groupées pour le salaire des joueurs de la LNH. Le graphique de gauche comprend seulement trois intervalles, alors que celui de droite en contient beaucoup plus. Lequel de ces deux graphiques représente le mieux les résultats ? La Figure 2.1 est un graphique plus utile que les histogrammes, mais aucune des deux n'est entièrement satisfaisante.

FIGURE 2.3 Exemples d'histogrammes où le nombre de catégories est inapproprié



Quiz rapide 2.5
Avec les graphiques de la Figure 2.3, peut-on avoir une idée (même approximative) du nombre de joueurs gagnant environ 500 000 $? Avec la Figure 2.1, est-ce plus facile ?

LES FORMES DE DISTRIBUTION

La distribution des fréquences et les graphiques qui la représentent nous permettent de connaître la forme que prend la distribution. Cette forme générale est un important élément descriptif des distributions. La Figure 2.4 nous montre six formes possibles.

La distribution unimodale

La *distribution unimodale* a une seule « bosse » indiquant que l'effectif pour une des valeurs (ou un seul intervalle de valeurs dans le cas des distributions groupées des effectifs) est plus grande que l'effectif de n'importe quelle autre valeur (ou intervalle de valeurs). La valeur sur l'abscisse qui est associée à cette bosse s'appelle le *mode*. *Lorsqu'une distribution contient*

une seule valeur, qui est la plus fréquente, la distribution est unimodale. Aussi, les effectifs pour les valeurs (ou intervalles) qui s'éloignent du mode deviennent graduellement plus petits. La distribution normale (la fameuse courbe en cloche que nous reverrons au chapitre 5) est une distribution unimodale.

La distribution bimodale (ou multimodale)

Contrairement à la distribution unimodale, la *distribution bimodale* contient deux modes. Dans ce cas, nous avons deux valeurs de la distribution qui sont à la fois fréquentes et les plus fortes de la distribution. Les distributions bimodales sont plus rares que les distributions unimodales. Une distribution bimodale indique généralement que nous avons deux sous-groupes d'observations distinctes dans la distribution. Par exemple, un histogramme décrivant la taille des joueurs de basket-ball et des jockeys sera presque certainement bimodale. Même si certains joueurs de basket-ball sont plus petits que d'autres, il y a fort à parier que tous seront plus grands que les jockeys. L'histogramme de cette distribution hypothétique aurait deux modes, l'un décrivant les jockeys, l'autre les joueurs de basket-ball. Lorsque nous avons plus de deux modes dans une distribution de fréquences, la distribution prend le nom de *distribution multimodale.*

Quiz rapide 2.6
Tenez pour acquis qu'il existe une distribution des connaissances en mathématiques. Vous testez les connaissances mathématiques de deux groupes d'étudiants, l'un provient du secondaire, et l'autre de l'université. Supposons que vous placiez les connaissances en mathématiques des deux groupes sur le même polygone, quelle sera la forme probable de cette distribution : unimodale ou bimodale ?

La distribution symétrique

Lorsque, dans une distribution, la fréquence des valeurs se répartit également des deux côtés de la valeur modale, nous disons que la distribution est *symétrique.* Lorsque la fréquence des valeurs ne se répartit pas également des deux côtés du mode, nous disons que la distribution est *asymétrique.*

La distribution asymétrique

La *distribution asymétrique* se distingue d'une distribution symétrique par la position du mode. Les distributions unimodales qui sont asymétriques ont des fréquences très différentes selon le côté où sont situés les effectifs plus grands. Une distribution *asymétrique positive* indique que les valeurs sont plus étirées du côté positif de l'abscisse. Une distribution *asymétrique négative* a des valeurs plus étirées du côté négatif de l'abscisse[1]. La distribution des salaires des hockeyeurs de la LNH est extrêmement asymétrique, et l'asymétrie est positive. La valeur la plus fréquente (le mode) se trouve à l'extrémité gauche de la distribution, et les valeurs s'étirent plus du côté positif de l'échelle. Quand la variable étudiée est le salaire, il est très fréquent d'obtenir une distribution asymétrique positive puisque la plupart des gens ont des salaires plus faibles que forts.

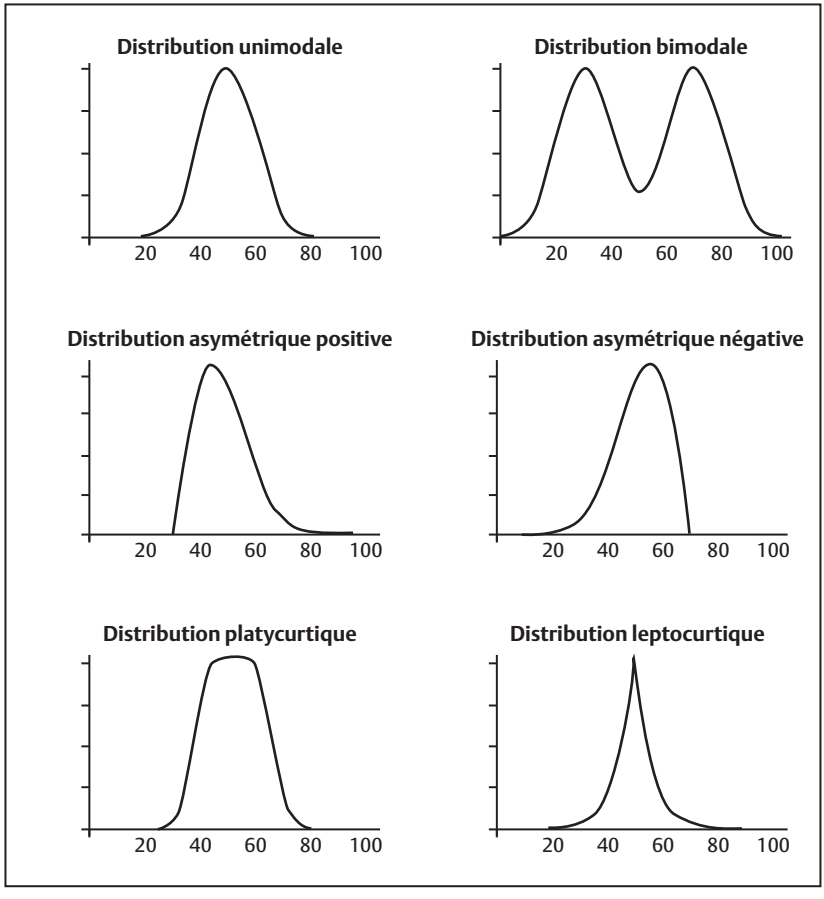
Une technique rapide pour distinguer une asymétrie positive d'une asymétrie négative consiste à examiner la forme du polygone (voir la Figure 2.4). Nous identifions la forme de l'asymétrie par la direction dans laquelle le polygone « pointe ». Lorsque la distribution pointe vers les valeurs faibles de la distribution (vers la gauche du graphique), nous disons que la distribution est asymétrique négative. Dans le cas contraire, la distribution est asymétrique positive.

Le degré d'aplatissement : leptocurtique et platycurtique

Le degré d'aplatissement d'une distribution indique avec quel degré la distribution de fréquences est aplatie ou pointue. Par exemple, les deux dernières distributions de la Figure 2.4 représentent une distribution qui est très plate (*distribution platycurtique*) et une qui est très pointue (*distribution leptocurtique*).

1. On se souviendra que, par convention, on met les valeurs faibles de la variable dans la partie gauche de l'abscisse (la partie « négative » de l'abscisse) et les valeurs fortes (positives), à droite.

FIGURE 2.4 Différentes formes de distribution de fréquences



Dans une distribution platycurtique, les valeurs de la distribution sont très étalées. La taille des effectifs est répartie plus également à travers les différentes valeurs de la variable, indiquant que les catégories contiennent des fréquences plus similaires. À l'inverse, pour la distribution leptocurtique, les valeurs sont très concentrées autour du mode : ainsi, il existe beaucoup d'observations proches du mode, et la fréquence des observations diminue rapidement au fur et à mesure que l'on s'éloigne de la valeur modale. La distribution des salaires des joueurs de la LNH est leptocurtique (aussi bien qu'asymétrique). Trois joueurs sur quatre (77 %) reçoivent des salaires

égaux ou inférieurs à 2 200 000 $, tandis que les autres (23 %) ont des salaires se situant entre 2 200 000 et 11 000 000 $.

Quiz rapide 2.7
Vous avez une distribution des absences au travail des employés d'une compagnie. La grande majorité d'entre eux s'absente entre 0 et 4 jours par année. Mais une minorité s'absente plus souvent, certains jusqu'à 50 jours. Quelle sera la forme probable de cette distribution : symétrique, asymétrique positive, asymétrique négative ?

Tableau 2.2
Notes obtenues à deux examens (en %)

Examen partiel	*Examen final*	*Examen partiel*	*Examen final*
30	33	65	71
32	42	67	73
35	44	70	74
46	52	71	75
49	55	71	76
49	57	72	77
50	61	74	77
52	62	75	78
55	62	75	79
56	64	75	81
59	65	76	82
61	66	76	82
62	66	77	84
63	67	78	86
64	69	87	88
65	71	90	92

LA DISTRIBUTION DES FRÉQUENCES : UN EXEMPLE COMPLET

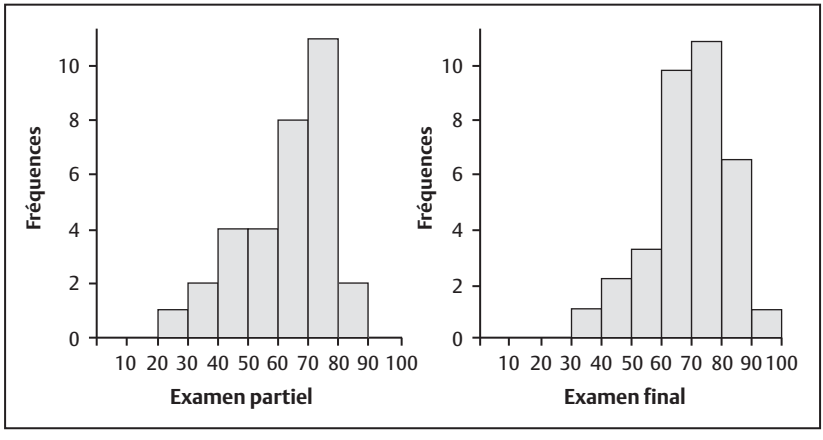
Le Tableau 2.2 présente les notes obtenues par 32 étudiants à des examens. Les notes s'étalent entre 30 et 90 pour l'examen partiel, et entre 33 et 92 pour l'examen final.

Le Tableau 2.3 reprend les données du Tableau 2.2, qu'il présente sous la forme d'une distribution groupée. La Figure 2.5 est l'histogramme groupé pour ces résultats. Nous pouvons voir que, bien que les notes de l'examen partiel s'étalent de la catégorie 20 à 30 jusqu'à la catégorie 80 à 90, la majorité des étudiants obtient des notes se situant entre 60 et 80. Quant à l'examen final, la répartition semble située un peu plus à droite (de la catégorie 30 à 40 jusqu'à la catégorie 90 à 100). Elle est aussi plus dispersée puisque la majorité des étudiants ont des notes entre 60 et 90.

Tableau 2.3
Distribution des fréquences et des pourcentages groupés et cumulatifs pour deux examens

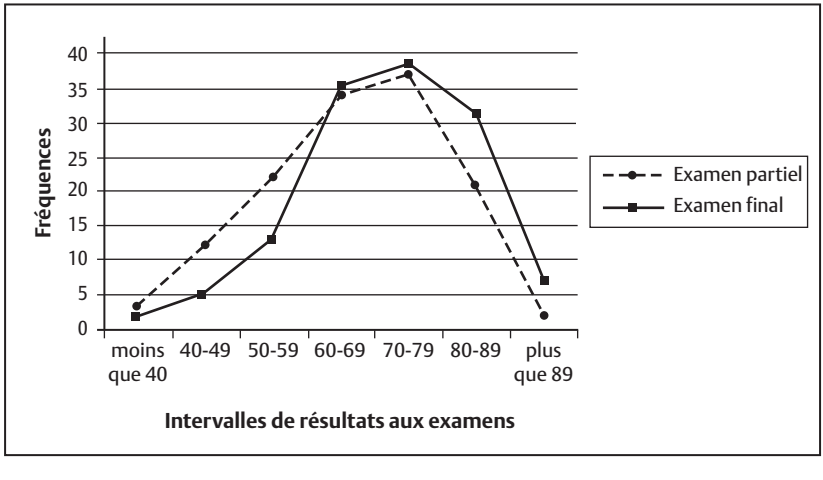
Intervalles de notes	*Fréquence examen partiel*	*Fréquence examen final*	*Pourcentage examen partiel*	*Pourcentage examen final*	*Pourcentage cumulatif examen partiel*	*Pourcentage cumulatif examen final*
plus de 20 à 30	1	0	3,1 %	0,0 %	3,1	0,0
plus de 30 à 40	2	1	6,3 %	3,1 %	9,4	3,1
plus de 40 à 50	4	2	12,5 %	6,3 %	21,9	9,4
plus de 50 à 60	4	3	12,5 %	9,4 %	34,4	18,8
plus de 60 à 70	8	9	25,0 %	28,1 %	59,4	46,9
plus de 70 à 80	11	10	34,4 %	31,3 %	93,8	78,1
plus de 80 à 90	2	6	6,3 %	18,8 %	100,0	96,9
plus de 90 à 100	0	1	0,0 %	3,1 %	100,0	100,0

FIGURE 2.5 Distribution des notes pour les examens partiel et final



Le polygone est particulièrement utile lorsqu'il s'agit de placer plusieurs distributions des effectifs sur le même graphique afin de les comparer. La Figure 2.6 est un exemple des polygones tracés à partir de la distribution des notes obtenues par les étudiants aux deux examens. Elle présente simultanément (sur le même graphique) deux polygones de fréquences, l'un qui décrit le résultat à l'examen final (en pointillé) et l'autre, les notes obtenues à l'examen partiel (la ligne solide). On trouvera les données pour ces polygones au Tableau 2.3.

FIGURE 2.6 Polygones pour les notes à deux examens



La comparaison visuelle de ces deux polygones indique plus de similarités que de différences entre les résultats aux deux examens. La majorité des étudiants obtient entre 50 et 90, et les notes très fortes et très faibles sont rares dans les deux cas. De plus, dans les deux cas, la fréquence la plus forte se retrouve pour la même catégorie de résultats aux examens (70-80). Par contre, les deux examens ne produisent pas des résultats identiques. On observe que plus d'étudiants obtiennent des notes très fortes (80-90 ; 90 et plus) à l'examen final qu'à l'examen partiel et, à l'inverse, plus d'étudiants obtiennent des résultats faibles à l'examen partiel qu'à l'examen final (40-50 ; 50-60). L'examen final était-il plus facile que l'examen partiel ? Les distributions permettent de tirer rapidement une conclusion préliminaire. En revanche, si l'on est tenté d'en tirer une conclusion définitive, il faut attendre. Il faut d'abord apprendre les procédures d'inférences statistiques qui sont discutées dans les chapitres suivants.

SOMMAIRE DU CHAPITRE

La description statistique la plus simple consiste à élaborer une distribution des effectifs. La distribution simple des effectifs présente le nombre d'observations pour chacune des valeurs d'une variable. Lorsque l'on a une grande étendue de valeurs possibles dans une distribution, il est préférable de construire une distribution groupée des effectifs. Dans ce cas, les réponses similaires sont mises dans une même catégorie et l'on compte le nombre d'observations qui tombent dans chacune d'elles. À partir de la distribution simple ou groupée des effectifs, il est possible de calculer la proportion et le pourcentage. Les distributions relatives expriment, pour chaque valeur de la distribution, le nombre d'observations qu'elles contiennent par rapport au nombre total d'observations. L'histogramme et le polygone se servent de la même information (la distribution des effectifs et ses dérivés, tels que les distributions relatives) afin de produire une « image » visuelle de la répartition. Ainsi, la distribution aussi bien que ses représentations graphiques permettent d'arriver aux mêmes conclusions.

EXERCICES DE COMPRÉHENSION

1. Lorsque nous organisons un ensemble de données par ordre croissant et que nous indiquons à côté de ces données la fréquence qui y est associée, nous construisons _____.
   a) une distribution des effectifs
   b) un histogramme
   c) un graphique des effectifs
   d) aucune de ces réponses
2. Généralement, en combien de classes les données doivent-elles être regroupées ?
   a) de 5 à 10
   b) de 10 à 20
   c) de 5 à 15
   d) de 15 à 30
3. Vous avez une distribution dont la valeur la plus petite est 22 et la valeur la plus élevée est 86. Supposons que vous décidiez de regrouper ces données en 8 classes. Quelle sera l'étendue de chaque intervalle de classe ?
   a) 8
   b) 9
   c) 10
   d) 19
4. Parmi les 5 000 professeurs d'université au Canada, 1 000 sont des professeurs adjoints. Quelle est la proportion de professeurs adjoints dans les universités canadiennes ?
   a) 1 000
   b) 20 %
   c) 0,20
   d) Impossible à calculer, puisque l'on ne connaît pas le nombre de professeurs dans les autres catégories.
5. Pour la question 4, le pourcentage de professeurs adjoints est de _____.

6. Le premier intervalle compte toutes les valeurs entre 10 et 20, et le deuxième compte toutes celles entre 20 et 30. Une personne obtient 20. Dans quelle catégorie allez-vous la placer ?
   a) La catégorie 1.
   b) La catégorie 2.
   c) À la fois dans la catégorie 1 et dans la catégorie 2.
   d) Dans ni l'une ni l'autre : les intervalles ne sont pas correctement définis.

7. Une valeur est plus fréquente que n'importe quelle autre dans une certaine distribution.
   a) La distribution est obligatoirement multimodale.
   b) La distribution est obligatoirement unimodale.
   c) La distribution est obligatoirement leptocurtique.
   d) La distribution est obligatoirement symétrique.

8. Nous mesurons le nombre de questions de raisonnement mathématique auxquelles un groupe d'élèves du primaire et un groupe de professeurs de mathématiques à l'université ont répondu correctement. La distribution est probablement _____.
   a) platycurtique
   b) symétrique
   c) négativement asymétrique
   d) bimodale

9. Dans cet examen, 90 % des étudiants obtiennent entre 70 et 72. La distribution des notes est fort probablement _____.
   a) platycurtique
   b) leptocurtique
   c) négativement asymétrique
   d) positivement asymétrique

Réponses

1. a
2. b
3. a
4. c
5. 20 %
6. d
7. b
8. d
9. b

Page laissée blanche

CHAPITRE 3
LES STATISTIQUES DESCRIPTIVES

Page laissée blanche

CHAPITRE 3

LES STATISTIQUES DESCRIPTIVES

Dans les deux chapitres précédents, nous avons appris à décrire un échantillon en ordonnant ses valeurs et en comptabilisant ses fréquences absolues ou relatives. Ces distributions permettent d'organiser et de simplifier la masse des observations afin de s'en faire une image globale. Cependant, ce n'est là qu'un premier pas. Il faut ensuite obtenir des prises sur ces données brutes en exécutant des calculs qui réduisent la distribution à quelques valeurs chiffrées qui la synthétisent. Ces indicateurs chiffrés s'appellent des statistiques[1] et l'on nomme *statistiques descriptives* l'ensemble de ces indicateurs. Elles vont servir à simplifier et à organiser les informations dans le but d'en faciliter l'interprétation.

LES STATISTIQUES DE LA TENDANCE CENTRALE

En premier lieu, il est important de définir le concept de *tendance centrale* d'une distribution. La tendance centrale est la valeur la plus typique de la distribution, celle qui la résume le mieux. Elle sert à répondre à des questions telles que : quel est le salaire « typique » d'un joueur de la LNH ? Quelle est la note « typique » des étudiants à un examen ? Quel est le taux d'absentéisme « typique » d'un employé en Italie ? Combien d'enfants la famille nord-américaine « typique » compte-t-elle ? La distribution des effectifs (chapitre 2) nous donne toutes les informations que contient une banque

1. La définition formelle de la « statistique » est discutée au chapitre 8.

de données ; la tendance centrale réduit cette masse à la seule valeur qui la décrit le mieux.

Les joueurs de hockey n'ont pas tous le même salaire, et les familles nord-américaines n'ont pas toutes le même nombre d'enfants. Néanmoins, la connaissance de la valeur typique est une façon très pratique de se faire une idée globale du salaire des joueurs de hockey ou du nombre d'enfants des familles nord-américaines. De manière générale, c'est la statistique de tendance centrale qui aura le plus d'influence sur les décisions prises par chacun dans sa vie.

Il existe plusieurs statistiques de la tendance centrale. Nous en décrirons trois : le mode, la médiane et la moyenne arithmétique.

L'hiver au Québec : manteau ou non ?

En hiver, les étudiants québécois n'ont jamais besoin d'écouter la météo pour décider s'ils doivent ou non porter un manteau pour se rendre à l'université. Ils savent que, « typiquement », il fait froid. Statistiquement, ces étudiants savent que la tendance centrale de la distribution des températures hivernales leur indique qu'« il fait froid » ! Pour compléter cette image, on pourrait essayer de décrire la température à Montréal pendant le mois de février. Armés d'une distribution des effectifs des températures pour le mois de février, on serait en mesure de répondre qu'il fait –40 °C pendant 5 % des jours, –20 °C pendant 15 % des jours, 0 °C pendant 40 % des jours, etc. C'est une information très précise. Cependant, il serait beaucoup plus pratique de répondre, plus simplement, que la température typique pour le mois de février est par exemple de –5 °C. C'est celle qui décrit le mieux la température au cours de ce mois. La mesure de la tendance centrale est alors une manière simple et pratique de décrire une distribution complète tout en sachant, bien sûr, que cette valeur n'est qu'une représentation de la distribution et, par conséquent, qu'elle offre une information moins précise que la distribution complète.

Le mode

Le mode (parfois noté Mo) est la valeur de la distribution dont la fréquence est la plus grande. Le mode, dans le cas d'une distribution groupée de données, est l'intervalle contenant le plus d'observations. Le Tableau 3.1 rapporte la distribution de fréquence simple pour les notes obtenues par 32 étudiants à un examen partiel. La note la plus fréquente étant 75 (3 étudiants l'obtiennent et c'est la seule note aussi fréquente), le mode est égal à 75. Le mode se trouve en examinant la fréquence des valeurs, mais c'est la valeur et non la fréquence qui est le mode. Ainsi, au Tableau 3.1, le mode

est 75 (la note sur la variable « note à l'examen ») et non pas 3 (qui est la fréquence de cette valeur, son effectif).

Une distribution qui contient une seule valeur dominante (la plus fréquente) est une distribution *unimodale*. Il est possible que deux valeurs soient égales et les plus fréquentes dans l'échantillon, la distribution est alors *bimodale* (il y a deux modes). Si la distribution contient plus de deux valeurs qui sont les plus égales et fréquentes, nous parlons alors d'une distribution *multimodale*.

Quiz rapide 3.1
Au Tableau 3.1, éliminez de la distribution les étudiants qui obtiennent une note de 75. Déterminez ensuite le ou les modes pour la distribution des notes restantes. Quelle est alors la forme de cette distribution ?

La présence d'une distribution bimodale indique parfois que deux groupes distincts se trouvent à l'intérieur de la distribution. Par exemple, si on mesure la taille d'une centaine d'hommes et d'une centaine de femmes, il y a de fortes chances pour que la distribution soit bimodale et comprenne une taille typique (modale) pour les femmes et une autre, différente, pour les hommes. Il en est ainsi, car la taille typique des femmes et des hommes n'est pas, en général, la même.

Comment trouver le mode ?

Pour trouver le mode, il suffit d'examiner les effectifs pour chaque valeur de la mesure. Le mode est la valeur qui est associée à l'effectif le plus grand. Aucun calcul n'est requis.

Critique du mode comme statistique de la tendance centrale

Le mode est une valeur de tendance centrale pratique, car il se trouve facilement et il s'agit invariablement d'une valeur existant véritablement dans une distribution. Dans le cas du Tableau 3.1, le mode est 75, et 75 est une véritable note obtenue par des étudiants. Si l'on désire interviewer (pour un article dans un journal, par exemple) l'étudiant dont la connaissance en statistique est « typique », nous choisirons une personne ayant obtenu

le résultat modal (c'est-à-dire celle qui a obtenu 75), car nous savons que nous allons effectivement trouver une telle personne. Cela n'est pas nécessairement le cas avec la médiane ou la moyenne (deux autres mesures de la tendance centrale), comme nous le verrons un peu plus loin.

Cependant, le mode n'est pas la mesure de tendance centrale par excellence. Pour trouver le mode, nous n'avons besoin que d'une seule information, soit la valeur associée à l'effectif le plus grand. Aucune des autres valeurs de la distribution ne l'affecte. Étant défini par seulement une partie de toute l'information disponible, le mode n'est pas toujours la valeur qui décrit le mieux la distribution. L'addition ou le retrait de quelques observations (ou même d'une seule parfois) peut considérablement changer le mode ou ne pas le modifier : en répondant au Quiz rapide 3.1, on a pu remarquer qu'en retirant les trois notes 75 de la distribution, la distribution devenait multimodale (les autres modes étant 49, 65, 71, 76). Quelle serait alors la « véritable » tendance centrale, la note qui décrirait le mieux la performance typique des étudiants à l'examen : 49, 65, 71 ou 76 ? Le mode ne peut pas nous aider à trouver la réponse.

De plus, en ajoutant des observations à l'échantillon, il est possible que le mode ne change pas. À titre d'illustration, si les équipes de la LNH engageaient 100 joueurs de plus, et que chacun d'eux recevait 50 000 000 $ en salaire, le mode ne changerait pas. Il serait toujours de 500 000 $! Le mode est donc une mesure peu démocratique puisque seulement une partie des valeurs l'affecte et que les autres ne comptent pas.

La médiane

La médiane (parfois notée Md) est la mesure de la tendance centrale qui permet de définir la valeur qui coupe la distribution en deux parties, chacune ayant le même nombre d'observations. La note médiane pour la distribution du Tableau 3.1 est 65. Puisque nous avons 32 étudiants dans la distribution, la valeur de la médiane devrait être celle qui coupe l'échantillon en deux, avec 16 étudiants d'un côté et 16 étudiants de l'autre. Si l'on compte 16 notes à partir de la plus petite (29), sans oublier que la note 49 est obtenue par plus d'un étudiant, on remarque que la note 65 coupe la distribution en deux groupes égaux. Dans ce cas, la médiane (Md) est 65.

Tableau 3.1
Distribution des notes à l'examen partiel

Notes	*Fréquence*	*%*	*% cumulatif*
29	1	3,1%	3,1%
30	1	3,1%	6,3%
35	1	3,1%	9,4%
46	1	3,1%	12,5%
49	2	6,3%	18,8%
50	1	3,1%	21,9%
52	1	3,1%	25,0%
55	1	3,1%	28,1%
56	1	3,1%	31,3%
59	1	3,1%	34,4%
61	1	3,1%	37,5%
62	1	3,1%	40,6%
63	1	3,1%	43,8%
64	1	3,1%	46,9%
65	2	6,3%	53,1%
67	1	3,1%	56,3%
70	1	3,1%	59,4%
71	2	6,3%	65,6%
72	1	3,1%	68,8%
74	1	3,1%	71,9%
75	3	9,4%	81,3%
76	2	6,3%	87,5%
77	1	3,1%	90,6%
78	1	3,1%	93,8%
87	1	3,1%	96,9%
90	1	3,1%	100,0%
Total	32	100,0%	
Mode (Mo)	75		
Médiane (Md)	65		
Moyenne (M)	63,25		

Comment trouver la médiane ?

La médiane est correctement calculée lorsque la moitié des observations se trouve au-dessus d'elle et l'autre moitié en dessous. L'observation qui se trouve au milieu d'une distribution est donc la médiane. Pour la trouver, il faut mettre les observations par ordre croissant. La procédure est légèrement différente selon que la distribution contient un nombre pair ou impair d'observations.

Lorsque l'échantillon contient un nombre impair d'observations,

a) on ajoute « 1 » au nombre total d'observations N ;
b) on divise ce total par 2 ;
c) la médiane est la valeur de l'observation qui se trouve à la position calculée à l'étape b.

Illustration : prenons les résultats obtenus par les sept premiers étudiants à l'examen (Tableau 3.1). Ces étudiants ont obtenu les notes suivantes : 29, 30, 35, 46, 49, 49 et 50. Quelle est la médiane pour ces sept observations (N = 7) ?

a) 7 + 1 = 8 ;
b) 8/2 = 4 ;
c) La quatrième observation de la distribution est la médiane. Le quatrième étudiant a obtenu 46 à l'examen. La médiane Md de cette distribution est donc 46.

Vérification des calculs : puisque la médiane est la valeur qui sépare la distribution en deux groupes égaux, lorsqu'elle est correctement calculée, il doit y avoir un nombre égal de personnes obtenant des notes au-dessus et en dessous de la médiane (Md = 46). Trois observations se trouvent au-dessus de 46 (49, 49 et 50) et trois observations se trouvent en dessous (29, 30 et 35). La médiane est donc à la bonne place.

Lorsque l'échantillon contient un nombre pair d'observations,

a) on ajoute « 1 » au nombre total d'observations N ;
b) on divise ce total par 2 (ce calcul donne un chiffre qui se termine par 0,5) ;
c) la médiane se situe entre la valeur de l'observation se trouvant à la position indiquée à l'étape b en enlevant 0,5 et l'observation se trouvant à la position indiquée à l'étape b en ajoutant 0,5. Par exemple, si N = 6 ; 6 + 1 = 7/2 = 3,5. La médiane se situe entre la 3^{e} et la 4^{e} observation.

Illustration : prenons les résultats obtenus par les six premiers étudiants à l'examen (Tableau 3.1). Ces étudiants ont obtenu les notes suivantes : 29, 30, 35, 46, 49 et 49.

a) 6 + 1 = 7 ;
b) 7/2 = 3,5 ; la médiane se trouve entre la note obtenue par le 3e et le 4e étudiant ;
c) L'étudiant à la 3e position a obtenu 35 et celui à la 4e position a obtenu 46 ;
d) La valeur intermédiaire entre 35 et 46 est (35 + 46)/2 = 81/2 = 40,5. La médiane est 40,5.

Vérification des calculs : trois observations se trouvent au-dessus de Md = 40,5 (46, 49 et 49) et trois observations se trouvent en dessous (29, 30 et 35). La médiane est donc à la bonne place.

Concernant les 32 étudiants dont les notes sont inscrites au Tableau 3.1, la médiane se situe entre les 16e et 17e étudiants. Puisque ces deux étudiants (16e et 17e) obtiennent la même note (65), la médiane est la moyenne des deux valeurs, c'est-à-dire 65 [(65 + 65)/2 = 65].

Critique de la médiane comme statistique de la tendance centrale

L'inconvénient principal de la médiane comme mesure de la tendance centrale est qu'elle ne se sert que d'une parcelle de l'information contenue dans la distribution, soit la position relative des observations. Par exemple, les deux distributions suivantes ont exactement la même médiane bien qu'elles soient fort différentes :

Échantillon X : 100, 110, 120, 130, 140

Échantillon Y : 100, 110, 120, 130, 1 000 000

N'utilisant pas toute l'information contenue dans la distribution, la médiane est, en général, une indication moins utile pour définir la tendance centrale.

En contrepartie, cette faiblesse est parfois un avantage. La médiane est une statistique de tendance centrale qui n'est pas affectée par les valeurs qui sont très différentes des autres. Lorsqu'une distribution contient quelques valeurs extrêmement différentes des autres valeurs, il est souvent préférable

de se servir de la médiane pour définir la tendance centrale. En reprenant l'illustration des échantillons X et Y ci-dessus, nous voyons que la médiane est de 120 pour les deux distributions. Si on calcule les moyennes de ces deux distributions (section suivante), celles-ci seront radicalement différentes. Dans un tel cas, la médiane est une meilleure estimation de la valeur typique que ne peut l'être la moyenne.

En science économique par exemple, le revenu médian est beaucoup plus utilisé que le revenu moyen, puisqu'une poignée de personnes ont des revenus dépassant les milliards de dollars (ce qui représente des salaires extrêmement différents de ceux que l'on rencontre habituellement). Le calcul de la moyenne décrirait fort mal le salaire « typique ».

La médiane est principalement utilisée lorsque l'on désire diviser un échantillon en deux groupes de taille identique, dans le but de faire une comparaison entre les deux groupes sur une autre variable. Par exemple, pour déterminer si le nombre d'heures d'étude affecte la note à l'examen, on crée deux groupes, l'un composé des étudiants qui ont obtenu une note sous la médiane, l'autre composé des étudiants qui ont obtenu une note au-dessus de la médiane. On peut maintenant comparer le temps d'étude pour chacun des deux groupes séparément.

La médiane est utile quand on veut obtenir une statistique de la tendance centrale, mais qu'il manque des observations. Par exemple, supposons que les valeurs du Tableau 3.1 représentent non pas les notes à un examen, mais le nombre de minutes que chaque personne prend pour résoudre un problème. Nous voulons trouver le temps typique requis pour y parvenir. Imaginons maintenant une 33^{e} personne qui n'a jamais terminé son problème. Pour déterminer la moyenne (ce que nous verrons plus loin), il faut connaître le temps pris par chaque personne. Or, puisque nous ne connaissons pas le temps requis par cette 33^{e} observation, il devient impossible de calculer la moyenne, sauf si nous la retirons de la distribution. En se servant de la médiane comme mesure de la tendance centrale, l'élimination de cette observation n'est plus nécessaire. Puisque nous avons 33 personnes, la médiane est le temps requis par la 17^{e} personne car $N = 33$ et $(33 + 1)/2 = 17$, et nous pouvons conclure, dans ce cas, que la moitié des personnes prend *moins* de 65 minutes et l'autre moitié prend *plus* de 65 minutes pour résoudre le problème.

Quiz rapide 3.2
Trouvez la médiane pour les cinq et six dernières observations du Tableau 3.1. Ajoutez dans les deux cas une dernière observation inconnue. Obtenez-vous un résultat différent de celui que vous aviez trouvé ?

La moyenne arithmétique

La moyenne arithmétique (parfois notée M) est probablement la statistique la plus utile et la plus fréquemment utilisée, aussi bien dans la vie scientifique et professionnelle que dans la vie de tous les jours[2]. Il suffit de penser, par exemple, à une note scolaire moyenne. Facile à calculer, la moyenne possède un ensemble de propriétés et de caractéristiques qui en font la valeur de la tendance centrale représentant le mieux la distribution et qui, par conséquent, est celle qu'on utilise généralement le plus.

Comment trouver la moyenne ?

Pour trouver la moyenne, il suffit d'additionner la valeur de chaque observation et de diviser ce total par le nombre d'observations. La formule pour trouver la moyenne M est :

$$\mathbf{M} = \left(\sum_{i=1}^{N} \mathbf{X_i}\right)/\mathbf{N} \qquad \text{Formule 3.1}$$

X_i est la valeur obtenue sur la variable X pour chaque observation i (i allant de 1 à N, la dernière personne), $\sum$ (sigma majuscule) est le symbole qui indique une sommation et N est le nombre total d'observations.

La formule se lit de la manière suivante : la moyenne (M) de la variable X est égale à la somme ($\sum$) des observations (X_i) divisée par le nombre (N) d'observations. À partir du Tableau 3.1, le Tableau 3.2 en donne un exemple en calculant la moyenne obtenue à l'examen partiel par 11 étudiants.

2. Nous décrivons la moyenne arithmétique (ou la moyenne tout court), mais il existe deux autres sortes de moyenne : la moyenne géométrique et la moyenne harmonique. Ces deux dernières formes de la moyenne sont expliquées dans les textes statistiques plus avancés.

Tableau 3.2
Calcul de la moyenne

Notes à l'examen	65	65	67	70	71	73	74	75	75	75	82

Somme $\sum X_i = 792$
$N = 11$
$M = 792/11 = \mathbf{72}$

Critique de la moyenne comme statistique de la tendance centrale

Lorsqu'il s'agit de trouver la valeur typique d'une distribution, la moyenne a beaucoup plus d'avantages que d'inconvénients, mais, néanmoins, elle a deux inconvénients principaux.

D'une part, la moyenne est souvent une valeur abstraite que l'on ne retrouvera pas nécessairement dans les données. Par exemple, la moyenne de la note obtenue à l'examen (Tableau 3.1) est 70,9. Si on étudie la distribution des notes, on constate que personne n'a obtenu cette note à l'examen. Si la femme canadienne moyenne a 1,24 enfant, et qu'un journaliste souhaite faire un reportage sur elle, même en cherchant longtemps, il aura bien du mal à la trouver !

D'autre part, lorsque la distribution des données est très asymétrique, la moyenne présente une image qui peut être trompeuse. Un bel exemple nous est donné par les salaires des joueurs de hockey de la LNH. Le Tableau 3.3 montre leur salaire moyen, médian et modal. Dans ce tableau, nous voyons que le salaire moyen (1 700 000 $) représente plus que le triple du salaire modal (500 000 $). Même s'il est indéniable qu'en moyenne les joueurs de hockey gagnent 1 700 000 $, le salaire le plus fréquent (le mode) n'est qu'une fraction de ce montant, et le salaire médian lui aussi est bien inférieur (1 000 000 $) au salaire moyen. Dans ce cas, il serait plus raisonnable de dire que le salaire typique des joueurs de la LNH se situe davantage aux alentours de 500 000 que de 1 700 000 $.

Malgré ces inconvénients, la moyenne est néanmoins l'estimation par excellence de la tendance centrale d'un échantillon. Voyons pourquoi.

Tableau 3.3
Moyenne, médiane et mode des salaires des joueurs de la LNH, 2002-2003

Statistiques de la tendance centrale	*Salaires*
Moyenne	1 708 305,82 $
Médiane	1 000 000,00 $
Mode	500 000,00 $

La moyenne utilise toutes les informations disponibles. La valeur de la tendance centrale doit être une représentation aussi parfaite que possible de la distribution. Le mode ne se sert que d'une parcelle des valeurs de la distribution (seule la valeur la plus fréquente est prise en considération). La médiane ne compte que la position des observations. La valeur des observations individuelles n'est pas pertinente. Pour calculer la moyenne, par contre, on a besoin de la totalité de l'information contenue dans la distribution. Puisque chaque valeur de la distribution contribue à la moyenne, c'est elle qui décrit le mieux la distribution complète. Chaque valeur de la distribution, sans exception, « a son mot à dire » lorsqu'il s'agit de calculer la moyenne. La moyenne est donc la statistique de la tendance centrale qui est la plus démocratique !

La moyenne est la statistique de la tendance centrale qui fait le moins d'erreurs. Le second avantage de la moyenne provient du fait qu'elle fait le moins d'erreurs lorsqu'elle est utilisée pour « prédire » chaque valeur de la distribution. On se rappellera que la tendance centrale doit indiquer la valeur typique, c'est-à-dire la valeur qui décrit le mieux toutes les autres valeurs de la distribution. Reprenons les notes du Tableau 3.2 dans le Tableau 3.4.

Tableau 3.4
Erreur moyenne : comparaison des valeurs de la tendance centrale : M, Md, Mo

Numéro de l'observation	*Notes à l'examen*	$X_i - M$ *Erreur à la moyenne (M = 72)*	$X_i - Md$ *Erreur à la médiane (Md = 73)*	$X_i - Mo$ *Erreur au mode (Mo = 75)*
1	**65**	65 – 72 = – 7	65 – 73 = – 8	65 – 75 = – 10
2	**65**	65 – 72 = – 7	65 – 73 = – 8	65 – 75 = – 10
3	**67**	67 – 72 = – 5	67 – 73 = – 6	67 – 75 = – 8
4	**70**	70 – 72 = – 2	70 – 73 = – 3	70 – 75 = – 5
5	**71**	71 – 72 = – 1	71 – 73 = – 2	71 – 75 = – 4
6	**73**	73 – 72 = +1	73 – 73 = 0	73 – 75 = – 2
7	**74**	74 – 72 = +2	74 – 73 = +1	74 – 75 = – 1
8	**75**	75 – 72 = +3	75 – 73 = +2	75 – 75 = 0
9	**75**	75 – 72 = +3	75 – 73 = +2	75 – 75 = 0
10	**75**	75 – 72 = +3	75 – 73 = +2	75 – 75 = 0
11	**82**	82 – 72 = +10	82 – 73 = +9	82 – 75 = +7
Total	792	0	– 11	– 27
N	11	11	11	11
Moyenne	792/11 = **72**			

Dans le Tableau 3.4, la moyenne est M = 72, la médiane est Md = 73 et le mode est Mo = 75 ; N = 11. Laquelle de ces trois statistiques de la tendance centrale est la plus représentative de toutes les valeurs de la distribution ? Pour répondre à cette question, il faut définir l'expression « la plus représentative ». Dans ce dessein, on choisit l'écart par rapport à la mesure de tendance centrale, c'est-à-dire la différence entre la valeur réelle de chaque observation et la valeur de la tendance centrale. Cette différence s'appelle l'*erreur*. Ainsi, la meilleure mesure de tendance centrale devrait être celle qui fait le moins d'erreurs lorsque l'on s'en sert pour prédire chaque valeur de la distribution. On peut faire l'exercice avec les données reproduites

au Tableau 3.4. On prend chaque valeur de la distribution, de laquelle on soustrait respectivement la moyenne, la médiane et le mode. Plus grande est cette différence, plus grande est l'erreur produite par cette statistique.

Au Tableau 3.4, nous observons que la première observation obtient une valeur réelle de 65 alors que la moyenne, la médiane et le mode sont respectivement de 72, 73 et 75. Si la moyenne représente parfaitement cette observation, elle devrait avoir la même valeur (65) que l'observation. Puisque la moyenne est égale à 72, il est clair que la moyenne fait une erreur de –7 (65 – 72 = –7). Le signe négatif signifie que la moyenne surestime la donnée (la véritable valeur de l'observation est plus faible que la moyenne). Lorsque la différence produit un signe positif, cela signifie que la moyenne sous-estime la donnée (la valeur de l'observation est plus grande).

Comparons maintenant l'erreur faite par la médiane, le mode et la moyenne lorsqu'on les utilise pour « prédire » la première observation du Tableau 3.4. L'erreur faite par la moyenne, dans ce premier cas (–7) est plus petite que les erreurs occasionnées par la médiane et le mode (–8 et –10 respectivement). Cela n'est pas toujours le cas pour toutes les observations : pour l'observation 7, par exemple, les erreurs faites par la moyenne sont plus fortes (+2) que celles faites respectivement par la médiane et le mode (+1 et –1).

Le Tableau 3.4 montre les erreurs pour toutes les observations. Nous pouvons alors déterminer l'erreur produite par chaque mesure de la tendance centrale. Nous faisons la somme des erreurs et nous observons que l'erreur totale faite par la moyenne vaut 0 alors que celles du mode et de la médiane valent respectivement –11 et –27. Si nous calculons l'erreur moyenne faite par la moyenne, la médiane et le mode, nous trouvons respectivement 0, –1 et –2,47.

La mesure de tendance centrale qui produit le moins d'erreur totale ou la plus petite erreur moyenne est celle qui décrit le mieux la distribution et, clairement, la moyenne en fait le moins. Ce résultat n'est pas un accident : *invariablement, l'erreur totale (et l'erreur moyenne) produite par la moyenne est égale à zéro* et, sauf pour le cas où la moyenne, la médiane et le mode sont identiques, l'erreur faite par le mode et la médiane sera plus grande. Il s'ensuit que la moyenne est la mesure de tendance centrale qui

représente le mieux les données d'un échantillon. Pour ceux qui aiment l'algèbre, l'encadré en fait la preuve mathématique.

L'erreur moyenne autour de la moyenne est égale à zéro : une preuve mathématique

Il est possible de démontrer en termes mathématiques que, peu importe l'échantillon, la somme des erreurs entre chaque donnée (X_i) et sa moyenne (M_X) est toujours nulle.

$$\begin{aligned}\frac{1}{N}\sum_i (X_i - M_X) &= \frac{1}{N}\left(\sum_i X_i - \sum_i M_X\right)\\ &= \frac{1}{N}\left(NM_X - \sum_i M_X\right)\\ &= \frac{1}{N}(NM_X - NM_X)\\ &= \frac{1}{N}0 = 0\end{aligned}$$

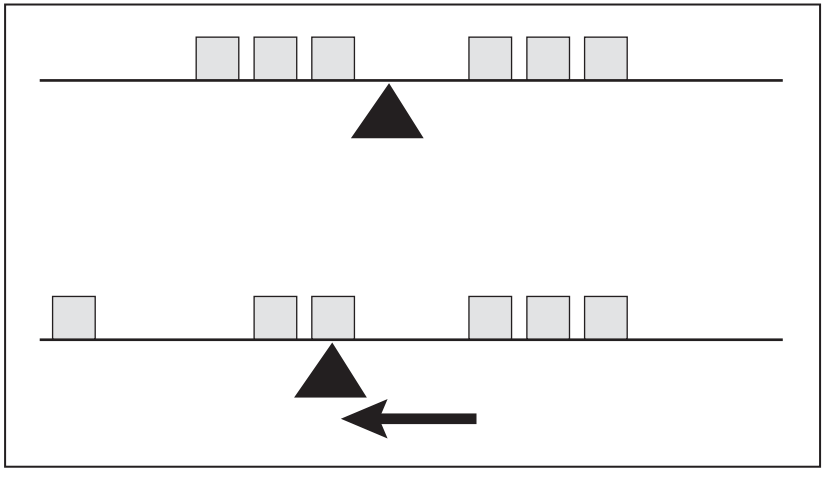
Invariablement, l'erreur moyenne sera, elle aussi, égale à zéro.

Quiz rapide 3.3
Générez une distribution contenant 5 données et vérifiez que la moyenne ne fait aucune erreur de prédiction en moyenne. Le mode et la médiane en font-ils plus ?

La moyenne est le point d'équilibre d'une distribution. L'erreur de prédiction moyenne est toujours égale à zéro quand on utilise la moyenne pour « prédire » chaque valeur de la distribution. Pour que cela soit vrai, il faut que la somme des erreurs supérieures à la moyenne et la somme des erreurs inférieures à la moyenne soient égales. Par conséquent, la moyenne est souvent interprétée comme étant *le point d'équilibre d'une distribution*. La Figure 3.1 propose une représentation graphique de la situation. Imaginons que les cubes sont des enfants de même poids assis sur une balançoire à bascule. Le triangle représente le point d'équilibre de la balançoire et les enfants sont disposés des deux côtés de ce point d'équilibre. Dans la situation A, nous voyons que la barre est en équilibre lorsque les deux groupes d'enfants sont exactement à la même distance du point d'équilibre (le triangle). Dans la situation B, nous déplaçons un des enfants près de l'extrémité gauche de la balançoire à bascule. Pour garder

la barre horizontale, il devient nécessaire de déplacer le point d'équilibre vers la gauche, plus près de l'enfant que nous avons déplacé. La moyenne agit comme le triangle de la Figure 3.1 : elle a tendance à se déplacer vers les valeurs les plus extrêmes de la distribution.

FIGURE 3.1 L a moyenne comme point d'équilibre d'une distribution



La Figure 3.2 reprend la même idée, mais, cette fois, en montrant la façon dont la moyenne, la médiane et le mode sont influencés par trois formes de distribution (voir le chapitre 2) : une distribution symétrique, une distribution asymétrique négative et une distribution asymétrique positive. Dans la situation A, on remarque que la moyenne, la médiane et le mode coïncident tous exactement. Lorsque les trois valeurs de la tendance centrale d'une distribution coïncident, la distribution est symétrique.

La situation B montre une distribution asymétrique. La moyenne est maintenant déplacée vers la droite, vers les observations extrêmes qui se trouvent du côté positif de l'abscisse. Lorsque la moyenne est décalée vers la droite de l'abscisse par rapport à la médiane, la distribution est asymétrique positive.

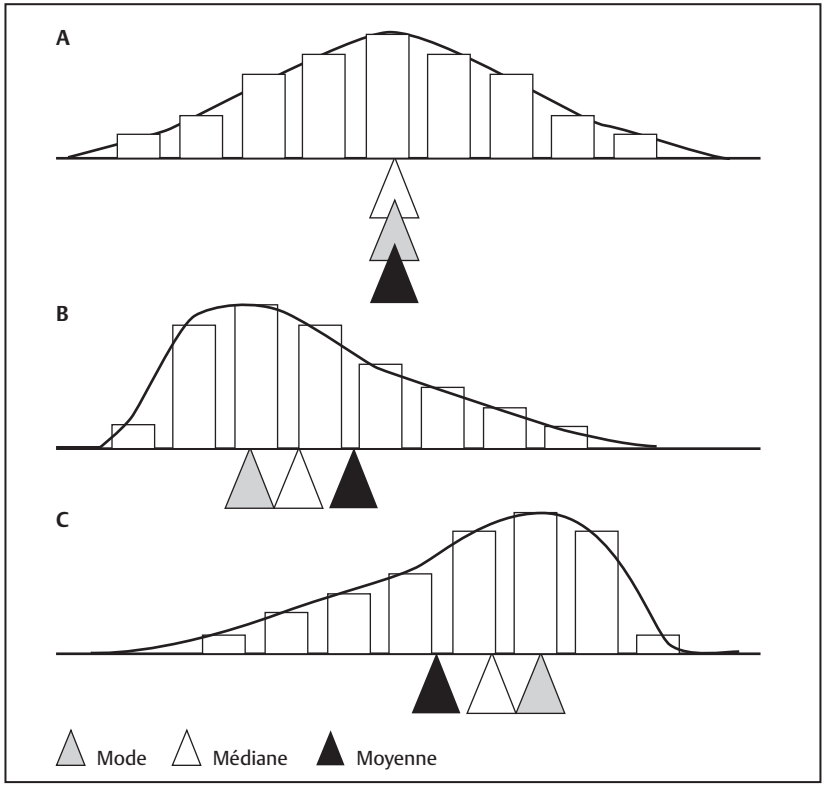
Dans la situation C, l'asymétrie est négative parce que la moyenne est décalée vers la gauche (la partie ayant des valeurs plus faibles) de l'abscisse.

Ainsi, l'asymétrie fait référence à la position de la moyenne par rapport à la médiane. Lorsque l'asymétrie est positive, la moyenne se trouve à la

droite de la médiane, et elle se trouve à sa gauche lorsque la distribution est asymétrique négative. Ainsi, lorsque la moyenne est numériquement supérieure à la médiane, la distribution est asymétrique positive.

À l'inverse, lorsque la moyenne est numériquement inférieure à la médiane, la distribution est asymétrique négative.

FIGURE 3.2 Mode, médiane et moyenne pour différentes formes de distribution



Quiz rapide 3.4
La distribution des salaires des joueurs de hockey est asymétrique. Est-elle positivement ou négativement asymétrique ? Pouvez-vous déduire, à partir de l'asymétrie, si la moyenne est plus ou moins forte que la médiane ?

LES MESURES DE DISPERSION

La moyenne est certes une des statistiques les plus importantes que l'on puisse calculer afin de comprendre une distribution. C'est une synthèse qui donne la meilleure représentation d'une distribution. Mais ce n'est pas parce qu'elle produit la « meilleure » estimation de toutes les valeurs d'une distribution qu'elle est une « bonne » estimation de la distribution (le salaire des joueurs de la LNH est un exemple de ce concept).

L'hiver au Québec : autobus ou métro ?

Supposons que, pour se rendre à un cours, on ait le choix entre l'autobus et le métro. Lequel prendre ? Tout dépend du temps de trajet moyen de l'un et l'autre mode de transport. S'il faut en moyenne 30 minutes en métro et 45 minutes en autobus, alors on prend certainement le métro. Imaginons par contre que les deux modes de transport prennent en moyenne 30 minutes. Doit-on en préférer un ? Supposons que l'autobus met entre 10 et 50 minutes pour parcourir le trajet, alors que le métro met entre 25 et 35 minutes. Puisqu'il est impératif d'être à l'heure à tous ses cours, il vaut mieux éviter l'autobus qui peut réserver de mauvaises surprises. Dans cet exemple, la fiabilité dans la durée du trajet est indiquée par la variabilité : plus la durée est variable, moins on a de chances que la durée moyenne soit la durée réelle du trajet.

Prenons un deuxième exemple : comptons le nombre de nez (oui !) sur le visage de chaque étudiant dans une classe et calculons la moyenne du nombre de nez. Il n'est pas nécessaire d'avoir un ordinateur pour savoir que la moyenne de la variable « nombre de nez » sera égale à 1. Maintenant, utilisons cette moyenne pour prédire le nombre de nez qu'un étudiant, aléatoirement choisi, possède. Dans ce cas, il est quasi certain que la moyenne sera une estimation parfaite du nombre de nez de cette personne (nous n'avons tous habituellement qu'un nez). Répétons l'expérience, mais cette fois, analysons non pas le nombre de nez, mais la taille des étudiants. Calculons la moyenne (disons qu'on obtient 1,70 m) et essayons de prédire la taille d'un étudiant choisi au hasard. Puisque la moyenne est la meilleure estimation, nous allons prédire que cette personne mesure 1,70 m. Mais, à moins d'avoir beaucoup de chance, il est probable que l'étudiant choisi aura une taille différente. Dans ce dernier cas, la moyenne est une moins bonne estimation de la taille, même si elle reste la meilleure estimation disponible.

Qu'est-ce qui fait la différence entre une bonne et une moins bonne estimation ? Si la distribution contient des valeurs très similaires (voire identiques, comme le nombre de nez), la moyenne est une bonne (et à la limite, une parfaite) estimation des valeurs de l'échantillon. Si la distribution contient des valeurs qui diffèrent beaucoup entre elles, la moyenne est une moins bonne estimation. Pour décrire adéquatement une distribution, il faut par conséquent trouver un moyen de quantifier non seulement sa moyenne, mais aussi le degré de différence entre les observations.

L'étendue

Nous avons vu, en construisant la distribution des salaires des joueurs de hockey que, même si la moyenne des salaires est d'environ 1 700 000 $, certains joueurs gagnent moins de 200 000 $ et d'autres reçoivent jusqu'à 11 000 000 $. À l'aide de la construction d'une distribution des effectifs, il est facile de déterminer le salaire le plus élevé et le salaire le plus faible. En comparant ces deux extrêmes (165 000 et 11 000 000 $), il est clair que les salaires peuvent être très différents.

La différence entre les deux extrêmes d'une distribution produit une première statistique qui reflète le degré de dispersion (de différence). Cette statistique, *la différence entre la valeur maximale et minimale,* s'appelle l'*étendue.*

Comment calculer l'étendue ?

L'étendue se calcule en soustrayant la valeur la plus faible de la valeur la plus forte d'une distribution. Il est à remarquer que la fréquence des observations n'est pas pertinente pour ce calcul.

$$\text{Étendue} = X_{(max)} - X_{(min)} \qquad \text{Formule 3.2}$$

où $X_{(max)}$ est la valeur la plus grande observée dans la distribution et $X_{(min)}$ la plus petite.

Puisque le joueur le mieux payé de la LNH reçoit la somme de 11 000 000 $ et que le moins bien payé reçoit 165 000 $, l'étendue est $X_{(max)} - X_{(min)}$ = 11 000 000 $ – 165 000 $ = 10 835 000 $. Les salaires payés aux

joueurs de la LNH varient et la différence entre le mieux payé et le moins bien payé est très grande.

Quiz rapide 3.5
Quelle est l'étendue des salaires pour les joueurs de Montréal ? Pour les joueurs d'Atlanta ? Si vous n'utilisez pas le site Internet (www.pum.umontreal.ca/ca/fiches/978-2-7606-2113-8.html), quelle est l'étendue pour la distribution du Tableau 3.2 ?

Critique de l'étendue comme statistique de la dispersion

La principale force de l'étendue comme mesure de la dispersion est sa facilité de calcul. En revanche, l'étendue est une mesure grossière de la dispersion, car elle n'utilise qu'une infime partie des informations contenues dans la distribution, en l'occurrence seulement les deux observations extrêmes. Ainsi, si tous les joueurs de la LNH recevaient un salaire de 165 000 $, sauf un qui recevrait 11 000 000 $, l'étendue serait identique à celle que nous venons de calculer. Pourtant, les deux distributions ne seraient vraiment pas les mêmes. De plus, l'étendue est une valeur relativement instable. En fait, une observation à elle seule peut faire changer considérablement l'étendue. Par exemple, si nous ajoutons un joueur dont le salaire est de 20 000 000 $, l'étendue serait maintenant le double (passant de 10 835 000 à 19 835 000 $). Ainsi, si l'étendue est utile pour nous fournir une statistique rapide pour évaluer le niveau de différence qui existe dans la distribution, il faut savoir que cette statistique a de fortes chances d'être instable. Lorsque nous travaillons avec des distributions construites avec un petit nombre d'observations, l'étendue sera presque certainement instable.

L'étendue interquartile

Puisque l'étendue est toujours sensible aux valeurs extrêmes d'un échantillon, nous pouvons améliorer la technique en calculant une étendue différente qui prend le nom d'*étendue interquartile*. Au lieu de comparer la valeur la plus grande et la valeur la plus petite de l'échantillon, l'étendue interquartile compare la différence entre deux autres valeurs, qui, elles,

sont plus stables. Nous savons que les observations tendent à être plus rares aux extrémités de la distribution et qu'elles sont plus fréquentes autour de la moyenne. Alors si nous calculons les étendues à partir des valeurs plus proches de la moyenne, le résultat obtenu aura tendance à être plus stable. Par convention, on définit « plus proche de la moyenne » 50 % des observations qui se situent autour de la médiane (+25 % et –25 %).

Comment trouver l'étendue interquartile ?

Nous verrons la définition exacte du quartile au prochain chapitre. En principe, l'étendue interquartile se calcule en retirant de la distribution 25 % des scores les plus faibles, 25 % des scores les plus élevés, puis en calculant l'étendue sur les données restantes. En examinant les pourcentages cumulatifs au Tableau 3.1, nous voyons que 25 % des étudiants obtiennent une note égale ou inférieure à 52 et que 72 % des étudiants obtiennent une note égale ou inférieure à 74 (72 % est le pourcentage le plus proche de 75 % dans le tableau). Nous pouvons maintenant calculer l'étendue interquartile, la différence entre ces deux quantités. L'étendue interquartile est proche de 22 (74 –52 = 22).

Critique de l'étendue interquartile comme statistique de la dispersion

L'étendue interquartile est plus stable que l'étendue. Elle est particulièrement utile lorsque nous travaillons avec des distributions très asymétriques où quelques observations peuvent se trouver très loin de la moyenne. Par exemple, pour les salaires des joueurs de hockey, l'étendue est supérieure à 10 000 000 $, mais l'étendue interquartile est de 1 500 000 $. De là, nous pouvons conclure que, même si les salaires couvrent un très large éventail, la différence entre les salaires de la majorité des joueurs n'est pas aussi grande (étendue interquartile).

L'étendue interquartile est plus stable que l'étendue, car l'ajout d'un joueur avec un salaire très élevé ou très faible ne la changera pas. Néanmoins, l'étendue interquartile n'est pas la statistique de dispersion la plus stable, car elle n'utilise qu'une petite partie de l'information disponible (seulement les deux valeurs qui définissent 50 % des observations). Il faut

trouver une façon de mesurer la dispersion des valeurs qui prenne en considération toutes les valeurs de la distribution. La *variance autour de la moyenne* est la statistique qui remplit cette condition.

LA VARIANCE AUTOUR DE LA MOYENNE

La variance est liée, sous une forme ou une autre, à la quasi-totalité des règles et des techniques statistiques. Il importe de bien la comprendre, car cette statistique revient constamment dans ce livre, de même que dans tous les livres de statistiques. Pour intégrer ce concept, il faut préalablement comprendre *le concept de l'erreur autour de la moyenne*.

Le concept de l'erreur autour de la moyenne revisité

Bien que la moyenne soit la meilleure estimation des valeurs d'une distribution, nous voulons savoir à quel point la moyenne détermine *avec précision* chaque observation individuelle. La moyenne est bonne lorsque l'erreur, autrement dit *l'écart* entre chaque observation et la moyenne, est petite. Si les écarts entre les observations et la moyenne sont petits, cela implique que la différence entre les observations est petite. Lorsque les écarts entre les observations et la moyenne sont grands, la différence entre les observations est plus grande et la moyenne est une moins bonne estimation des valeurs individuelles de la variable.

Le Tableau 3.5 présente les données pour deux échantillons notés X et Y, chacun composé de trois observations. Ces deux échantillons ont une même moyenne de 60. Cependant, il est clair que les valeurs de la distribution X (59, 60, 61) sont très similaires alors que les valeurs de la distribution Y sont très différentes (40, 60, 80). Il est donc certain que la moyenne sera une bonne estimation pour X et une estimation beaucoup moins bonne pour Y. Essayons maintenant de concevoir une approche qui pourra confirmer quantitativement notre intuition.

Tableau 3.5
Le concept de l'erreur à la moyenne

	Distribution X		
	Score	***Erreur*** $(X_i - M_X)$	***Erreur quadratique*** $(X_i - M_X)^2$
	59	–1	1
	60	0	0
	61	+1	1
Somme	180	0	2
N	3	3	3
Moyenne	180/3 = 60	0	
Écart moyen	—	0/3 = 0	2/3 = 0,67
Variance $\Sigma(X_i - M_X)^2/(N-1)$	—	0/2 = 0	2/2 = 1
	Distribution Y		
	Score	***Erreur*** $Y_i - M_Y$	***Erreur quadratique*** $(X_i - M_X)^2$
	40	–20	400
	60	0	0
	80	+20	400
Somme	180	0	800
N	3	3	3
Moyenne	60	0	
Écart moyen	—	0/3 = 0	800/3 = 266,67
Variance $\Sigma(X_i - M_X)^2/(N-1)$	—	0/2 = 0	800/2 = 400

Calculons l'erreur produite lorsqu'on se sert de la moyenne pour estimer chaque donnée du Tableau 3.5. Comme nous l'avons vu plus tôt, l'erreur est la différence entre chaque valeur et la moyenne ; les erreurs apparaissent dans la colonne « Erreur », que nous notons $(X_i – M_X)$, où X_i

représente chaque observation i et M_X la moyenne de la distribution. Notez que les écarts sont positifs lorsque la valeur X_i est plus grande que la moyenne, et négatifs dans le cas contraire (par exemple 59–60 = –1, alors que 61–60 = +1).

En étudiant les erreurs dans le Tableau 3.5, on voit qu'elles sont plus grandes pour l'échantillon Y (–20,0 et +20) que pour l'échantillon X (–1,0 et +1). On peut donc conclure que la moyenne de la distribution X (M_X) fait moins d'erreurs lorsqu'elle est utilisée pour reproduire les valeurs individuelles alors que la moyenne pour la distribution Y (M_Y) en fait plus.

Pour synthétiser les écarts à un seul nombre, on peut calculer l'écart moyen : on additionne les écarts et on calcule leur moyenne avec la Formule 3.1. Or, nous tombons sur un obstacle : la somme des écarts à la moyenne est toujours zéro et une division par le nombre d'observations aboutit invariablement à un écart moyen de zéro !

$$SC = \sum_{i=1}^{N} (X_i - M_X) = 0 \qquad \text{Formule 3.3}$$

En effet, comme nous l'avons vu plus tôt, la moyenne étant le point d'équilibre d'une distribution, la somme des écarts positifs (sous-estimation) est invariablement égale à la somme des écarts négatifs (surestimation). La somme des écarts positifs et négatifs est obligatoirement égale à zéro. Calculons la somme des écarts pour la distribution X du Tableau 3.5 : +1 + 0 + (–1)= 0. Cette somme est bien zéro.

Puisque l'écart moyen à la moyenne est toujours égal à zéro, il s'agit d'une statistique inutile. Nous devons trouver une façon d'éliminer ce problème. Une solution possible est d'enlever le signe des écarts avant de calculer la moyenne. Enlever le signe est une opération mathématique qui s'appelle « prendre la valeur absolue », notée par des barres verticales | |. Si on recalcule l'écart moyen au Tableau 3.5 en ignorant le signe de chaque différence, on voit que l'écart absolu moyen pour la distribution X (c'est-à-dire 0,67) est moins grand que l'écart pour la distribution Y (13,33). Nous pouvons donc conclure que la moyenne est une meilleure estimation pour l'échantillon X qu'elle ne l'est pour l'échantillon Y, ce qui est conforme avec notre intuition dans ce cas. Cependant, les valeurs absolues ont des propriétés mathématiques peu pratiques. Examinons une autre solution.

On sait que la multiplication de deux valeurs négatives donne un produit positif. Ainsi, (–2) X (–2) = +4. Aussi, mettre une valeur au carré donne toujours une valeur positive, que la valeur originale soit négative ou positive. Ainsi, $(-2)^2 = (-2) \text{ X } (-2) = +4$. Ce fait ouvre la porte à une manière pratique de s'assurer que la somme des écarts n'est pas obligatoirement zéro. Il suffit de mettre chaque écart au carré. Cette quantité est appelée l'*écart quadratique* ou, plus simplement, l'*écart au carré*. Si nous additionnons ensemble tous les écarts au carré, nous obtenons une statistique qui prend le nom de *somme des écarts quadratiques* ou, plus simplement, *somme des carrés (SC)*. Cette statistique reflète le degré de différence entre les valeurs. La Formule 3.4 formalise cette statistique.

$$SC = \sum_{i=1}^{N} (X_i - M_X)^2 \qquad \text{Formule 3.4}$$

Il faut noter, au Tableau 3.5, que la somme des écarts au carré est plus grande pour la distribution Y ($SC_Y = 800$) que pour la distribution X ($SC_X = 2$). Cela indique qu'il existe plus de différence (au carré) entre la moyenne et les observations de la distribution Y qu'il n'en existe pour la distribution X, ce qui correspond, encore une fois, à notre intuition initiale. La somme des écarts au carré est toujours plus grande que zéro, sauf dans un cas. Le Quiz rapide 3.6 invite à déduire ce cas particulier.

Quiz rapide 3.6
Il existe un cas particulier où la somme des écarts quadratiques d'une distribution est égale à zéro. Lequel ?

Critique de la somme des erreurs au carré comme statistique de la dispersion

La somme des écarts au carré a des caractéristiques utiles :

a) elle est facile à calculer ;
b) plus les observations sont différentes, plus la somme des erreurs au carré est grande, ce qui indique que les valeurs de la distribution sont plus dispersées.

En calculant la somme des erreurs au carré (SC), on obtient en réalité deux informations. D'une part, elle indique le degré avec lequel la moyenne est une bonne ou une moins bonne estimation de chaque valeur de l'échan-

tillon : plus grande est la quantité SC, moins bonne est la moyenne. Par ailleurs, et cette information est peut-être encore plus importante, ce calcul indique dans quelle mesure la variable mesurée produit des observations qui ont des valeurs différentes, c'est-à-dire le degré de *variabilité* qui existe au sein d'une variable.

Mais la somme des erreurs au carré souffre d'un inconvénient important qui fait d'elle une mesure sous-optimale de la dispersion des valeurs d'une distribution : sa taille est simultanément influencée par

a) la taille des différences entre les observations et la moyenne ;
b) le nombre d'observations. Plus on a d'observations, plus la somme des écarts au carré est grande.

Il faut séparer ces deux influences. La *variance autour de la moyenne,* qui se nomme habituellement *la variance,* est la procédure statistique qui corrige le problème.

Le calcul de la variance autour de la moyenne

La variance d'une distribution, généralement notée par le symbole s^2, est définie par la Formule 3.5 :

$$s^2 = \sum_{i=1}^{N} (X_i - M_X)^2 / N - 1 \qquad \text{Formule 3.5}$$

où $\sum_{i=1}^{N} (X_i - M_X)^2$ est la somme des carrés, et N est le nombre d'observations.

La variance est la somme des carrés divisée par le nombre d'observations moins 1. En divisant par le nombre d'observations (moins 1), nous obtenons la différence moyenne (au carré), ce qui a pour effet de séparer les deux influences sur la somme des carrés : la taille des différences et le nombre d'observations[3]. La variance est la statistique qui indique le degré moyen de précision (au carré) de la moyenne pour estimer chaque valeur de l'échantillon. Plus grande est la dispersion des valeurs d'une distribu-

3. La variance est presque toujours obtenue en divisant par N – 1, mais parfois il faut la calculer en divisant par N. Ces deux façons de calculer la moyenne seront expliquées plus en détail au chapitre 8. Le choix de diviseur N ou N – 1 exige une compréhension des concepts de l'échantillon et de la population, concepts qui sont abordés aux chapitres 8 et 9.

tion, plus grande est la variance et, par conséquent, moins bonne est la moyenne comme estimateur de chaque observation.

Le calcul de la variance pour chacun des échantillons X et Y est donné au Tableau 3.5.

Quiz rapide 3.7
Ajoutez à la distribution X du Tableau 3.5 un nouveau score de 60. La variance augmente-t-elle ? Pourquoi ? Ensuite, ajoutez à X un score de 20. La variance augmente-t-elle ? Pourquoi ?

Critique de la variance comme statistique de la dispersion

Cette façon de conceptualiser la dispersion des observations, c'est-à-dire la variance, comporte plusieurs avantages.

- La variance prend en considération toutes les valeurs de la distribution, pas seulement ses extrêmes (comme le fait l'étendue ou l'étendue interquartile).
- Elle ne prend jamais de valeur négative (il est impossible que la différence soit plus petite que zéro) et elle est généralement plus grande que zéro (sauf, naturellement, si la variable est une constante).
- La variance est une statistique stable. Lorsque la distribution est composée d'une trentaine d'observations ou plus, l'ajout de valeurs supplémentaires ne changera pas beaucoup la variance, et ce, dans la majorité des situations.

La variance, comme mesure de dispersion, souffre d'un inconvénient majeur. Elle rapporte l'erreur moyenne *au carré*. Nous n'avons pas l'habitude de penser en termes d'erreurs au carré, ce qui rend son interprétation plutôt difficile. Si l'on prend, par exemple, le Tableau 3.6 qui présente les notes à un examen évalué sur 100. Le calcul de la variance donne $s^2 = 219{,}75$. Puisque les valeurs possibles pour les notes à l'examen sont habituellement entre 0 et 100, le chiffre 219,75 est difficile à interpréter. Nous ne pouvons certainement pas dire que la différence moyenne entre les notes obtenues à l'examen est 219,75 ! Pour arriver à une conclusion plus raisonnable, il faut exprimer la différence moyenne avec un chiffre qui reflète la variable originale avec plus de réalisme (dans le cas de la performance à l'examen, les chiffres qui décrivent la dispersion des notes devraient être entre 0 et 100). L'*écart-type* est la statistique qui répond à ce besoin.

Tableau 3.6
Distribution des notes à l'examen partiel (bis)

	Notes	$(X_i - M)$	$(X_i - M)^2$
	29	–34,25	1173,06
	30	–33,25	1105,56
	35	–28,25	798,06
	46	–17,25	297,56
	49	–14,25	203,06
	49	–14,25	203,06
	50	–13,25	175,56
	52	–11,25	126,56
	55	–8,25	68,06
	56	–7,25	52,56
	59	–4,25	18,06
	61	–2,25	5,06
	62	–1,25	1,56
	63	–0,25	0,06
	64	0,75	0,56
	65	1,75	3,06
	65	1,75	3,06
	67	3,75	14,06
	70	6,75	45,56
	71	7,75	60,06
	71	7,75	60,06
	72	8,75	76,56
	74	10,75	115,56
	75	11,75	138,06
	75	11,75	138,06
	75	11,75	138,06
	76	12,75	162,56
	76	12,75	162,56
	77	13,75	189,06
	78	14,75	217,56
	87	23,75	564,06
	90	26,75	715,56
Somme	2 024	0	7 032,00
N	**32**	32	31
Résultat	**63,25**	0	219,75
Nom de la statistique	**Moyenne**	**Écart moyen**	**Variance**

L'écart-type

La variance se calcule avec la Formule 3.5. La formule fait la somme des écarts au carré qui est divisée par le nombre d'observations (moins 1). Ainsi, la variance décrit « la différence moyenne au carré », une quantité qu'il est difficile de se représenter. Pour éliminer cette difficulté, il s'agit simplement d'extraire la racine carrée de la variance (Formule 3.6). Ce calcul élimine la mise au carré initiale. Ce faisant, nous obtenons la véritable différence moyenne entre les observations et la moyenne, une nouvelle statistique qui se nomme l'*écart-type*. Il s'agit d'une statistique qui est extrêmement importante car elle décrit la différence *typique* entre les observations et la moyenne.

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{N} (X_i - M)^2 \Big/ N - 1} \qquad \text{Formule 3.6}$$

Si on examine la distribution X du Tableau 3.5, $N = 3$, la moyenne est 60 et la différence typique entre la moyenne et les autres résultats (l'écart-type s) est égale à 1. Ainsi, au Tableau 3.5, nous concluons que la différence moyenne entre les notes et la moyenne n'est que de 1 point. Par contre, dans la distribution Y, les scores sont très différents les uns des autres. La variance étant $s^2 = 400$, l'écart-type est la racine carrée de 400, soit $s = 20$. Pour la distribution Y, la différence typique entre les notes des étudiants et la moyenne de la classe est de 20 points. Il est clair que la différence moyenne entre les observations et la moyenne est plus petite pour la distribution des notes X que pour la distribution Y. Autrement dit, la variance à l'examen X est plus petite que la variance à l'examen Y. Mais dans les deux cas, le calcul de l'écart-type donne une valeur qui se situe entre 0 et 100, ce qui correspond aux valeurs véritables de la distribution des notes.

Il faut remarquer que nous avons utilisé le mot « variance », même si l'écart-type est la valeur que nous avons utilisée pour la justifier. Ce détour est permis, car la variance et l'écart-type relèvent essentiellement du même concept : un grand écart-type ne peut provenir que d'une grande variance, et si l'on connaît l'une de ces statistiques, on connaît l'autre.

Quiz rapide 3.8
La variance d'une variable est égale à $s^2 = 1$. Quel sera son écart-type ?
Et si l'écart-type est égal à 2, quel sera sa variance ?

La variance d'un phénomène comme indicateur de son intérêt

Le concept de la variance est central non seulement en statistique, mais pour l'ensemble de l'exercice scientifique aussi bien que dans la vie de tous les jours. Pourquoi ?

Avez-vous déjà vécu dans le désert pendant l'été ? On aura sans doute remarqué qu'on écoute rarement les bulletins de météo. Pourquoi ? Parce qu'on sait que le lendemain sera chaud et sec ! Les bulletins de météo ne sont pas importants dans ce cas puisqu'ils diffusent invariablement la même information jour après jour. Statistiquement, il y a peu (pas ?) de variance dans la température en été dans le désert, et parce qu'il n'y a pas de variance, l'information au sujet de la température perd de son importance.

Dans ce chapitre, nous avons utilisé l'exemple farfelu du nombre de nez que les individus ont. Un article scientifique portant sur le nombre de nez que les humains possèdent a-t-il jamais été publié ? Jamais, n'est-ce pas ? Pourquoi ? Parce qu'il n'y a pas de variance à la variable « nombre de nez que les gens ont ». À l'inverse, pourquoi est-ce qu'on attend avec une certaine appréhension l'affichage des notes aux examens ? On le fait parce qu'il est possible d'obtenir la meilleure note, la pire note ou une note entre ces deux extrêmes. Les notes aux examens ont de la variance. Si on savait que les notes aux examens sont invariablement identiques, on ne se précipiterait pas pour les vérifier. Le principe général est le suivant : plus un phénomène (la température, le nombre de nez, les notes aux examens, etc.) démontre de la variance, plus il est intéressant.

Cela conduit à un paradoxe apparent. Moins un phénomène démontre de la variance, meilleure est la moyenne comme indicateur de chaque observation. Mais plus la moyenne est « bonne » (plus l'écart-type est petit), moins intéressant est le phénomène qu'elle décrit !

AUTRES STATISTIQUES DESCRIPTIVES

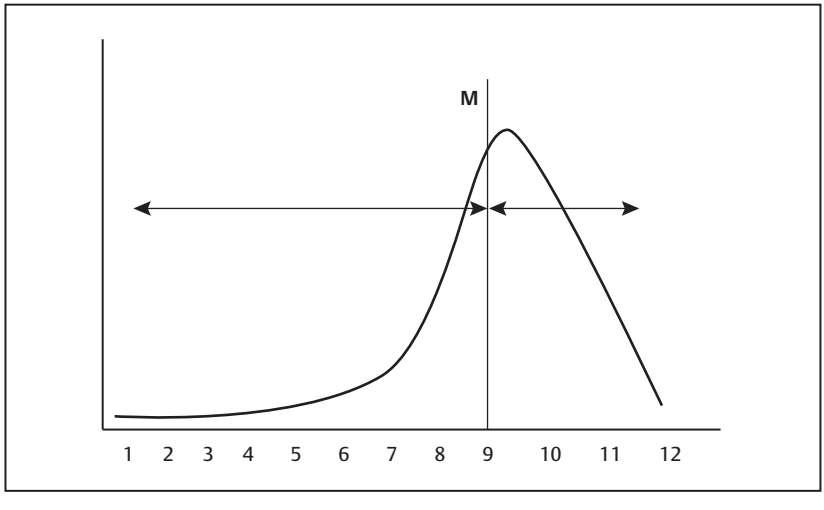
Pour bien décrire une distribution, il faut examiner sa moyenne et sa variance (ou écart-type). Règle générale, ces deux statistiques sont les plus intuitives. Mais il faut aussi prendre en considération la forme de la distribution : son degré d'asymétrie et son degré d'aplatissement (voir le chapitre 2). Commençons par l'asymétrie.

La plupart des tests statistiques comparent les moyennes parce que la moyenne est la valeur unique qui représente le mieux la distribution complète. Mais ces mêmes statistiques présument que la moyenne est une estimation également adéquate pour toutes les valeurs de la distribution, qu'elles se trouvent au-dessus ou en dessous de la moyenne. C'est-à-dire que la plupart des tests statistiques présument que la distribution est symétrique. Lorsque la distribution est asymétrique, le nombre d'observa-

tions se situant des deux cotés de la moyenne et leurs distances relatives à la moyenne ne sont pas égaux. Cela implique que la moyenne est une meilleure estimation des valeurs se situant d'un côté de la moyenne et une moins bonne estimation des valeurs se situant de l'autre coté.

Étudions la Figure 3.3 qui présente une distribution asymétrique (négative). Les valeurs possibles de la variable décrite varient entre 1 et 12, et sa moyenne est de 9. Calculons l'erreur maximale possible lorsque l'on utilise la moyenne comme estimateur de chaque valeur de la distribution. La valeur minimale est de 1, ce qui implique que l'erreur maximale possible est de –8 (1 – 9 = –8) pour les valeurs se situant en dessous de la moyenne. La valeur supérieure maximale est de 12. L'erreur maximale possible pour les valeurs situées au-dessus de la moyenne est alors de +3 (12 – 9 = +3). La distribution n'étant pas symétrique, la moyenne de cette distribution ne fait pas une surestimation ou une sous-estimation égale des valeurs de la distribution, ce qui viole une des présomptions de la plupart des tests statistiques.

FIGURE 3.3 Moyenne et erreurs possibles pour une distribution asymétrique



Mais les tests statistiques sont suffisamment robustes pour demeurer valides dans un cas d'asymétrie à condition que celle-ci ne soit pas « trop »

exagérée. Il nous faut donc une manière de calculer numériquement le niveau d'asymétrie. La Formule 3.7 établit le degré (aussi bien que le signe) de l'asymétrie (symbolisé par « Sk » en référence au terme anglais *skewness*). Nous présentons la formule, mais les logiciels d'analyses statistiques font les calculs requis automatiquement.

$$\mathrm{Sk}_X = \frac{\sum_i (X_i - M_X)^3}{s_X^3} \times \frac{N}{(N-1)(N-2)} \qquad \text{Formule 3.7}$$

où X_i est la i^e donnée, s_X^3 est l'écart-type à la puissance 3, M_X est la moyenne et N est le nombre d'observations dans la distribution.

Le résultat nous indique la direction de l'asymétrie. Il existe trois cas possibles : une asymétrie positive, négative ou nulle.

- Si $\mathrm{Sk}_X > 0$, sa valeur aura un signe positif, indiquant que l'asymétrie est positive (la distribution s'étale davantage vers les valeurs plus élevées de la variable).
- Si $\mathrm{Sk}_X = 0$, la distribution est parfaitement symétrique (les valeurs s'étalent uniformément et également vers les valeurs plus élevées et les valeurs plus faibles de la variable).
- Si $\mathrm{Sk}_X < 0$, sa valeur aura un signe négatif, indiquant que l'asymétrie est négative (la distribution s'étale davantage vers les valeurs plus faibles de la variable).

La Formule 3.7 produira un chiffre allant de zéro (lorsque la distribution est parfaitement symétrique), à un chiffre plus grand (positif) ou plus petit que zéro (négatif) lorsqu'elle est asymétrique. Certains tests statistiques peuvent être utilisés afin de déterminer si la valeur de l'asymétrie est plus ou moins éloignée de zéro. Lorsque ces tests révèlent un niveau trop fort d'asymétrie, certaines corrections mathématiques doivent être mises en œuvre afin de maintenir la validité des conclusions.

Quiz rapide 3.9
Revenez à la distribution des salaires des joueurs de la LNH (chapitre 2). Quel signe prendra la statistique de l'asymétrie : positif, négatif ou nul ?

Le degré d'aplatissement

Le degré d'aplatissement d'une distribution est une autre statistique utilisée pour la décrire. L'aplatissement réfère au degré de concentration près de la moyenne des valeurs de la distribution versus leur étalement plus loin de la moyenne. Comme pour l'asymétrie, la plupart des tests statistiques présument que le degré d'aplatissement d'une distribution n'est ni trop fort ni trop faible et, comme pour l'asymétrie, la violation de cette présomption déclenchera des procédures mathématiques additionnelles qui sont expliquées dans des textes plus avancés.

La description d'une distribution devra donc inclure, minimalement, une estimation du degré d'aplatissement (Ku, de l'anglais *Kurtosis*), ce qui se calcule avec la formule (carrément folle) qui suit :

$$\mathrm{Ku}_X = \frac{\sum_i (X_i - M)^4}{s_X^{\,4}} \times \frac{N(N+1)}{(N-1)(N-2)(N-3)} - 3\frac{(N-1)}{(N-2)(N-3)} \quad \text{Formule 3.8}$$

Cette statistique Ku nous indique le coefficient d'aplatissement d'une distribution.

Dans la pratique, et comme pour le calcul de l'asymétrie, nous confions le calcul du degré d'aplatissement aux ordinateurs. Un degré d'aplatissement de 0 indique une distribution ayant une rondeur typique (qu'on appelle mésocurtique). Si Ku_X est supérieur à 0, cela indique une distribution plus pointue (qu'on appelle leptocurtique) : les valeurs sont plus fortement concentrées autour de la valeur de la tendance centrale. Dans le cas contraire, qu'on appelle platycurtique, la valeur prend un signe négatif indiquant que les extrémités de la distribution sont plus épaisses, contenant plus d'observations qu'à la normale.

Comme pour la statistique de l'asymétrie, le degré d'aplatissement nécessite, lorsqu'il atteint une taille suffisante, des procédures correctives qui sont expliquées dans des textes plus avancés[4].

4. Il faudra se référer à des textes plus avancés tels que celui de Tabachnick et Fidell (2007) pour interpréter et corriger les problèmes occasionnés par l'asymétrie ou la curtose.

Quiz rapide 3.10
Certaines distributions, que l'on nomme des distributions « carrées », ont des effectifs de la même taille pour chaque valeur de la variable. Quel serait le signe de l'asymétrie et du degré d'aplatissement pour ces distributions : positif, négatif ou nul ?

Le coefficient de variabilité

Il est souvent très utile de déterminer si deux variables détiennent des niveaux de variabilité équivalents ou différents. Prenons un exemple hypothétique : on a une distribution qui décrit la taille en mètre d'un groupe de femmes. On construit alors une deuxième distribution qui décrit elle aussi la taille de ce même groupe de femmes, mais cette fois en centimètres. On sait que 1 m = 100 cm. En calculant la moyenne et l'écart-type pour ces deux distributions, on obtient les résultats suivants : $M_{mètre} = 1,6$; $s_{mètre} = 0,20$; $M_{centimètre} = 160,0$; $s_{centimètre} = 20,0$. Comparons maintenant les deux écarts-types afin de déterminer si la taille mesurée en mètre varie plus que la taille mesurée en centimètres. (Bien sûr cela n'est pas le cas mais c'est un jeu.) À l'examen des chiffres décrivant les écarts-types, on pourrait conclure que la taille mesurée en centimètres a plus de variabilité (s = 20) que la taille mesurée en mètre (s = 0,20), ce qui n'est pas très sensé. Bien entendu, la différence entre ces deux écarts types provient du simple fait que les échelles de mesure (mètre et centimètres) ne sont pas numériquement les mêmes. Mais comment pouvons-nous « prouver » que cela est le cas ? Une statistique utile pour y parvenir se nomme *le coefficient de variabilité*. Ce coefficient se trouve en calculant le rapport entre l'écart-type et la moyenne. La Formule 3.9 suit :

$$CV = s/M \qquad \text{Formule 3.9}$$

Après avoir calculé le coefficient de variabilité, on peut maintenant comparer les deux résultats. Pour les résultats exprimés en mètre, le coefficient de variabilité devient : 0,2/1,6 = 0,125. Pour les résultats exprimés en centimètres, le coefficient de variabilité devient : 20/160 = 0,125. Les coefficients de variabilité étant identiques, nous pouvons maintenant faire la preuve que les deux mesures (mètre et centimètres) ont le même niveau de variabilité.

Quiz rapide 3.11
La moyenne de l'examen A est M = 50 et sa variance est de 100. Quel est son coefficient de variabilité ? Pour l'examen B, la moyenne est M = 60 et son écart-type est s = 15. La variabilité des résultats pour l'examen A est-elle plus grande, moins grande ou égale à celle des résultats pour l'examen B ?

Ainsi, le coefficient de variabilité nous permet de déterminer si des variables différentes ont le même degré de variabilité, et si la moyenne de chacune des variables est une représentation également ou inégalement adéquate de chacune des distributions.

SOMMAIRE DU CHAPITRE

Dans ce chapitre, nous avons étudié trois mesures de la tendance centrale d'une distribution. Le mode est la valeur de la distribution qui est la plus fréquente. La médiane est la valeur de la distribution qui la coupe en deux parties, chacune comprenant un nombre égal d'observations. La moyenne est le point d'équilibre d'une distribution. Comparativement aux deux autres mesures de tendance centrale, la moyenne est la seule qui utilise toute l'information de la distribution et elle est celle qui estime toutes les valeurs de la distribution en faisant le moins d'erreurs. La moyenne est généralement la statistique la plus utile et la plus pratique pour tirer des conclusions au sujet d'une distribution. La moyenne permet aussi de saisir le concept d'asymétrie (positive et négative). Nous avons aussi examiné les statistiques qui mesurent le degré de différence entre les observations. L'étendue et l'étendue interquartile sont deux mesures qui fournissent une information grossière de ces différences. Le calcul des erreurs à la moyenne est à la base de statistiques beaucoup plus convaincantes : la variance et l'écart-type. La variance indique dans quelle mesure la moyenne est une bonne ou une mauvaise estimation de chacune des valeurs d'une distribution. L'écart-type fait la même chose, sauf que les valeurs numériques qu'il prend sont plus faciles à interpréter. Les distributions peuvent aussi être décrites par leur degré d'asymétrie et par leur degré d'aplatissement et, enfin, par leur coefficient de variabilité.

EXERCICES DE COMPRÉHENSION

1. La moyenne d'un échantillon est :
   a) une statistique descriptive
   b) une statistique d'inférence
   c) une constante
   d) une variable

2. Parmi les expressions suivantes, laquelle présente une statistique correctement utilisée ?
   a) Sur le plan des statistiques des ventes, le mois dernier, nous avons vendu un manteau de 1 000 $.
   b) À l'Université de Montréal, Jeanne n'est qu'une statistique de plus.
   c) Le coût moyen pour instruire un étudiant ou une étudiante est de 10 000 $.
   d) Toutes ces réponses.

3. Nous calculons la moyenne pour une distribution très asymétrique négative :
   a) La moyenne sera numériquement plus faible que la médiane.
   b) La moyenne n'est pas la meilleure estimation de la tendance centrale dans ce cas.
   c) Dans ce cas, la moyenne n'est pas calculable.
   d) La somme des écarts étant zéro, la moyenne sera égale à zéro.

4. Nous calculons l'écart qui existe entre chaque observation et la moyenne de la distribution à laquelle l'observation appartient. Nous faisons la somme de tous les écarts dans la distribution. Cette somme sera _____.
   a) positive (plus grande que zéro)
   b) négative (moins grande que zéro)
   c) zéro
   d) positive, négative ou zéro, selon l'asymétrie de la distribution.

5. La distribution X est symétrique alors que la distribution Y est asymétrique. Nous désirons couper chaque distribution en deux groupes égaux. La mesure de tendance centrale qui est appropriée pour la distribution X est _____, alors que pour la distribution Y, il s'agirait _____.
   a) la moyenne ; du mode
   b) la médiane ; de la médiane
   c) la moyenne ; du mode ou de la médiane
   d) la médiane ; du mode si l'asymétrie est positive, de la médiane si l'asymétrie est négative

6. La distribution est très leptocurtique. Par conséquent, la moyenne de cette distribution sera _____.
   a) une très mauvaise estimation de chaque valeur de la distribution
   b) une très bonne estimation de chaque valeur de la distribution
   c) adéquate, mais la médiane sera une meilleure estimation
   d) impossible à déterminer

7. La distribution est très platycurtique. Par conséquent, la moyenne de cette distribution sera _____.
   a) une très mauvaise estimation de chaque valeur de la distribution
   b) une très bonne estimation de chaque valeur de la distribution
   c) adéquate, mais la médiane sera une meilleure estimation
   d) impossible à déterminer

8. La différence « typique » qui existe entre chaque observation et la moyenne de la distribution s'appelle :
   a) l'étendue
   b) la variance
   c) l'écart-type
   d) l'écart

9. La variable X a une moyenne de 10 et un écart-type de 2. Pour la variable Y, la moyenne est de 100 et l'écart-type est de 20. Laquelle de ces deux variables démontre le plus de variabilité ?
   a) La variable X.
   b) La variable Y.
   c) Les deux sont égales.
   d) Impossible à déterminer, vu les informations fournies.

Réponses

1. a
2. c
3. a
4. c
5. b
6. b
7. a
8. c
9. c

Page laissée blanche

CHAPITRE 4
LA POSITION RELATIVE DES OBSERVATIONS

Page laissée blanche

CHAPITRE 4

LA POSITION RELATIVE DES OBSERVATIONS

Au chapitre 2, nous avons vu comment obtenir une distribution à partir d'un ensemble d'observations. Dans ce chapitre, nous faisons l'inverse : nous allons décrire les observations à partir de la distribution. Ces nouvelles statistiques permettent de situer une observation par rapport à la distribution ou par rapport à la moyenne. Les procédures statistiques qui positionnent les observations sont utiles en particulier dans les cas où il faut faire une interprétation ou prendre une décision au sujet d'une personne.

Supposons que nous voyons une personne qui mesure 2,50 m. Il est évident que cette personne est très grande. Comment sommes-nous arrivés à cette conclusion ? Nous savons que la très grande majorité des gens mesure à peu près 1,70 m. Implicitement, nous avons comparé la taille de cette personne à la taille des autres personnes que nous voyons quotidiennement. Autrement dit, l'interprétation (la personne est « grande ») que nous faisons d'une observation (mesure : 2,50 m) est élaborée à partir d'une *comparaison* avec les autres observations que nous avons à notre disposition. Les statistiques de positionnement que nous allons voir dans ce chapitre nous permettent de faire ces comparaisons.

Ces statistiques de positionnement peuvent jouer un rôle dans la vie de chacun. Par exemple, il est fort probable qu'un étudiant a été admis à l'université parce que son dossier scolaire a été jugé bon, c'est-à-dire meilleur que celui d'un autre étudiant qui, lui, n'a pas été admis. Si on a l'ambition de poursuivre des études supérieures, ou qu'on espère être admis dans une

école de médecine ou dans n'importe quel autre programme contingenté, ou encore qu'on a pour objectif d'obtenir un poste dans une entreprise prestigieuse, il est certain que notre dossier et nos compétences seront analysés et comparés à ceux d'autres personnes qui ont les mêmes aspirations. Les procédures statistiques que nous allons maintenant étudier sont appliquées dans de telles situations.

Une blague

Nous surprenons une conversation entre Jean et Paul, deux entraîneurs d'athlètes.

Jean : Mon athlète a terminé sa course en moins d'une minute.

Paul : Wow !… tu dois être fier de lui.

Jean : Oui, mais c'était une course de 100 mètres !

Paul : Hum… C'est effectivement moins bon.

Jean : Et il est arrivé deuxième.

Paul : Magnifique !

Jean : Mais il n'y avait que deux athlètes dans la course !

Une anecdote

Quelques semaines après avoir commencé sa scolarité de doctorat, un des auteurs de ce livre a subi son premier examen de statistiques avancées. L'examen était composé de cinq questions complexes. Après trois heures d'acharnement, il n'a réussi à répondre qu'à deux d'entre elles. S'attendant au pire, il appréhendait l'affichage des notes et son anxiété allait croissant. À sa stupéfaction, la note affichée à côté de son nom était un providentiel A+ ! Croyant que cette note ne pouvait être que le fruit d'une blague ou d'une erreur, et prenant son courage à deux mains, il prit rendez-vous avec le professeur pour en avoir le cœur net. Le professeur lui confirma sa note en ajoutant qu'il avait obtenu cette excellente note parce que les autres étudiants n'avaient même pas réussi à répondre à une seule question ! Comparativement à celle des autres étudiants, sa performance avait été exceptionnelle.

Souvent, c'est la position d'une observation par rapport aux autres observations (et non pas l'observation prise isolément) qui détermine l'interprétation qui pourra en être faite. Bien qu'il existe plusieurs façons de déterminer la position d'une observation par rapport aux autres, le *rang*, le *rang percentile* et les *valeurs étalons* sont celles qui sont le plus souvent utilisées.

- Le rang, une mesure ordinale, indique la position de chaque observation sur une échelle allant de 1 à N, où N indique le nombre total d'observations.

- Le rang percentile indique la position d'une observation en la comparant à la proportion, ou le pourcentage, des observations qui lui sont égales ou inférieures.
- La valeur étalon indique la position de chaque observation par rapport à la moyenne. Est-elle en dessous ou au-dessus de la moyenne ? Est-elle près ou éloignée de la moyenne ?

Le rang percentile et la valeur étalon sont des statistiques très utiles, en particulier lorsqu'il s'agit de faire des comparaisons entre plusieurs personnes sur la même variable (telle personne est-elle plus forte en mathématiques que telle autre ?), ou pour la même personne sur plusieurs variables (X est-elle plus forte en chimie ou en géographie ?).

LE RANG ABSOLU

Le *rang absolu*, ou plus simplement le *rang*, est la position d'une observation par rapport aux autres observations de la distribution. Le rang donne la position par rapport au meilleur (ou au pire), mais comme dans la conversation entre les deux entraîneurs de la blague ci-dessus, si on n'indique pas combien il y a d'observations dans la liste, l'information obtenue peut être trompeuse.

Comment calculer le rang

Il faut trier les observations (par ordre croissant ou par ordre décroissant) puis les numéroter de 1 jusqu'à N. Le numéro assigné à chaque observation est le rang. En général, nous attribuons le rang 1 à la « meilleure » performance, mais, selon le problème, nous pouvons choisir d'attribuer le rang 1 à la valeur la plus petite ou encore à la valeur la plus grande. Pour les compétitions de vitesse, nous attribuons le rang 1 (et la médaille d'or) à l'athlète qui prend *le moins* de temps pour terminer l'épreuve. Dans ce cas, c'est la valeur la plus petite de la variable *temps* qui occupe le rang 1 et, par conséquent, l'athlète qui prend le plus de temps pour terminer sa course obtient le dernier rang. Mais, en athlétisme, le rang 1 est attribué à celui ou celle qui a obtenu *le plus* de points et le dernier rang est attribué à l'athlète en ayant reçu le moins. Il y a autant de rangs qu'il y a d'observations dans une

distribution et, en principe, on attribue un rang différent à chaque observation. Ainsi, en présence de 100 observations, la dernière obtient le rang 100 et la première, le rang 1.

Un cas particulier se produit lorsque plusieurs observations sont identiques. On ne peut pas attribuer des rangs différents à ces observations puisqu'elles sont identiques. Dans ce cas, il faut attribuer le rang mitoyen à ces observations. Le Tableau 4.1 présente les notes obtenues à un examen par 32 étudiants. Dans ce cas, nous avons choisi d'attribuer le rang 1 à la note la plus faible (29) et le dernier rang (32, puisque nous avons un total de 32 étudiants) à la personne ayant obtenu la meilleure note (90)[1]. On remarque que les personnes 5 et 6 obtiennent la même note (49) à l'examen. Puisque ces deux étudiants obtiennent la même note, ils doivent détenir le même rang absolu. Dans ce cas, on attribue le rang mitoyen à chacune de ces deux valeurs, soit le rang 5,5 : (5 + 6)/2 = 5,5. Le Tableau 4.1 présente les rangs de tous les étudiants à l'examen.

Quiz rapide 4.1
Nous connaissons le salaire des 679 joueurs de hockey. M. X touche 11 millions et il est le joueur de hockey le mieux payé. En présumant que tous les salaires sont différents, quel sera le rang associé au salaire de cet athlète (le rang 1 est celui du joueur ayant le plus bas salaire) ? Des gens d'affaires américains ont des salaires supérieurs à 11 000 000 $. Si nous construisons une distribution qui comprend le salaire des joueurs de hockey ainsi que celui des gens d'affaires, est-ce que le rang attribué à M. X changera ?

Critique du rang comme mesure de la position

L'utilisation du rang absolu comme mesure de positionnement est lacunaire pour deux raisons. D'une part, on doit indiquer la taille de l'échantillon pour que cette information soit signifiante. Une deuxième place sur 100 n'est pas une deuxième place sur 2 ! D'autre part, en calculant le rang absolu, on a traduit une variable mesurée avec une échelle à intervalles ou une échelle de rapport en variable mesurée sur une échelle ordinale. En

1. Nous choisissons le rang 1 pour la note la plus faible ici, car il s'agit d'identifier les étudiants qui ont le plus besoin d'aide. Dans ce cas, le rang détermine l'ordre de priorité.

convertissant en variable ordinale des variables à intervalles ou de rapport, on sacrifie beaucoup d'informations. Avec les variables ordinales, comme nous l'avons vu au chapitre 1, il est impossible de savoir si les rangs attribués à deux personnes reflètent une grande ou une petite différence entre la performance de ces deux personnes.

Le calcul du rang percentile permettra de résoudre partiellement ces problèmes.

LE RANG PERCENTILE

Les *rangs percentiles* (ou plus simplement les *percentiles*) font partie des statistiques les plus utilisées lorsqu'il s'agit de rapporter des résultats obtenus à un test standardisé, comme les mesures d'intelligence ou d'aptitude. Le percentile situe une valeur par rapport à toutes les autres valeurs. Il indique la proportion (ou le pourcentage) des observations qui sont égales ou inférieures à chaque valeur d'une distribution. Par exemple, si quelqu'un obtient 70 % à un examen et que cette note se situe au 50^{e} percentile, cela indique que 50 % des étudiants ont obtenu une note égale ou inférieure à la sienne et que 50 % des notes à l'examen lui sont supérieures. Si sa note se situe au 99^{e} percentile, 99 % de la classe a obtenu une note égale ou inférieure à la sienne et seulement 1 % des étudiants ont obtenu une note qui lui est supérieure.

Comment calculer le rang percentile

Formellement, le rang percentile d'une valeur se définit par le pourcentage de personnes qui tombent sous cette valeur, plus la moitié du pourcentage de personnes qui tombent exactement sur cette valeur. Pour construire un tableau des percentiles, quatre étapes sont nécessaires. Les trois premières étapes consistent à créer des distributions (de fréquences ou de pourcentages) cumulatives comme nous l'avons vu au chapitre 2. La quatrième étape consiste à effectuer une correction arithmétique. Le Tableau 4.1 présente la distribution des notes obtenues à un examen par les étudiants et le rang absolu aussi bien que le rang percentile associés à chaque note.

Tableau 4.1
Rang absolu et rang percentile pour les notes à un examen

Notes à un examen	*Fréquence*	*Rang absolu*	*Pourcentage*	*Pourcentage cumulatif*	*Rang percentile*
29	1	1	3,1 %	3,1 %	0 + (0,5 × 3,1) = 2
30	1	2	3,1 %	6,3 %	3,1 + (0,5 × 3,1) = 5
35	1	3	3,1 %	9,4 %	6,3 + (0,5 × 3,1) = 8
46	1	4	3,1 %	12,5 %	9,4 + (0,5 × 3,1) = 11
49	2	$\frac{(5+6)}{2} = 5{,}5$	6,3 %	18,8 %	12,5 + (0,5 × 6,3) = 16
50	1	7	3,1 %	21,9 %	18,8 + (0,5 × 3,1) = 20
52	1	8	3,1 %	25 %	21,9 + (0,5 × 3,1) = 23
55	1	9	3,1 %	28,1 %	25 + (0,5 × 3,1) = 27
56	1	10	3,1 %	31,3 %	28,1 + (0,5 × 3,1) = 30
59	1	11	3,1 %	34,4 %	31,3 + (0,5 × 3,1) = 33
61	1	12	3,1 %	37,5 %	34,4 + (0,5 × 3,1) = 36
62	1	13	3,1 %	40,6 %	37,5 + (0,5 × 3,1) = 39
63	1	14	3,1 %	43,8 %	40,6 + (0,5 × 3,1) = 42
64	1	15	3,1 %	46,9 %	43,8 + (0,5 × 3,1) = 45
65	2	$\frac{(16+17)}{2} = 16{,}5$	6,3 %	53,1 %	46,9 + (0,5 × 6,3) = 50
67	1	18	3,1 %	56,3 %	53,1 + (0,5 × 3,1) = 55
70	1	19	3,1 %	59,4 %	56,3 + (0,5 × 3,1) = 58
71	2	$\frac{(20+21)}{2} = 20{,}5$	6,3 %	65,6 %	59,4 + (0,5 × 6,3) = 63
72	1	22	3,1 %	68,8 %	65,6 + (0,5 × 3,1) = 67
74	1	23	3,1 %	71,9 %	68,8 + (0,5 × 3,1) = 70
75	3	$\frac{(24+25+26)}{2} = 25$	9,4 %	81,3 %	71,9 + (0,5 × 9,4) = 77
76	2	$\frac{(27+28)}{2} = 27{,}5$	6,3 %	87,5 %	81,3 + (0,5 × 6,3) = 84
77	1	29	3,1 %	90,6 %	87,5 + (0,5 × 3,1) = 89
78	1	30	3,1 %	93,8 %	90,6 + (0,5 × 3,1) = 92
87	1	31	3,1 %	96,9 %	93,8 + (0,5 × 3,1) = 95
90	1	32	3,1 %	100 %	96,9 + (0,5 × 3,1) = 98

1. On compile la fréquence des notes. Au Tableau 4.1, colonne 2, on voit qu'une personne a obtenu 29 à l'examen, qu'une autre a obtenu une note de 30, que deux étudiants ont obtenu 49, etc.
2. On convertit la distribution des effectifs en pourcentage (quatrième colonne). En convertissant chacune de ces fréquences en pourcentage, on voit qu'un étudiant représente 3,1 % des observations (on a un total de 32 étudiants et 1/32 = 0,03125, ou 3,1 %). Par exemple, les notes de 29 et de 30 sont obtenues respectivement par 3,1 % des étudiants alors que 2 étudiants (2/32 = 0,0625, ou 6,3 % du total) ont obtenu 49.
3. On cumule les pourcentages pour obtenir le pourcentage cumulatif (cinquième colonne). Ainsi, on voit que 3,1 % des étudiants ont obtenu 29 et que 6,3 % des étudiants ont obtenu 30 ou moins à leur examen, et que 18,8 % des étudiants ont une note de 49 ou moins. À la dernière ligne du Tableau 4.1, on voit que 100 % des étudiants ont obtenu 90 ou moins à l'examen.
4. On applique maintenant la correction arithmétique qui produit le rang percentile final pour chaque observation. Le rang percentile se définit comme le pourcentage cumulatif des observations se situant sous chaque valeur, plus *la moitié* du pourcentage des observations qui se situent exactement à cette valeur. La Formule 4.1 décrit la procédure et la sixième colonne du Tableau 4.1 indique le résultat des calculs.

Rang percentile de X = pourcentage cumulatif inférieur à $X + 1/2 \times$ pourcentage à X Formule 4.1

Le dernier terme de la Formule 4.1 est la correction arithmétique qu'il est nécessaire de faire pour estimer le pourcentage de personnes se situant à la valeur X ou en dessous. Puisque les observations se situent exactement à cette valeur, on suppose que si la mesure avait été plus précise (quelques décimales de plus), la moitié des observations auraient obtenu un score légèrement supérieur, et l'autre moitié, un score légèrement inférieur.

Illustration du rang percentile

La dernière colonne du Tableau 4.1 donne le rang percentile pour chaque valeur de la distribution. Établissons le percentile associé à la note la plus faible de la distribution (29). Nous voyons que 3,1 % des étudiants ont obtenu 29, et qu'aucun n'a obtenu de note plus basse. En nous servant de la Formule 4.1, nous pouvons calculer le rang percentile pour la note de 29. Puisque nous n'avons aucune valeur inférieure à 29, le pourcentage des valeurs sous 29 est égal à 0. Mais 3,1 % des étudiants obtiennent une note de 29. Le rang percentile tel qu'il est défini par la Formule 4.1 devient alors $0 + (0{,}5 \times 3{,}1) = 1{,}55$ %. Le rang percentile est donc 1,55. En arrondissant, le rang percentile sera approximativement 2. L'étudiant qui a eu 29 à son examen obtient une note égale ou supérieure à seulement 2 % des étudiants. Par conséquent, 98 % des étudiants ont obtenu une note supérieure à 29.

Pour la note de 30, nous additionnons la quantité 3,1 % (le pourcentage cumulatif d'observations se situant en bas de 30) plus la moitié de 3,1 % (la moitié du pourcentage de personnes ayant obtenu 30 à l'examen) : 3,1 % + 0,5 (3,1 %) = 4,56 %, que nous arrondissons au rang percentile 5. Nous répétons cette opération pour chacune des valeurs de la banque de données. Nous constatons d'abord que 31 étudiants sur 32 (96,9 %) ont obtenu à l'examen une note inférieure à la note la plus forte, soit 90. Un seul étudiant (3,1 % du total) a obtenu la meilleure note (90). Nous appliquons la formule pour trouver le rang percentile de la note 90 : 96,9 % + 0,5 (3,1 %) = 96,9 % + 1,55 % = 98,45, que nous arrondissons à 98. L'étudiant qui a mérité la note de 90 à l'examen a obtenu une note égale ou supérieure à celle obtenue par environ 98 % des étudiants du cours.

Le rang percentile maximal est moins grand que 100. Cela vient du fait qu'il est logiquement impossible qu'une observation d'une distribution soit supérieure à 100 % des observations (cela voudrait dire que cette note est plus forte qu'elle-même !). Par ailleurs, on remarque que le calcul du rang percentile donne des valeurs approximatives : la description qu'il fait de la position d'une observation inclut un certain niveau d'imprécision. Nous y reviendrons plus tard dans ce chapitre. Lorsque les percentiles sont construits sur un grand échantillon, ces imprécisions deviennent négligeables.

Trois autres façons d'estimer les rangs percentiles

Il existe un raccourci pour calculer le rang percentile qui consiste à enlever 0,5 au rang absolu occupé par l'observation puis à diviser par N :

$$\text{Rang percentile de } X = \frac{\text{Rang } X - 0{,}5}{N} > \times 100\,\% \qquad \text{Formule 4.1b}$$

Les logiciels Excel et SPSS utilisent des méthodes légèrement différentes :

$$\text{(Excel)} \quad \text{Rang percentile de } X = \frac{\text{Rang } X - 1}{N - 1} \times 100\,\% \qquad \text{Formule 4.1c}$$

$$\text{(SPSS)} \quad \text{Rang percentile de } X = \frac{\text{Rang } X}{N + 1} \times 100\,\% \qquad \text{Formule 4.1d}$$

Par exemple, pour la note de 90, le rang absolu est 32. Le percentile devient donc,

selon notre approche, $31{,}5/32 \times 100\,\% = 98{,}4\,\%$

selon Excel, $31/31 \times 100\,\% = 100\,\%$

selon SPSS, $32/33 \times 100\,\% = 96{,}97\,\%$.

L'approche d'Excel est à déconseiller, puisqu'elle donne des rangs percentiles de 100 % ; l'approche de SPSS tend à sous-estimer légèrement le rang percentile.

Quiz rapide 4.2
Calculez avec la Formule 4.1b le rang percentile de la note 65 du Tableau 4.1. Trouvez-vous le même résultat ?

Il est possible de se servir des percentiles pour déterminer à rebours une valeur critère. Par exemple, si nous voulons que 40 % des étudiants aient la mention « échec » et 60 %, la mention « succès », il faut trouver la note dont le rang percentile serait 40. Dans le Tableau 4.1, cette note serait entre 62 et 63 (disons 62,5). Un autre exemple : nous pouvons nous référer au Tableau 4.1 pour déterminer la note qui correspond au 50e percentile. Nous cherchons la ligne qui identifie le 50e percentile. Dans ce cas, la note est 65. Nous pouvons alors affirmer que la moitié des étudiants a obtenu une note égale ou inférieure à 65.

Les rangs percentiles souvent requis sont les 25e, 50e et 75e. On les appelle aussi les quartiles, car ils définissent 4 zones : les scores se situant à chacun des rangs percentiles 25, 50 et 75 ou en dessous, et les autres. Les quartiles servent aussi pour calculer l'étendue interquartile (voir le chapitre 3). Il est parfois également utile de diviser la distribution en 10 zones. Nous appelons chaque zone un décile. Comme avec les quartiles, les déciles définissent les valeurs de la variable associées à 10 %, 20 %, etc., des observations de la distribution.

Par exemple, 50 % des joueurs de hockey ont un salaire égal ou inférieur à quel salaire ? En d'autres termes, quel salaire se situe au 50e centile des salaires des joueurs de la LNH ? Dans la base de données *NHLSalaire2002–2003* du site Internet (www.pum.umontreal.ca/ca/fiches/978-2-7606-2113-8.html), c'est le salaire de 1 000 000 $. Mais on sait aussi que la médiane est la valeur de tendance centrale qui coupe la distribution en deux parts égales : 50 % des observations se situant en dessous ou au-dessus d'elle, ce qui revient à dire que la médiane et le percentile 50 ont exactement la même valeur. Ainsi, le joueur de hockey dont le salaire le situe au 50e centile est au moins aussi bien payé que la moitié des joueurs. En revanche, 50 % de ses collègues sont mieux payés. Le Tableau 4.2 donne les quartiles et les déciles pour les joueurs de la LNH. Ainsi, 10 % des joueurs gagnent 450 000 $ ou moins, 20 % gagnent 550 000 $ ou moins, et 90 % des hockeyeurs gagnent 3 600 000 $ ou moins. Par soustraction, seulement 10 % des joueurs gagnent plus de 3 600 000 $.

Souvent, il est pratique de se servir des percentiles afin de déterminer la valeur originale qui correspond à un percentile donné. Par exemple, nous pourrions faire appel au Tableau 4.1 pour déterminer la note à l'examen qui correspond au 75e percentile. Au Tableau 4.1, le percentile 75 n'existe pas et le percentile le plus proche de la valeur recherchée est 77 qui, lui, correspond à la note de 75 % à l'examen. Nous pouvons alors conclure que la note de 75 % correspond approximativement au percentile 75. Il est aussi possible de faire une interpolation pour calculer une valeur plus précise. Mais sauf dans des situations exceptionnelles, cette précision mathématique n'est pas requise.

Quiz rapide 4.3
À partir des données du Tableau 4.1, quels seraient les quartiles ? Pourquoi les quartiles (quart signifiant « quatre ») n'ont que trois nombres ?

L'utilisation des percentiles pour interpréter des mesures

La plupart des résultats individuels que l'on obtient sur des mesures psychologiques, telles que les tests de personnalité, d'habileté cognitive et d'aptitude, ne sont interprétables que lorsqu'ils sont comparés à un *tableau normatif*. Une fois établis, ces tableaux normatifs sont généralement intégrés aux manuels techniques qui accompagnent les tests normalisés.

Tableau 4.2
Déciles et quartiles pour les salaires des joueurs de la LNH

Déciles		*Quartiles*	
10 %	450 000		
20 %	550 000	25 %	600 000
30 %	667 435		
40 %	800 000		
50 %	1 000 000	50 %	1 000 000
60 %	1 300 000		
70 %	1 800 000	75 %	2 100 000
80 %	2 500 000		
90 %	3 600 000		

Les tableaux normatifs sont des tableaux à double entrée indiquant dans une première colonne chaque score qu'il est possible d'obtenir sur la mesure et, dans une deuxième colonne, le percentile associé à ce score, tel qu'il est établi à partir de grands échantillons.

Chaque performance au test peut maintenant être interprétée en se référant directement au tableau normatif. Si la personne obtient le score X et que celui-ci est associé au 10e percentile dans le tableau normatif, on dit que sa performance la situe au 10e percentile sur la mesure. Cette performance est plutôt faible, car 90 % des gens obtiennent un résultat plus fort. Si la performance X se situe au 90e percentile, on tire la conclusion inverse.

Les tableaux normatifs sont souvent segmentés en fonction du sexe, de l'âge ou d'autres caractéristiques importantes pour la compréhension et l'interprétation d'un résultat individuel. Par exemple, les tests qui mesurent certaines habiletés physiques, telle la force, sont accompagnés de tableaux normatifs séparés selon le sexe, ce qui permet une interprétation plus raisonnable. Un homme capable de soulever 50 kg pourrait se trouver au 50e percentile (il est aussi fort que l'homme médian) alors que cette même performance pourrait situer une femme au 80e percentile (elle détient une force égale ou supérieure à celle de 80 % des femmes).

Critique du rang percentile comme mesure de positionnement

Le rang percentile possède deux avantages. Il s'agit d'une statistique facile à calculer et facilement comprise par les non-spécialistes. Pour cette raison, les tableaux normatifs sont souvent exprimés en percentiles. Par exemple, la taille des bébés et leur poids sont souvent exprimés en rangs percentiles : lorsqu'un parent apprend que son nouveau-né pèse 6 kg et que cela le situe au 99e percentile, il comprend facilement que seulement quelques bébés pèsent plus que son nouveau-né.

Par contre, le rang percentile a aussi des inconvénients. Un inconvénient majeur provient du fait qu'il peut mener à des interprétations trompeuses lorsqu'il est basé sur des distributions qui ne sont pas symétriques ou sur des distributions qui comprennent un faible nombre d'observations. Dans ces cas, il faut interpréter les percentiles avec prudence. Si on étudie le Tableau 4.3, on repère les trois notes suivantes : 74, 75, 76. Objectivement parlant, les performances à l'examen de ces trois étudiants sont très semblables. Pourtant, les percentiles associés à ces notes sont très différents. Avec une note de 74, la performance de cet étudiant le situe au 50e centile : une performance sans grand éclat. Mais s'il avait obtenu seulement 1 ou 2 points de plus, nous aurions conclu (rangs percentiles de 75 et 84 respectivement) que sa performance était très bonne ou même excellente. Une petite différence dans les valeurs brutes peut donc mener à de grandes différences dans les percentiles.

De la même façon, un salaire de joueur de hockey dans la LNH de 400 000 $ diffère peu d'un salaire de 500 000 $. Or, le rang percentile du premier est de 4,5 alors que le rang percentile du second est de 18,3. En fait, presque 15 % des salaires sont agglutinés dans cette zone étroite, ce qui rend l'interprétation plus difficile. Pour les salaires des joueurs de la LNH, cela n'est pas un résultat surprenant puisque cette distribution est très asymétrique.

Ce genre d'asymétrie dans les distributions survient plus fréquemment lorsque les distributions sont construites avec un faible nombre d'observations. Par conséquent, le rang percentile est une statistique qui doit être interprétée avec beaucoup de prudence lorsqu'elle est construite sur de petits échantillons.

Heureusement, la plupart des tests standardisés (tests d'intelligence ou de personnalité, par exemple) sont utilisés avec des tableaux normatifs qui analysent les performances de grands groupes de répondants (souvent des milliers, et rarement moins que des centaines). Par conséquent, les percentiles qui sont associés à ces tableaux peuvent être interprétés sans grand risque de distorsion.

Comme nous venons de le voir, les rangs percentiles peuvent mener à des interprétations douteuses. Cette difficulté est partiellement attribuable au fait que les percentiles ne se servent que d'une portion de l'information provenant de la distribution, soit la fréquence relative des observations. La moyenne et la variance des observations ne sont pas directement prises en considération. En mettant à profit ces informations supplémentaires pour établir la position des observations, les résultats obtenus seront beaucoup plus intéressants. Nous nous tournons maintenant vers les valeurs étalons qui permettent de positionner les valeurs, peu importe la forme de la distribution dont elles proviennent.

Tableau 4.3
Les percentiles pour une distribution asymétrique

Notes	*Fréquence*	*Rang absolu*	*Rang percentile*
30	1	1	1,6
50	1	2	4,7
62	1	3	7,8
63	5	4 à 8 : 6	18,8
64	1	9	26,6
70	1	10	29,7
74	12	11 à 22 : 16,5	50,0
75	4	23 à 26 : 24,5	75,0
76	2	27 à 28 : 27,5	84,3
77	1	29	89,1
78	1	30	92,2
87	1	31	95,3
90	1	32	98,4

LA VALEUR ÉTALON Z

Le rang percentile indique la position de n'importe quelle observation par rapport aux autres observations de la distribution. La *valeur étalon* indique aussi la position d'une observation, mais, cette fois, par rapport à la moyenne de la distribution. La *valeur étalon Z* (parfois appelée la *cote Z*) est probablement la valeur de positionnement la plus fréquemment utilisée. Nous l'appelons une valeur *standardisée*.

La valeur étalon Z convertit les valeurs initiales en valeurs standardisées. Ces valeurs standardisées peuvent être négatives, positives ou égales à zéro. Lorsqu'une *valeur étalon Z est positive*, cela indique que *l'observation se trouve au-dessus* de la moyenne. Lorsqu'elle est *négative, l'observation est inférieure à la moyenne* et lorsque la *valeur étalon Z est égale à zéro, l'observation se trouve exactement à la moyenne*. Ainsi, si trois étudiants obtiennent respectivement 60, 70 et 80 à un examen ayant 70 pour moyenne et que nous convertissons ces notes en valeurs étalons Z, la note 60 prendra un signe négatif (sous la moyenne), la note 70 sera de zéro (égale à la moyenne) et la note 80 sera positive (supérieure à la moyenne).

La valeur étalon Z peut varier entre moins l'infini et plus l'infini. Plus une valeur étalon Z est loin de zéro, plus la valeur brute qui lui correspond est distante de la moyenne. Ainsi, si le salaire de Jules, converti en valeur étalon Z, est égal à +2, mais que le salaire de Marie est de +1 (en valeur Z), cela indique non seulement que les deux salaires sont plus élevés que le salaire moyen (parce que les deux valeurs Z sont positives), mais que le salaire de Jules (Z = +2) est, lui, plus élevé que le salaire de Marie (Z = +1).

Comme pour le percentile, la caractéristique la plus importante d'une valeur étalon est qu'elle permet de faire des comparaisons directes entre la performance d'une personne sur plusieurs variables et la performance de plusieurs personnes sur une même variable. Ainsi, si quelqu'un veut savoir s'il est plus fort en mathématiques qu'en chimie, il lui serait possible de convertir ses notes dans ces deux cours en valeurs étalons (par exemple Z). Si le Z pour les mathématiques est égal à +1 et que sa note en chimie est de Z = –1, cela indique : a) qu'il est plus fort en mathématiques qu'en chimie ; et b) qu'il se situe *au-dessus* de la moyenne en mathématiques et *sous* la

moyenne en chimie. La valeur étalon positionne l'information en prenant en considération la moyenne et l'écart-type de la distribution.

Comment calculer la valeur étalon Z

La valeur étalon Z compare chaque observation initiale à la moyenne. Donc, pour convertir une valeur d'une distribution en valeur étalon Z, il nous faut connaître cette valeur aussi bien que la moyenne de la distribution. Le point de départ est l'écart entre la donnée, par exemple X_i, et la moyenne ($X_i - M_X$). L'écart sera plus grand pour les valeurs X_i situées loin de la moyenne et il sera plus petit pour les observations situées plus près de la moyenne. Naturellement, l'écart sera de zéro pour les observations qui se trouvent directement à la moyenne.

Tableau 4.4
Températures en degrés Celsius, en degrés Fahrenheit et en valeurs étalons Z pour un pays fictif

Mois	*X* °*C*	$X - M_X$	Z_X	*Y* °*F*	$Y - M_Y$	Z_Y
Février	0	–15	–1,39	32	–27	–1,39
Mars	5	–10	–0,93	41	–18	–0,93
Avril	10	–5	–0,46	50	–9	–0,46
Mai	15	0	0,00	59	0	0,00
Juin	20	5	0,46	68	9	0,46
Juillet	25	10	0,93	77	18	0,93
Août	30	15	1,39	86	27	1,39
Somme	105	—	0	413	—	0
N	7	—	7	7	—	7
Moyenne	15,0	—	0,0	59,0	—	0,0
Écart-type	10,8	—	1,0	19,4	—	1,0

Le Tableau 4.4 présente les températures en Celsius et en Fahrenheit pour un pays fictif. En février, il fait 0 °C, soit 32 °F. Il s'agit de la même

température (le même degré de chaleur), mais le chiffre la décrivant en degrés Celsius est différent de celui qui la décrit en degrés Fahrenheit, puisque les échelles de mesure en Celsius et en Fahrenheit sont différentes. (Pour convertir les degrés Celsius en degrés Fahrenheit, on multiplie les degrés Celsius par 9/5 et on ajoute 32.) Le Tableau 4.4 présente la température moyenne et l'écart-type des températures pour les sept mois qui sont mentionnés.

Pour déterminer la position de chaque température par rapport à la température moyenne, on calcule d'abord l'écart entre chaque température et sa moyenne ($X - M_X$ pour les degrés Celsius et $Y - M_Y$ pour les degrés Fahrenheit). En observant la taille de ces différences, on voit que la température en février est sous la moyenne [$X_{février} - M_X = 0 - 15 = -15$ et $Y_{février} - M_Y = 32-59 = -27$]. Pouvons-nous conclure que, comparativement à la moyenne, le mois de février est moins froid en degrés Celsius qu'il ne l'est en degrés Fahrenheit ? Bien sûr que non. Il faut corriger ces valeurs afin de prendre en considération les deux échelles de mesure. La correction se fait en divisant l'écart obtenu ($X_i - M_X$; et $Y_i - M_Y$) par l'écart-type de la distribution dont l'observation provient.

La valeur étalon se construit en exprimant la différence observée (entre chaque observation et la moyenne) par rapport à la différence typique, que nous connaissons comme étant l'écart-type (voir le chapitre 3). La distance entre une observation et la moyenne est-elle plus grande, moins grande ou aussi grande que l'écart-type ? La Formule 4.2 décrit le calcul de la valeur étalon Z :

$$Z_X = \frac{(X_i - M_x)}{s_x} \qquad \text{Formule 4.2}$$

où $(X_i - M_X)$ est l'écart entre une observation et la moyenne, et s_X est l'écart-type de la distribution X.

Calculons la valeur étalon Z pour la température du mois d'avril en degrés Celsius (qui est $X_{avril} = 10$ au Tableau 4.4). Nous connaissons la température moyenne ($M_{Celsius} = 15$) et son écart-type ($s_{Celsius} = 10{,}8$). Nous entrons les chiffres dans la Formule 4.2.

$$Z_{avril} = \frac{(10 - 15)}{10{,}8} = -5 / 10{,}8 = -0{,}46.$$

La température du mois d'avril, exprimée sous une forme standardisée (valeur étalon Z), est donc Z = –0,46. Puisque la valeur Z est négative, nous savons que la température en avril se situe au-dessous de la température moyenne pour tous les mois de l'année.

Le Tableau 4.4 indique la température en valeurs étalons Z. On peut remarquer que lorsqu'elles sont traduites en valeurs étalons Z, les températures en degrés Celsius et en degrés Fahrenheit sont identiques. Nous disons que la valeur étalon est standardisée parce que chaque valeur est exprimée par rapport à un dénominateur commun, en l'occurrence l'écart-type.

La valeur étalon Z répond à la question suivante : l'observation X_1 est-elle aussi différente de la moyenne que l'est l'observation X_2 ? Si la réponse est affirmative (les valeurs Z calculées pour X_1 et pour X_2 sont identiques), on conclut alors que les deux observations occupent exactement la même position relative sur les deux échelles de mesure. Cette caractéristique des valeurs étalons Z est particulièrement utile lorsqu'il s'agit de comparer plusieurs performances produites par la même personne.

Supposons qu'un étudiant a obtenu 70 % à l'examen d'anthropologie et 80 % en littérature. Est-il meilleur en littérature qu'en anthropologie ? Convertissons ces deux performances en valeurs étalons Z. La moyenne et l'écart-type de l'examen d'anthropologie sont M = 50 et s = 10. La valeur Z pour cette performance est (70 – 50)/10 = +2. Sa note étant très supérieure à la moyenne, nous concluons que cet étudiant est bon (même très bon) en anthropologie. La moyenne et l'écart-type pour l'examen de littérature sont 65 et s = 15. Encore une fois, nous calculons la valeur étalon Z = (70 – 65)/15 = +0,33. Sa note en littérature est plutôt proche de la moyenne en littérature (Z = +0,33). La note en anthropologie étant beaucoup plus forte que la note moyenne de sa classe et la note en littérature, étant proche de la moyenne, nous pouvons alors conclure que l'étudiant est plus fort en anthropologie qu'il ne l'est en littérature.

Nous pouvons aussi faire appel aux valeurs étalons pour comparer deux observations sur la même variable. Revenons à l'examen d'anthropologie. L'étudiant en question a donc obtenu 70 % et une amie à lui a obtenu 80%. Si nous standardisons les deux performances (M = 50, s = 10) et calculons leur valeur étalon Z, nous voyons que la note de l'amie se situe à Z = +3, alors que la note de l'étudiant se situe à Z = +2. Tous les deux ont obtenu

des notes au-dessus de la moyenne (les Z obtiennent des signes positifs), mais l'amie a mieux réussi parce que sa performance se situe plus loin de la moyenne que celle de l'étudiant.

Au Tableau 4.4, la température exprimée en valeur étalon Z est identique, qu'elle soit originellement mesurée en Fahrenheit ou en Celsius. La température du mois de février, exprimée en Celsius ou en Fahrenheit, est exactement à la même distance de la température moyenne (le score Z est –1,39 dans les deux cas). Que l'on pense Celsius ou Fahrenheit, lorsqu'il fait froid, il fait froid !

Deux particularités des valeurs étalons Z

Deux particularités des valeurs étalons Z rendent cette statistique fort utile.

1. La moyenne d'une distribution exprimée en valeur étalon Z est toujours égale à 0.

$$M_Z = \frac{\sum Z_i}{N_z} = 0 \qquad \text{Formule 4.3}$$

Prenons une observation qui se trouve exactement à la moyenne de sa distribution. Calculons la valeur étalon Z pour cette observation. La moyenne de la distribution et la valeur de l'observation étant identiques, la différence entre les deux (X= -M) est égale à zéro. Puisqu'elle est zéro, la valeur étalon Z qui lui correspond devient, elle aussi, zéro. La moyenne étant le point d'équilibre d'une distribution, il y aura autant de valeurs sous la moyenne que de valeurs au-dessus de la moyenne, si bien qu'elles s'annulent. Exprimés en scores Z, les négatifs et les positifs s'annulent, produisant un Z moyen de zéro.

Quiz rapide 4.4
Calculez la moyenne des valeurs étalons Z pour la distribution de températures en degrés Fahrenheit et pour la distribution en degrés Celsius. Les moyennes diffèrent-elles ?

2. L'écart-type d'une distribution exprimée en valeurs étalons Z est toujours égal à 1.

Calculons l'écart-type des valeurs étalons Z pour les températures mesurées en Fahrenheit. Faisons appel à la formule habituelle pour le calcul de l'écart-type. Comme pour n'importe quelle variable, l'écart-type des valeurs étalons Z représente la différence moyenne entre les observations et la moyenne de leur distribution. Or, la moyenne des valeurs étalons M_Z est invariablement égale à 0. Il est donc inutile de soustraire M_Z, et la formule devient :

$$s_z = \sqrt{\frac{\sum_{i=1}^{N} (Z_i - M_z)^2}{N-1}} = \sqrt{\frac{\sum_{i=1}^{N} Z_i^2}{N-1}} \qquad \text{Formule 4.4}$$

Il n'y a donc qu'à mettre au carré chaque valeur Z, en faire la somme, puis diviser par $N - 1$. Finalement, nous calculons alors la racine carrée du résultat pour obtenir l'écart-type. Pour les données du Tableau 4.4, ceci donne :

$$\begin{aligned} s_z &= \sqrt{\frac{(-1{,}39)^2 + (-0{,}93)^2 + (-0{,}46)^2 + (0)^2 + (0{,}46)^2 + (0{,}93)^2 + (1{,}39)^2}{(7-1)}} \\ &= \sqrt{\frac{1{,}932 + 0{,}864 + 0{,}212 + 0{,}212 + 0{,}864 + 1{,}932}{6}} \\ &= \sqrt{\frac{6}{6}} = \sqrt{1} = 1 \end{aligned}$$

Quiz rapide 4.5
Supposons une distribution X où M = 100 et s = 20, et une distribution Y où M = 100 et s = 10. Les deux variables détiennent-elles le même niveau de variabilité ? Transformons chaque valeur de chaque distribution en valeur étalon Z et calculons le coefficient de variabilité pour chacune des deux distributions X et Y. Le coefficient de variabilité sera-t-il le même ou différent pour les distributions X et Y lorsque ces dernières seront exprimées en valeurs étalons Z ?

Nous voyons maintenant pourquoi la valeur étalon Z est une valeur si populaire : tous les échantillons, lorsqu'ils sont exprimés en valeurs étalons Z, détiennent la même moyenne (0) et le même écart-type (1). Grâce à cette transformation, les valeurs obtenues par une personne sur n'importe quelles variables sont directement comparables à condition qu'elles

soient toutes standardisées. Si cette personne obtient des valeurs étalons Z de 0 sur deux variables, elle se situe à la moyenne des deux variables. Si elle obtient –1 sur la variable X et +1 sur la variable Y, elle se situe à un écart-type sous la moyenne sur X et à un écart-type au-dessus de la moyenne sur Y.

Comment ramener une valeur étalon à sa valeur initiale brute

À partir d'une valeur étalon Z_i, il est possible de trouver sa valeur brute X_i, si on connaît la moyenne M_X et l'écart-type s_X. Il s'agit d'une simple transformation algébrique de la Formule 4.2 qui devient la Formule 4.5. Isolons X_i à partir de la formule de calcul de la cote Z :

$$Z_i = (X_i - M_X) / s_X \qquad \text{Formule 4.2}$$

$$X_i = (Z_i \times s_X) + M_X \qquad \text{Formule 4.5}$$

Au Tableau 4.4, prenons la valeur étalon Z pour le mois d'avril (Z_{avril} = –0,46) et calculons sa température en Celsius. Nous savons que la température moyenne en Celsius est de 15 degrés et que son écart-type est de 10,8, ce qui donne

$$X_{avril} = (-0{,}46 \times 10{,}8) + 15$$
$$= 10$$

AUTRES VALEURS ÉTALONS

Bien que la valeur étalon Z soit très souvent utilisée pour faire des comparaisons, elle souffre d'un inconvénient « politique ». Supposons qu'un psychologue scolaire présente à un parent le résultat du test de QI administré à son enfant. La performance au test de l'enfant le place à la moyenne (qui est 100 pour ce test de QI), et par conséquent l'enfant obtient une cote Z de 0. Si le parent comprenait les statistiques et les tests de QI, il n'y aurait aucun problème à lui dire que son enfant a obtenu une performance de Z = 0, car il comprendrait que son enfant est doté d'un niveau d'intelligence « moyen ».

Mais supposons que le parent est un néophyte en statistique. Si le psychologue lui dit que son enfant a un QI Z = 0, il pourrait croire que son enfant n'est pas intelligent ! Pour cette raison, lorsqu'il s'agit de présenter des résultats aux non-spécialistes, il est préférable de les présenter avec des chiffres qui sont moins susceptibles d'être mal interprétés. Puisque la grande majorité des gens sont habitués aux résultats notés sur 100, il est préférable de présenter les résultats obtenus avec des chiffres qui reflètent cette échelle. La *valeur étalon T* correspond à cette représentation des observations. Cette valeur est utilisée très fréquemment pour positionner les résultats sur les tests de personnalité, et les spécialistes en évaluation psychologique y ont souvent recours.

La valeur étalon T a une moyenne de 50 et un écart-type de 10. Tous les échantillons, lorsqu'ils sont exprimés en valeurs T, ont cette moyenne et cet écart-type. Ce même principe est identique avec la valeur étalon Z ($M_Z = 0$, $s_Z = 1$). Donc, un enfant qui a un QI le situant à la moyenne obtient une performance de T = 50 (et de Z = 0) sur son test de QI. Les gens reconnaissant facilement qu'un résultat de 50 indique un résultat moyen, le parent non statisticien sera moins prompt à faire une interprétation erronée du résultat. Les valeurs T inférieures ou supérieures à 50 sont respectivement en dessous ou au-dessus de la moyenne.

Comment calculer la valeur étalon T

Pour calculer les valeurs étalons T à partir des valeurs brutes, il est plus facile de préalablement convertir ces valeurs brutes en valeurs étalons Z. La Formule 4.6 convertit une valeur étalon Z en valeur étalon T.

$$T = (10 \times Z) + 50 \qquad \text{Formule 4.6}$$

Z est la performance exprimée en valeur étalon Z, 10 est l'écart-type des valeurs T et 50 est la moyenne des valeurs T. Calculons la valeur étalon T pour la température du mois de mars inscrite au Tableau 4.4. La température en Fahrenheit est 41, ce qui se traduit en valeur étalon $Z_{mars} = -0{,}93$.

$$\begin{aligned} T_{mars} &= (10 \times -0{,}93) + 50 \\ &= (-9{,}30 + 50) \\ &= 40{,}7 \end{aligned}$$

Quiz rapide 4.6
Calculez la valeur étalon T pour la température du mois d'avril à partir de sa température en Celsius et en Fahrenheit.

Créer sa propre valeur étalon

Il n'y a rien de magique dans les valeurs étalons T ou Z. En fait, selon la situation, on peut créer sa propre valeur étalon. Dans tous les cas, il s'agit de déterminer la moyenne et l'écart-type désirés. D'abord, on convertit les valeurs réelles en valeurs étalons Z, puis on exprime ces valeurs en fonction de la moyenne et de l'écart-type choisis.

Illustration : un chef d'entreprise en Allemagne veut informer chaque membre du personnel sur son salaire par rapport aux autres salaires offerts par la compagnie. Puisque les salaires sont exprimés en milliers d'euros, il risque des ennuis en indiquant au salarié moyen que son salaire est de 0 (s'il choisit la valeur étalon Z) ou 50 (s'il choisit T). Il faut donc créer une nouvelle valeur étalon nommée *E*. On décide arbitrairement que cette nouvelle statistique a une moyenne de 30 000 et un écart-type de 10 000. Quel est maintenant le salaire de trois employés ? Gerhart a un salaire moyen, Rudolf a un salaire le situant à deux écarts-types au-dessus de la moyenne, et Willie a un salaire à un écart-type sous la moyenne. Exprimons ces trois salaires en valeurs étalons E. D'abord, nous convertissons les salaires en valeurs étalons Z. Pour Gerhart Z = 0 ; pour Rudolf Z = +2, pour Willie Z = –1. Nous calculons maintenant les observations en valeurs étalons E.

Gerhart : $E = 10\,000 \times 0 + 30\,000 = 30\,000$

Rudolf : $E = 10\,000 \times +2 + 30\,000 = 50\,000$

Willie : $E = 10\,000 \times -1 + 30\,000 = 20\,000$

Nous venons d'inventer une nouvelle statistique, la valeur étalon E. On ne la retrouvera dans aucun autre livre de statistiques ; néanmoins, elle est tout aussi valide que les valeurs étalons Z et T.

Quiz rapide 4.7
Quel serait le salaire d'Ingrid en valeur étalon E si son salaire se trouvait à Z = +1,5 ?

Un dernier mot sur les valeurs étalons

La transformation d'une observation en une valeur étalon est une transformation linéaire. Cela implique que le fait de convertir une distribution d'observations en valeurs étalons n'aura *aucun impact* sur la forme de la distribution. Si la distribution originale est asymétrique, la distribution des valeurs Z le sera autant. La transformation des valeurs brutes en valeurs étalons ne peut en aucun cas rendre « normale » (voir le chapitre 5) une distribution qui ne l'est pas. Pour cette raison, lorsque nous travaillons avec une distribution qui est convertie en valeurs étalons Z ou T, on ne doit pas parler de « normalisation », mais de « standardisation ». Il est impossible de se servir du processus de standardisation (Z ou T) pour produire une distribution symétrique à partir d'une distribution qui ne l'est pas.

SOMMAIRE DU CHAPITRE

Le positionnement d'une observation permet de faire des comparaisons entre plusieurs personnes sur la même variable ou la position d'une personne sur plusieurs variables. Le rang, le rang percentile et la valeur étalon sont trois techniques qui permettent de trouver la position d'une observation. Le rang percentile positionne une observation par rapport aux autres observations de la distribution. La valeur étalon (Z, T ou autre) positionne chaque observation par rapport à la moyenne. Le percentile est particulièrement utile lorsqu'il s'agit de présenter les résultats aux personnes qui ne connaissent pas les statistiques. Cependant, dans certaines situations, il peut mener à des interprétations et des conclusions problématiques. La valeur étalon, en particulier la statistique Z, bien qu'un peu plus difficile à calculer et à comprendre, est plus polyvalente. Ces mesures de positionnement sont très souvent utilisées en pratique lorsqu'il s'agit d'évaluer un individu sur le plan psychologique ou pour élaborer un diagnostic dans le domaine de l'éducation.

EXERCICES DE COMPRÉHENSION

1. Pour être admis dans une université, les étudiants doivent passer un examen d'admission. Cette université n'accepte que les étudiants qui se classent parmi les 10 % des étudiants qui ont eu les meilleurs résultats. Jeanne est admise. Nous pouvons alors déduire que son percentile pour l'examen d'admission est _____.
   a) au minimum 10
   b) plus de 10 et moins de 90
   c) au moins 90
   d) plus de 90

2. La meilleure note obtenue à l'examen est 98 et la pire note est 12. Le rang absolu pour la personne qui obtient la meilleure note est _____, alors que le rang absolu détenu par la personne ayant obtenu la pire note est _____.
   a) 1 ; 100
   b) 12 ; 98
   c) 98 ; 12
   d) 1 ; impossible à déterminer

3. Nous avons une distribution décrivant les degrés de pauvreté dans plusieurs villes nord-américaines. Nous voulons positionner chaque ville relativement aux autres sur le plan de la pauvreté. Quelle serait la technique qui sacrifierait le plus d'informations ?
   a) La moyenne de la distribution.
   b) La valeur étalon Z.
   c) Le percentile.
   d) Le rang absolu.

4. Le salaire de Jules est de 25 000 ∈ ; or, la moyenne des salaires en France est de 25 000 ∈ . La distribution des salaires est normale. Compte tenu de ces informations, quelle est la position du salaire de Jules en valeur étalon Z, en percentile, en valeur étalon T ?
   a) Z = 0 ; percentile = 50 ; T = 50.
   b) Z = 0 ; le percentile ne peut pas être déterminé ; T = 25 000.
   c) Z ne peut pas être déterminé ; le percentile ne peut pas être déterminé ; T ne peut pas être déterminé.
   d) Z ne peut pas être déterminé ; percentile = 50 ; T = 0.

5. On calcule les valeurs étalons Z pour toute une distribution de valeurs. On calcule la somme des valeurs étalons Z qui sont positives et la somme des valeurs étalons Z qui sont négatives. On additionne une somme à l'autre et le résultat est _____.

a) 1
b) 0
c) –1
d) impossible à determiner

6. Des 10 étudiantes qui suivent ce cours, 9 obtiennent une note entre 70 et 72, et 1 étudiante, Florence, obtient 90. Nous convertissons chacun de ces résultats en valeur étalon Z. En valeur étalon Z, la performance de Florence est _____.

a) Z négatif proche de zéro
b) Z positif proche de zéro
c) Z négatif loin de zéro
d) Z positif loin de zéro

7. Nous examinons le taux de criminalité dans une trentaine de pays. Le taux de criminalité moyen est de 1 000 crimes/1 million d'habitants. Nous convertissons ces taux de criminalité en valeur étalon Z. Le pays A obtient un taux de criminalité de Z = 0, alors que le pays B obtient un Z = –2. Le taux de criminalité du pays A est _____, alors que le taux de criminalité du pays B est _____.

a) essentiellement zéro ; très fort
b) essentiellement zéro ; très faible
c) de 1 000 crimes/1 million d'habitants ; moins de 1 000 crimes/1 million d'habitants
d) de 1 000 crimes/1 million d'habitants ; de plus de 1 000 crimes/1 million d'habitants

8. Dans une compagnie, chaque mois, le vendeur qui réalise le plus de ventes mérite un voyage à Tahiti. Vous êtes le directeur de cette compagnie et vous devez choisir la personne qui ira à Tahiti. Quelle est la statistique de positionnement la plus appropriée dans ce cas ?

a) Le rang absolu.
b) Le percentile.

c) La valeur étalon Z.
d) Aucun de ces choix.

9. Nous comptabilisons le nombre de livres publiés par les professeurs de 10 grandes universités de recherche au Canada. L'Université de Montréal se situe au deuxième rang de cette distribution, dont nous ignorons la forme. Si nous convertissons toutes les valeurs de cette distribution en valeurs étalons Z et en percentiles, la cote Z de l'Université de Montréal sera _____ et son percentile sera _____.
a) positive ; plus grand que 50
b) négative ; plus petit que 50
c) positive ou négative ; au-dessus ou au-dessous de 50
d) de zéro ; au-dessus ou au-dessous de 50

Réponses

1. c
2. d
3. d (le nombre total d'étudiants n'est pas indiqué)
4. a
5. b
6. d
7. c
8. a
9. c (N. B. L'énoncé n'indique pas si le rang 2 définit un nombre élevé ou faible de livres publiés.)

CHAPITRE 5
LA DISTRIBUTION NORMALE

Page laissée blanche

CHAPITRE 5

LA DISTRIBUTION NORMALE

La distribution normale joue un rôle central en statistiques. D'une part, la forme de cette distribution décrit un grand nombre de caractéristiques physiques, sociologiques et psychologiques. Nous l'appelons « normale » puisque, d'après Quételet (voir le texte ci-dessous), il s'agit de la distribution « habituelle ». D'autre part, la distribution normale est importante parce que nous en savons beaucoup à son sujet. Notre connaissance des caractéristiques de la distribution normale a permis l'élaboration d'un ensemble de tests statistiques sophistiqués (que nous verrons dans les chapitres ultérieurs). La compréhension de la distribution normale et de ses caractéristiques est essentielle pour l'étude des statistiques, en particulier les statistiques qui nous permettent de faire des inférences.

Adolphe Quételet et Carl Friedrich Gauss

Au XIX^e^ siècle, le mathématicien Adolphe Quételet fait une découverte importante : en examinant la distribution des effectifs de la taille des recrues de l'armée française, il remarque que quelques soldats sont très petits et quelques-uns très grands, les autres se situant entre ces deux extrêmes. Le graphique de polygone de la taille produit une courbe en forme de cloche. Mais plus important encore, Quételet remarque que la distribution de la taille des soldats français ressemble comme deux gouttes d'eau à la distribution du tour de poitrine des soldats écossais ! Pourtant, il s'agit de deux mesures différentes (une longueur et une circonférence) et de deux groupes différents (des Français et des Écossais). Quételet ne trouve pas de raisons pouvant raisonnablement expliquer cette coïncidence et en déduit qu'il s'agit de la distribution habituelle à laquelle on pourrait « normalement » s'attendre. Par conséquent, nous donnons le nom de « distribution normale » à cette distribution dont le polygone prend la forme d'une cloche.

C'est le très célèbre mathématicien Carl Friedrich Gauss qui a expliqué pourquoi la distribution normale est si habituelle. En son honneur, nous donnons un second nom à la distribution normale : la distribution gaussienne.

Lorsqu'une population est normale, il est possible de déterminer :

- le rang percentile d'une observation à partir d'une valeur étalon Z et vice-versa ;
- la proportion des observations qui se situent au-dessus ou en dessous d'une valeur ou entre deux valeurs ;
- si le résultat d'une expérience est probable ou improbable (ce qu'on appelle un résultat statistiquement significatif ; voir chapitres 8 et 9).

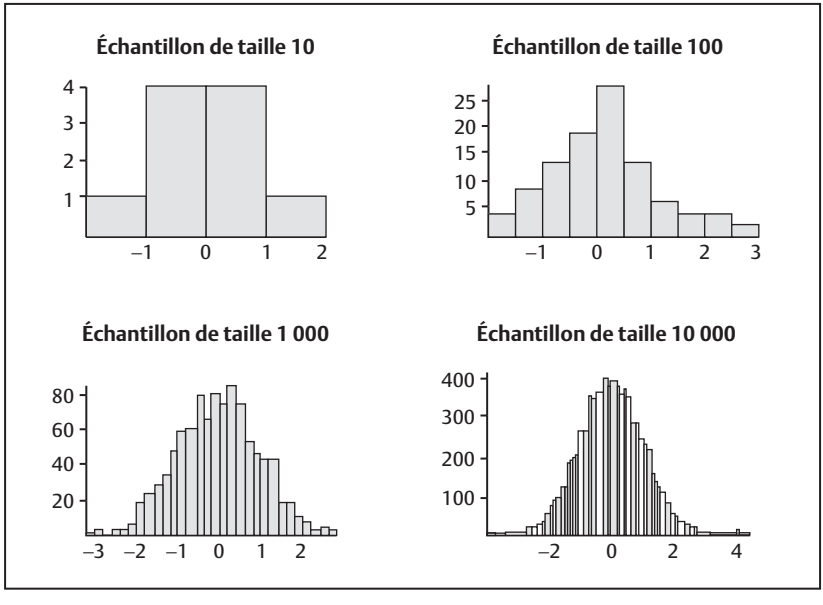
QUELQUES CONSEILS DE PRUDENCE EN GUISE DE PRÉAMBULE

Bien qu'un nombre important de phénomènes soient normalement distribués, tous ne le sont pas. Les temps de réponse et les salaires sont des cas où la distribution n'est pas normale. Lorsque nous travaillons avec des variables qui, clairement, ne sont pas normalement distribuées, les procédures décrites aux chapitres 8 à 12 ne peuvent pas être utilisées. Dans ce cas, il faut préférer les analyses non paramétriques (qui sont traitées au chapitre 13). Heureusement, la normalité est une présomption raisonnable pour la vaste majorité des phénomènes, en particulier ceux que l'on trouve en sciences sociales.

Une distribution parfaitement normale est une conception théorique que nous ne retrouvons dans la nature que lorsque nous analysons des populations entières. Puisqu'il nous est généralement impossible de mesurer une population entière, nous n'analysons, en général, qu'une partie de ces informations, que nous appelons un échantillon[1]. Lorsque l'échantillon est très petit, sa distribution a peu de ressemblance avec la distribution normale. Mais au fur et à mesure que le nombre d'observations augmente, la distribution de l'échantillon ressemble de plus en plus à la distribution parfaitement normale. Un échantillon comprenant plusieurs millions d'observations ne sera pas parfaitement normal, mais il sera plus proche de la normalité qu'un échantillon comprenant des milliers d'observations. Cependant, la ressemblance avec la distribution normale sera dans ces deux cas excellente. La Figure 5.1 présente quatre échantillons comprenant des nombres différents d'observations (N).

1. Les concepts de population et d'échantillon sont approfondis dans les chapitres 8 et 9.

FIGURE 5.1 Exemples d'échantillons de tailles variables tirés d'une population normale



Chacun de ces échantillons est extrait aléatoirement d'une population d'observations qui est normalement distribuée. Dans la Figure 5.1, la courbe en forme de cloche est beaucoup plus clairement identifiable pour les distributions comprenant des effectifs plus grands (N = 1 000, N = 10 000) que celles ayant des effectifs plus petits (N = 10, N = 100). Mais même lorsque le nombre d'observations est très petit (N = 10), nous commençons, néanmoins, à y reconnaître une forme « normale ». Enfin, on peut noter que la différence dans la forme de la courbe entre N = 10 et N = 100 est plus marquée que la différence entre les courbes N = 1 000 et N = 10 000. Lorsque les distributions contiennent déjà beaucoup de données, l'ajout d'observations additionnelles affectera peu la forme de la distribution.

DÉFINITION DE LA DISTRIBUTION NORMALE

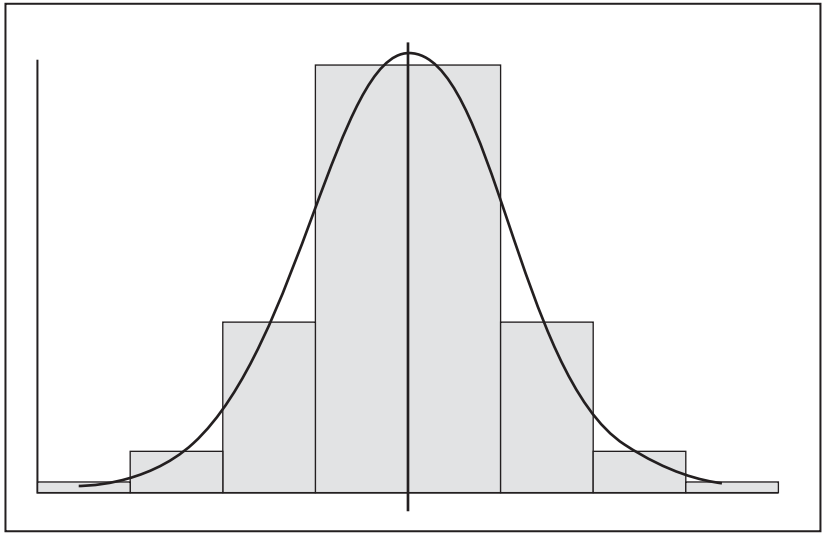
Une distribution est normale lorsqu'elle répond à trois critères :

1. elle est construite sur une variable continue ;
2. elle est unimodale ;

3. elle est symétrique ; les effectifs se réduisant au fur et à mesure que l'on s'éloigne de la moyenne sans jamais arriver à zéro. Par conséquent, la moyenne, la médiane et le mode coïncident tous (sont identiques) dans une distribution normale.

La Figure 5.2 présente une distribution normale. On remarque qu'elle est unimodale et que la ligne verticale représente la position des trois statistiques de valeurs centrales (la moyenne, le mode et la médiane). Les trois valeurs étant identiques, elles sont représentées par la même ligne verticale. Dans la Figure 5.3, plusieurs distributions sont représentées.

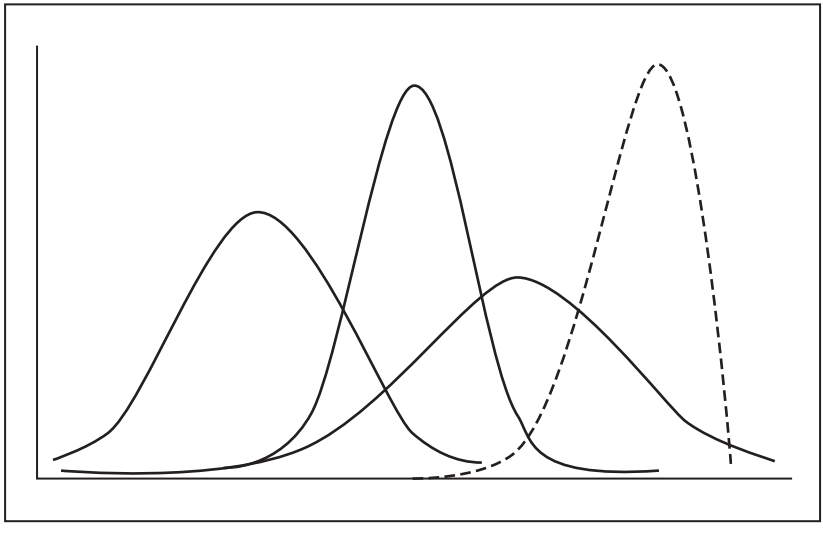
FIGURE 5.2 Distribution de fréquences normale (et son polygone)



Pensons à un cercle. Il existe un nombre infini de cercles possibles — certains étant plus grands que d'autres. Néanmoins, tous les cercles partagent obligatoirement une caractéristique : ils sont ronds. Analogiquement, il existe un nombre infini de courbes normales qui diffèrent toutes, mais qui partagent néanmoins les caractéristiques qui définissent la normalité (unimodale, continue, symétrique, M = Md = Mo). À la Figure 5.3, on remarque trois distributions (les traits noirs pleins) qui sont toutes normales, même si elles ne sont pas identiques. Les moyennes de ces trois distributions sont différentes et leurs écarts types le sont aussi. La distribution en pointillé, par contre, n'est pas une distribution normale. Le Quiz rapide 5.1

invite le lecteur à dire en quoi la courbe en pointillé de la Figure 5.3 n'est pas normale.

FIGURE 5.3 Quelques exemples de distributions



Quiz rapide 5.1
Pourquoi la courbe en pointillé de la Figure 5.3 n'est-elle pas une distribution normale ?

LA DENSITÉ SOUS LA COURBE

Le polygone des fréquences et l'histogramme sont des représentations graphiques de la fréquence (ou de la proportion) des observations se situant à chaque valeur d'une variable. Les statisticiens utilisent le terme de *densité* pour décrire la proportion des observations pour les différentes valeurs d'une distribution. Lorsqu'une distribution est normale, il est possible de déduire seulement à partir de sa moyenne et de son écart-type la *proportion* ou la *densité* des observations qui se trouvent entre chaque valeur de la variable et sa moyenne. Il est aussi possible de déterminer la densité des observations qui sont inférieures ou supérieures à n'importe quelle valeur aussi bien que la densité des observations qui se trouvent entre deux valeurs.

Si le poids des enfants de six ans suit une distribution normale et qu'on connaît sa moyenne et son écart-type, il est possible de déduire la proportion (la densité) des enfants qui pèsent plus de 40 kg, moins de 30 kg, ou la proportion des enfants qui pèsent entre 30 et 40 kg. De plus, connaissant la densité, il devient possible de déterminer la probabilité d'obtenir n'importe quelle valeur : par exemple, si les notes en chimie sont distribuées normalement, nous pouvons établir la probabilité d'obtenir 90 % au prochain examen. Enfin, grâce à la distribution normale, il nous est possible de convertir les valeurs étalons Z attribuées à chaque observation en percentiles et vice-versa. Examinons d'abord le concept de la densité des observations.

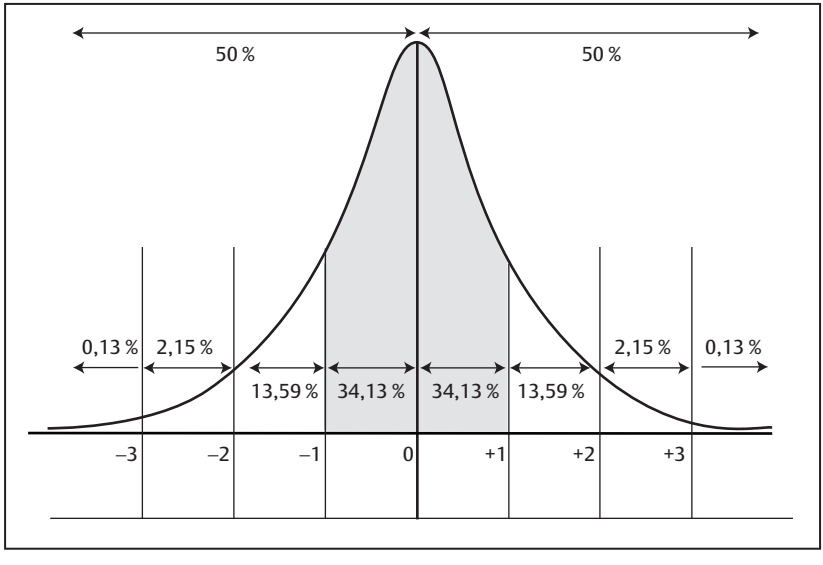
La médiane est la valeur qui divise la distribution en deux groupes égaux. Il y a autant d'observations au-dessus qu'en dessous de la médiane. Pour les distributions normales, la médiane et la moyenne sont égales. Parce que la médiane et la moyenne coïncident, la proportion des observations se trouvant au-dessus et en dessous de la moyenne est égale aussi. Ainsi, pour les distributions normales, la proportion (la densité) des observations se situant au-dessus et en dessous de la moyenne est égale à 0,50.

Par ailleurs, lorsqu'on observe une distribution normale comme celle de la Figure 5.2, on voit que, plus on s'éloigne de la moyenne, moins il y a d'observations. La proportion (densité) des observations se réduit au fur et à mesure que l'on s'éloigne de la valeur centrale. Grâce aux travaux de Gauss, nous connaissons la proportion exacte des observations qui se trouvent à différents points de la distribution lorsque celle-ci est parfaitement normale.

Quiz rapide 5.2
Imaginez une distribution unimodale, symétrique, mais leptocurtique. Est-ce que la densité sous cette courbe serait la même que celle que l'on trouve à la Figure 5.4 ?

La Figure 5.4 montre la répartition de la densité des observations de la distribution normale. On voit que 34,13 % des observations se situent entre la moyenne et les valeurs se trouvant à un écart-type au-dessus de la moyenne ; 13,59 % des observations se situent entre +1 et +2 écarts types de la moyenne ; 2,15 % des observations sont entre +2 et +3 écarts types ; enfin, très peu d'observations (0,13 %) se situent au-delà de +3 écarts types de la moyenne.

FIGURE 5.4 La densité (proportion) des observations sous la courbe normale



Mais ces proportions ne sont pas inventées ! Indépendamment de la moyenne ou de l'écart-type, pour toutes les distributions normales, 34,13 % des observations se trouvent entre la moyenne et ±1 écart-type, etc.

On peut, par exemple, supposer que la taille des hommes adultes suit une distribution normale avec une moyenne de 170 cm et un écart-type de 10 cm. À partir de ces deux seules informations, nous pouvons affirmer que 34,13 % des hommes mesurent entre 160 et 169,999 cm et 34,13 % mesurent entre 170 et 179,999 cm ; que 13,59 % des hommes mesurent entre 180 et 189,999 cm (ainsi qu'entre 150 et 159,999 cm), etc.

Puisque nous connaissons la densité d'une distribution normale, nous pouvons alors estimer la probabilité d'obtenir une observation se trouvant dans une certaine région. Prenons une observation aléatoire de la population. Cette observation pourrait provenir de n'importe quelle partie de la distribution : elle pourrait être au-dessus ou au-dessous de la moyenne et elle pourrait en être proche ou distante. Nous savons que 50 % des observations se situent au-dessus et 50 % se situent en dessous de la moyenne pour toutes les distributions normales. Ainsi nous pouvons affirmer qu'une observation tirée aléatoirement d'une distribution normale aura une chance sur deux

(p = 0,50) de se situer au-dessus (ou au-dessous) de la moyenne mais qu'elle aura une très faible chance de se retrouver très loin de la moyenne. Étudions ceci de plus près.

Nous savons que 34,13 % des observations d'une distribution se situent entre la moyenne et +1 écart-type (voir la Figure 5.4). Par conséquent, la probabilité que notre observation se situe entre la moyenne et +1 écart-type est p = 0,3413. Pour la distribution hypothétique de la taille, ayant une moyenne M = 170 cm et un écart-type s = 10 cm, il existe une probabilité p = 0,3413 qu'une observation prise au hasard soit entre 170 (M = 170) et 180 cm (s = 10 ; 170 +10 = 180). De la même manière, on sait que 34,13 % (p = 0,3413) des observations se retrouvent entre la moyenne et un écart-type en dessous d'elle. Par conséquent, il y a une probabilité p = 0,3413 qu'une observation tirée au hasard de cette distribution de la taille se situe entre 160 et 170 cm.

Quiz rapide 5.3
La moyenne d'une distribution normale est M = 100 et son écart-type s = 20. Nous tirons aléatoirement une observation de cette distribution. En vous référant à la Figure 5.4, quelle est la probabilité que cette observation soit plus grande que 140 ?

De plus, comme la distribution normale est symétrique, les mêmes proportions se retrouvent pour les valeurs au-dessus et en dessous de la moyenne. En additionnant les observations qui sont en dessous de la moyenne et celles qui se trouvent au-dessus de la moyenne, nous voyons que 68,26 % des observations se trouvent entre –1 et +1 écart-type de la moyenne (34,13 % + 34,13 % = 68,26 %) ; que 27,18 % des observations se trouvent entre –2 et –1 et entre +1 et +2 écarts types (13,59 % +13,59 %) ; et que 4,30 % des observations (2,15 %+2,15 %) sont entre –2 et –3 et entre +2 et +3 écarts types. Enfin, seulement une très petite proportion des observations (0,13 %+0,13 %=0,26 %) se situe en deçà de –3 et au-delà de +3 écarts types de la moyenne. La somme de ces proportions (68,26 + 27,18 + 4,30 + 0,26) = 100 %, ce qui confirme qu'elles incluent toutes les observations de cette distribution normale.

Ce même constat peut être formulé en termes probabilistes. Ainsi, si nous revenons à la distribution de la taille, dont la moyenne est de 170 cm

avec un écart-type de 10, nous pouvons établir que la probabilité d'avoir une taille entre 160 et 180 cm est p = (0,3413 + 0,3413) = 0,6826, d'avoir une taille entre 150 et 160 cm et entre 180 et 190 cm est p=(0,1359 + 0,1359) = 0,2718, et que la probabilité d'avoir une taille de moins de 150 cm et de plus de 190 cm est p = (0,0215 + 0,0215 + 0,0013 + 0,0013) = 0,0456. La probabilité qu'un homme soit très grand (plus de 190 cm) ou très petit (moins de 150 cm) est clairement petite. Si l'on additionne ces trois probabilités (0,6826 + 0,2718 + 0,0456), nous trouvons une probabilité p = 1,0, indiquant qu'un homme choisi aléatoirement aura certainement une taille !

Supposons une distribution normale pour un test de QI administré à 1 000 élèves. Supposons aussi que la moyenne du QI est de 100 et que l'écart-type des QI est de 15. Que pouvons-nous conclure au sujet du QI de ces élèves ?

1. Environ 500 élèves ont un QI supérieur à 100 et 500 élèves ont un QI inférieur à 100. Ainsi, la probabilité que l'étudiante X possède un QI supérieur à la moyenne est p = 0,50.
2. Environ 341 élèves ont un QI entre 100 et 115, et 341 élèves ont un QI entre 85 et 100. (La moyenne plus 1 écart-type est égale à 100 + 15 = 115 et la moyenne moins 1 écart-type vaut 100 – 15 = 85.) Puisque 34,13 % des QI se situent entre la moyenne et +1 écart-type, nous savons alors que 341 (34,13 % de 1 000 élèves = 341 approximativement) élèves ont un QI entre 100 et 115. Puisque 68,26 % des observations se situent entre -1 et +1 écart-type de la moyenne, un total d'environ 682 élèves ont un QI entre 85 et 115 (68,26 % de 1 000 = 682 approximativement).
3. Environ 136 (13,6 %) élèves ont un QI entre 115 et 130 (13,6 % de 1 000), et 136 ont un QI entre 70 et 85.
4. Environ 22 élèves ont un QI entre 130 et 145 (2,15 % de 1 000, soit 21,5), et 22 ont un QI entre 55 et 70.
5. Seulement 1 élève a un QI supérieur à 145 (0,13 % de 1 000 = 1,3) et seulement 1 élève a un QI inférieur à 55 (0,13 % de 1 000 = 1,3).

Quiz rapide 5.4
Supposons que pour une distribution normale, M = 10 et s = 2. Supposons que vous avez 100 observations. Combien de ces observations sont supérieures à la moyenne ? Combien se situent entre 10 et 14 ? Combien obtiennent une valeur inférieure à 8 ?

LA CONVERSION DES VALEURS ÉTALONS Z EN RANGS PERCENTILES

On se souvient que les valeurs étalons Z (ou T, etc.) et les percentiles sont utilisés pour trouver la position relative des observations. Lorsque les distributions sont normales, nous pouvons facilement traduire les valeurs étalons Z en percentiles et vice-versa. Certains tests psychologiques standardisés expriment les résultats en valeurs T ou, plus rarement, en valeurs étalons Z. Il est souvent préférable d'expliquer ces résultats à une personne en faisant appel aux percentiles, une information qui est plus facilement comprise. Lorsque les résultats sont exprimés en valeurs étalons T, il faut préalablement les convertir en valeurs étalons Z avant de les traduire en percentiles (voir le chapitre 4).

La logique de base se comprend facilement. Le percentile indique la proportion des observations égales ou inférieures à n'importe quelle valeur d'une distribution. La médiane indique la valeur qui coupe la distribution en deux parties égales. Puisqu'il s'agit d'une distribution normale, la moyenne et la médiane sont identiques. Donc, pour une distribution normale, 50 % des observations sont égales ou inférieures à la moyenne. Quel serait alors le percentile associé à une valeur se trouvant exactement à la moyenne ? Trouvons d'abord la valeur de cette observation en valeur étalon Z. Puisqu'elle se trouve à la moyenne, sa valeur Z est égale à zéro (voir le chapitre 4). Nous pouvons alors conclure que 50 % des valeurs de la distribution seront égales ou inférieures à Z = 0, ce qui définit un percentile de 50 pour cette observation. À partir de la cote Z, nous avons déduit le percentile !

Prenons maintenant une observation se situant à +1 écart-type de la moyenne (par exemple à 115 lorsque la moyenne M = 100 et l'écart-type s = 15). Cette observation se traduit par une valeur étalon Z de +1 [Z = (115 – 100)/15 = +1]. Nous savons, d'après la Figure 5.4 que 34,13 % des observations se trouvent entre la moyenne et +1 écart-type. Nous savons aussi que 50 % des observations se trouvent *en dessous de* la moyenne. Nous faisons la somme pour trouver que 50 % + 34,13 % = 84,13 %. Ce nombre représente la proportion des observations se trouvant à ou en dessous de +1 écart-type (Z = +1) de la moyenne. Puisque 84,13 % des observations se trouvent à cette valeur ou en dessous, il s'agit donc du rang percentile 84,13 ou, plus simplement, 84.

Procédons de la même façon pour une valeur se trouvant à +2 écarts types de la moyenne (c'est-à-dire 130 lorsque M = 100 et s = 15). Une observation se situant à 2 écarts types au-dessus de la moyenne a une côte Z de +2 [(130-100)/15 = +2]. Puisque la valeur est à +2 écarts types, elle doit être supérieure à la moyenne, et donc son rang percentile supérieur à 50. Nous savons, d'après la Figure 5.4, que 50 % des observations se trouvent en dessous de la moyenne, que 34,13 % se situent entre la moyenne et +1 écart-type et que 13,59 % des observations se trouvent entre +1 et +2 écarts types. Nous additionnons alors ces trois proportions : 50 % + 34,13 % + 13,59 % = 97,72 %. Nous concluons alors que 97,72 % des observations sont égales ou inférieures à 130. En arrondissant, cette observation se situe au rang percentile 98. Lorsque nous avons un QI de 130, il est égal ou supérieur à 97,72 % des QI de la population, et par soustraction (100 %-97,72 %), seulement 2,28 % des personnes détiennent un QI plus élevé.

Quiz rapide 5.5
Quel sera le rang percentile pour une observation se trouvant à plus de +3 écarts types de la moyenne ?

Souvenons-nous que, pour les distributions normales, 50 % des observations se situent de chaque côté de la moyenne. Trouvons maintenant le rang percentile d'une observation se situant à un écart-type *en dessous de* la moyenne. Puisque cette observation est inférieure à la moyenne, son rang percentile devra être plus petit que 50. Nous savons que 34,13 % des observations se trouvent entre la moyenne et cette observation. Donc, cette observation se situera à 50 % – 34,13 % = 15,87 % ou (en arrondissant) au rang percentile 16. La position en percentile d'une observation se situant à –2 écarts types de la moyenne sera de 2,28 %, puisque 13,59 % des observations sont entre –1 et –2 écarts types, le calcul est simple : 50 – 34,13 – 13,59 = 2,28 (percentile 2).

Quiz rapide 5.6
Quel sera le rang percentile de l'observation se trouvant à –3 écarts types de la moyenne ? En supposant qu'elle est normale, quelle est la proportion des observations se situant entre ± 3 écarts types de la moyenne ?

Comment trouver la densité des observations se situant entre deux valeurs ?

On peut déterminer la proportion des observations se trouvant entre deux valeurs de la distribution normale, à condition que les observations soient ou puissent être converties en valeurs étalons Z (parce que la moyenne et l'écart-type de la distribution sont connus). Il s'agit de trouver la densité sous la courbe pour les deux valeurs et de les soustraire. Prenons comme illustration deux performances à un examen, l'une se situant à la moyenne de la classe (Z = 0) et l'autre se situant à +1 écart-type (Z = +1). Les densités pour ces deux valeurs sont respectivement de 0,50 et de 0,8413. La différence entre les deux est de 0,3413, indiquant que 34,13 % des étudiants ont obtenu une note entre la moyenne et +1 écart-type. Par conséquent, la probabilité d'obtenir un résultat entre la moyenne et +1 écart-type est p = 0,3413.

LA CONVERSION DES RANGS PERCENTILES EN VALEURS ÉTALONS Z

Faisons l'inverse maintenant, en présumant toujours la normalité. On suppose qu'une observation se trouve au rang percentile 84. Quelle est sa position en valeur étalon Z ? Le rang percentile étant plus grand que 50, il est certain qu'elle se situe au-dessus de la moyenne et que, par conséquent, sa valeur étalon Z sera positive (supérieure à Z = 0). À partir de la Figure 5.4, on sait qu'approximativement 34 % des observations se trouvent entre la moyenne et une valeur qui est à +1 écart-type de la moyenne. Donc, lorsque le rang percentile est égal à 84, la valeur étalon Z est égale à +1. Au rang percentile 98, nous sommes à la valeur étalon +2. À l'inverse, un rang percentile de 15,87 (ou 16) implique que Z = –1, et un rang percentile de 2 implique que Z = –2. Le Tableau 5.1 résume ces relations. Dans la colonne de gauche, on lit la valeur étalon Z, et dans la colonne de droite, on lit la proportion des observations égales ou inférieures à cette valeur Z. Par exemple, 0,13 % des observations d'une distribution normale sont égales ou inférieures à une valeur située à Z = –3, et 99,87 % des observations sont égales ou inférieures à une observation dont la position en valeur étalon Z = +3.

Ces calculs sont plutôt simples lorsqu'on travaille avec des valeurs qui se situent exactement à ±1, ±2 ou ±3 écarts types de la moyenne, une fois ces

valeurs converties en scores Z. Mais que fait-on lorsqu'il s'agit d'observations qui ne tombent pas exactement sur ces valeurs ? Quel est le percentile pour une observation qui se situe à Z = +0,83 ou Z = –1,48 ? L'idéal serait d'avoir un tableau comme le Tableau 5.1, mais qui inclurait toutes les valeurs étalons Z possibles et la densité associée à chacune. Le tableau de la *proportion sous la courbe normale standardisée*, qui est reproduit intégralement dans l'annexe (Tableau A.1), a été construit pour répondre à ce besoin.

Tableau 5.1
Valeur étalon Z et rang percentile correspondant

Z	–3	–2	–1	0	+1	+2	+3
Rang percentile	0,13 %	2,28 %	15,87 %	50,00 %	84,13 %	97,72 %	99,87 %

Le tableau de la proportion sous la courbe normale standardisée

Il importe de savoir comment lire et interpréter le tableau de la proportion sous la courbe normale standardisée se trouvant en annexe. Le Tableau 5.2 en donne un extrait. Il comprend deux colonnes et un grand nombre de rangées. La colonne de gauche indique une suite de valeurs étalons Z allant de 0,00 à +3, alors que celle de droite indique la proportion des observations qui se trouvent à cette valeur ou en dessous[2].

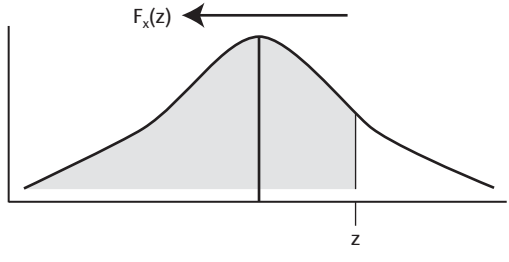
Supposons qu'on veuille estimer la proportion des observations qui se situent à la moyenne ou qui sont plus petites que la moyenne. On sait qu'une observation à la moyenne se situe à Z = 0. On trouve la valeur Z = 0 dans la colonne de gauche du Tableau 5.2, et de celle de droite, on lit la proportion des observations qui se trouvent à cette valeur Z = 0 ou en dessous de cette valeur. Dans ce cas, il s'agit de 0,50, indiquant que 50 % des observations se trouvent à la moyenne ou au-dessous d'elle. On pouvait s'attendre à ce résultat puisque, avec les distributions normales, 50 % des observations se trouvent de chaque côté de la moyenne. On peut alors affirmer que la densité des observations sous Z = 0,0 est 0,50 ou 50 %.

2. La plupart des tableaux de la densité sous la courbe normale vont de 0 à +4. Mais cela ne veut pas dire que le Z maximal est +4. Pour la distribution normale théorique, il n'y a pas de limite aux valeurs possibles.

Supposons que l'on désire connaître la densité des observations se situant à Z = 1,0 ou en dessous d'elle. Au Tableau 5.2, on trouve à gauche la valeur Z = 1, et à droite, 0,8413, indiquant que 84,13 % des observations se trouvent à Z = 1 ou en dessous d'elle. Si l'on désire connaître la densité des observations à Z = 0,06 ou en dessous, on peut trouver, au Tableau 5.2 et dans l'appendice, qu'il s'agit d'une proportion de 0,5239 ou de 52,39 % des observations.

On peut aussi faire appel au tableau de la densité sous la courbe normale pour déterminer le rang percentile pour n'importe quelle valeur Z. Quel serait le rang percentile pour la personne se situant à Z = 0,06 ? Puisque 52,39 % des observations se situent à Z = 0,06 ou moins, nous savons alors que le percentile associé à cette valeur est 52,39 ou, plus simplement, 52.

Tableau 5.2
Extrait du tableau de la proportion sous la courbe normale standardisée



z	$F_x(z)$
0,00	0,5000
0,06	0,5239
0,26	0,6026
0,50	0,6915
0,70	0,7580
0,76	0,7764
0,90	0,8159
1,00	0,8413
1,26	0,8962
1,50	0,9332
1,74	0,9591
2,00	0,9772
3,00	0,9987

On peut faire appel à ce tableau de la densité sous la courbe normale pour les valeurs qui se situent en dessous de la moyenne. Puisqu'elles se situent en dessous de la moyenne, leurs valeurs étalons Z prendront un signe négatif. Supposons que l'on désire déterminer la proportion des observations se situant à ou sous Z = –0,26. Pour l'instant, on peut ignorer le signe de cette valeur Z et trouver d'abord la densité qui y correspond (au Tableau 5.2 et à l'appendice, cette densité est de 0,6026). On sait que la distribution contient 100 % des observations. Il ne reste alors qu'à soustraire la densité répertoriée dans le tableau du total de la distribution ; à la valeur 100 % (ou p = 1,0). Dans ce cas, on aurait 1,0 – 0,6026 = 0,3974. Ainsi, avec une performance de Z = –0,26, cette observation est égale ou plus forte que 39,74 % des performances et le percentile associé à cette performance est 39,74, ou simplement 40.

Pour une valeur Z = –2, la densité qui y correspond dans le Tableau 5.2 est 0,9772. On soustrait cette valeur de 1,00 et on trouve 0,028. Ainsi, on peut conclure que 2,28 % des observations se trouveront à la valeur Z = –2 ou en dessous. On peut ainsi conclure que cette observation (Z = –2) se situe au centile 2,28 (ou 2).

SOMMAIRE DU CHAPITRE

La distribution normale est la forme habituelle que prend la distribution de plusieurs variables continues. Une distribution est normale lorsqu'elle est construite sur une variable continue qui est unimodale et qui détient une moyenne, un mode et une médiane identiques, ce qui lui garantit une forme symétrique. Lorsqu'une distribution est normale, nous pouvons savoir la proportion (la densité) des observations qui se trouvent en dessous ou au-dessus de la moyenne, la proportion qui se trouve entre n'importe quelle valeur et la moyenne, la proportion des observations qui y sont supérieures et, enfin, la proportion des observations se trouvant entre deux valeurs. Tant que nous connaissons la moyenne et l'écart-type d'une distribution normale, il est possible de déterminer pour n'importe quelle valeur exprimée en valeur étalon Z son rang percentile et, à partir du rang percentile, de déterminer sa valeur Z en faisant appel au tableau de la densité sous la courbe normale. Enfin, toutes ces valeurs peuvent être exprimées en termes probabilistes.

EXERCICES DE COMPRÉHENSION

1. Concernant la distribution normale, laquelle de ces affirmations est fausse ?
   a) Elle est la base de plusieurs analyses statistiques.
   b) Elle s'étend maximalement entre –4 et +4 écarts types.
   c) L'aire sous la courbe (la densité) correspond à la probabilité.
   d) Plus d'observations seront proches de la moyenne que loin d'elle.
2. Dans cette distribution, la moyenne, la médiane et le mode sont tous identiques. La distribution _____.
   a) n'est pas normale
   b) pourrait être normale
   c) est tout à fait normale
   d) impossible à déterminer
3. Nous convertissons toutes les données de cette distribution normale en valeurs étalons Z et nous examinons la distribution résultante.
   a) Sa moyenne est égale à 0.
   b) Sa variance est de 1.
   c) Elle est en forme de cloche.
   d) Toutes ces réponses sont justes.
4. Nous convertissons chaque valeur d'une distribution asymétrique négative en valeur étalon Z. Quelle sera la forme de la distribution de ces valeurs Z ?
   a) Normale
   b) Asymétrique négative
   c) Asymétrique positive
   d) Toutes ces réponses sont possibles.

Pour les questions 5 à 9, vous devez faire appel au tableau de la densité sous la courbe normale.

5. Cent étudiants ont subi un examen où la moyenne du groupe est de 75 % avec un écart-type de 10. Les résultats se distribuent normalement. Combien d'étudiants ont obtenu 75 % ou moins à l'examen ?
6. Combien d'étudiants ont obtenu entre 75 et 85 % à l'examen ?
7. Quelle est la probabilité qu'un étudiant ait une note supérieure à 95 % ?

8. Un étudiant obtient la note de 55 % à son examen. À combien d'écarts types de la moyenne est-ce que cette note se situe ?

9. Quel est le percentile pour l'étudiant qui a obtenu 55 % à son examen ?

Réponses

1. b
2. b (Notez que l'énoncé du problème n'inclut qu'un seul des trois critères qui définissent une distribution normale.)
3. d
4. b
5. 50
6. 34
7. (1 – 0,9772) = 0,0228
8. –2
9. 2 (2,28)

CHAPITRE 6
LA CORRÉLATION

Page laissée blanche

CHAPITRE 6

LA CORRÉLATION

Jusqu'à présent, nous avons appris à décrire les variables, les distributions et les observations à l'intérieur des variables. Nous abordons maintenant la relation qui existe entre les variables, et que l'on nomme la *corrélation*. La corrélation est une méthode qui permet de déterminer le degré de coïncidence entre deux variables.

Les corrélations jouent un rôle important dans la vie quotidienne. On peut remarquer qu'il pleut parfois lorsque le ciel est ennuagé, tandis qu'il ne pleut jamais en l'absence de nuages. On se rend compte qu'on tousse souvent lorsqu'on a un rhume, alors que cela n'arrive que rarement lorsqu'on n'a pas de rhume. Peut-être a-t-on aussi remarqué que les résultats aux examens s'améliorent lorsqu'on leur a consacré plus de temps d'étude ? En fait, on vient de noter qu'il existe une corrélation entre la présence de nuages et la pluie, le rhume et la toux ainsi que l'assiduité à l'étude et les résultats scolaires. Y a-t-il plus de pauvreté dans les plus grandes villes ? Le nombre de meurtres est-il plus grand dans les sociétés où les citoyens ont plus d'armes à feu ? On peut répondre à toutes ces questions par le biais de la corrélation. La corrélation est une procédure statistique qui permet de quantifier le degré avec lequel deux événements tendent à être reliés (la présence de nuages et la pluie ; le rhume et la toux ; les notes et le temps d'étude ; les meurtres et les armes à feu ; la pauvreté et la taille des villes). Pour établir cette relation, il est nécessaire d'avoir deux mesures pour chaque observation. Ainsi, si nous voulons calculer la corrélation entre le QI et les notes scolaires, nous devons avoir accès à un groupe de personnes pour lesquelles

nous possédons à la fois le QI et les notes scolaires. Si nous voulons établir la relation entre la taille des villes et le degré de pauvreté, nous devons avoir, pour chaque ville de la distribution, sa taille et son degré de pauvreté. Quel qu'il soit, le sujet d'analyse (une personne, une ville, une classe, etc.) doit fournir deux informations, l'une se rapportant à une variable, l'autre à une deuxième variable. Il existe plusieurs types de corrélations, mais celle que Karl Pearson a développée — et qu'on appelle *corrélation simple, corrélation d'ordre zéro, corrélation bivariée* ou *corrélation linéaire* — est celle qui, dans la pratique, est la plus utilisée.

LA CORRÉLATION DE PEARSON

La *corrélation de Pearson* est une procédure statistique qui produit un *coefficient de corrélation,* un index du degré de relation linéaire qui existe entre deux mesures (nous verrons plus tard ce qu'on entend par « linéaire »). Il y a plusieurs types de corrélations, mais celle dont nous discutons ici, la corrélation de Pearson, est utilisée lorsque nous désirons établir la relation qui existe entre des variables mesurées sur des échelles à intervalles ou des échelles de rapport. La corrélation de Pearson prend des valeurs variant entre –1 et +1. Nous disons que la corrélation est parfaite lorsqu'elle atteint des valeurs numériques extrêmes (+1 ou –1) et qu'elle est nulle quand le coefficient prend la valeur de 0. La relation peut être positive ou négative. Par exemple, la corrélation entre la présence de nuages et la pluie est positive, car plus il y a de nuages, plus grandes sont les chances qu'il pleuve. Souvent, comme dans le cas de la relation nuages-pluie, la relation n'est pas parfaite (il ne pleut pas toujours lorsque le ciel est couvert). Par exemple, bien qu'il existe une corrélation entre le niveau d'intelligence et le succès scolaire, la relation est loin d'être parfaite. Souvent, des étudiants intelligents ne réussissent pas aussi bien que des étudiants moins doués et vice-versa. Dans ce cas, la corrélation de Pearson prendra des valeurs positives, mais moins grandes que +1 (par exemple +0,50 ou +0,12).

Y a-t-il une relation entre la satisfaction au travail et l'absentéisme ? Oui, mais la corrélation est négative (par exemple –0,20). Dans ce cas, plus les gens sont satisfaits, moins ils s'absentent. La valeur –0,20 (la relation satisfaction-absence) est non seulement négative, mais elle est aussi moins

grande que la relation entre les nuages et la pluie (0,50), car très souvent, nous allons au travail même lorsque nous n'aimons pas cela et, parfois, nous nous absentons même lorsque nous adorons notre travail. Enfin, certains phénomènes ne sont pas liés. Y a-t-il une relation entre la quantité de crème glacée vendue à New York chaque jour de l'été et le nombre de naissances à Montréal ayant lieu les mêmes jours ? Il y a fort à parier qu'une telle relation n'existe pas. La corrélation entre la consommation de crème glacée et le taux de natalité sera alors proche de 0,0. De manière similaire, l'habileté sociale et l'intelligence ne sont pas en corrélation.

La corrélation de Pearson est un indice pratique qui nous renseigne simultanément sur deux aspects de la relation (linéaire) entre deux variables :

1. *La magnitude de la relation :* plus la corrélation est proche de +1 ou de –1, plus elle est forte.
2. *La direction de la relation :* une *corrélation positive* indique que plus les valeurs d'une variable sont grandes, plus les valeurs de l'autre variable seront grandes aussi. Une *corrélation négative* implique que plus les valeurs d'une variable augmentent, plus elles se réduisent pour la deuxième variable.

La corrélation de Pearson est représentée par le symbole r_{xy}. Elle se calcule entre seulement deux variables à la fois, que nous représentons généralement par les symboles X et Y. Pour cette raison, nous lui donnons parfois le nom de *corrélation bivariée* : la relation entre deux variables. Si la corrélation entre deux variables X et Y est égale à 0,5, nous écrivons : $r_{xy} = 0{,}50$.

LA LOGIQUE QUI SOUS-TEND LE CALCUL DE LA CORRÉLATION

La corrélation quantifie le niveau de similarité entre deux variables. Le problème consiste donc à trouver une façon de définir mathématiquement la similarité. Une manière évidente serait de vérifier si les sujets produisent la même réponse (numérique) pour deux variables. Lorsque les valeurs obtenues pour une variable tendent à être reproduites sur une autre, il y a une relation forte entre les variables. Une solution au calcul de la corrélation serait alors de calculer la différence entre les valeurs de chaque variable. S'il n'existait pas de différence entre les valeurs des deux variables pour chaque observation, nous pourrions dire que la corrélation est parfaite.

Par exemple, supposons que nous avons la note obtenue par un groupe d'étudiants à deux examens. Si les étudiants obtiennent exactement la même note aux deux examens, il est facile de conclure que la relation (la corrélation) entre les deux examens est parfaite.

Supposons maintenant que nous désirons calculer la corrélation entre deux examens, mais qu'un examen est noté sur 100 et l'autre sur 20. Le Tableau 6.1 présente les données.

Tableau 6.1
Notes obtenues à deux examens par les mêmes étudiants

Étudiant	*Note sur 100*	*Note sur 20*
A	95	19,0
B	87	17,4
C	74	14,8
D	56	11,2
E	43	8,6

Aucun étudiant n'obtient la même note aux deux examens parce que l'échelle de mesure n'est pas la même pour les deux variables : les notes au premier examen (notes sur 100) peuvent varier entre 0 et 100 tandis que l'étendue pour la deuxième variable est de 0 à 20. Si nous comparons les deux séries de résultats en les soustrayant, la différence entre les notes obtenues aux deux examens ne sera jamais zéro. Par conséquent, nous devrions conclure qu'il n'existe pas de similitude (de « corrélation ») entre les notes aux deux examens.

Quiz rapide 6.1
Quelle est la coordonnée de l'étudiant B au Tableau 6.1 ?

Prenons un autre exemple. On se doute bien qu'il existe une relation entre l'ancienneté et le salaire : les employés détenant plus d'expérience reçoivent généralement un salaire plus élevé. Or, le salaire est chiffré en milliers de dollars alors que les années d'expérience sont mesurées en quelques

années. La simple différence entre année et salaire ne sera jamais égale à zéro, et nous devrions conclure qu'il n'y a pas de relation entre ces deux variables, ce qui n'est pas sensé.

Donc, si nous basons le calcul de la corrélation sur la simple différence numérique obtenue entre deux mesures, la conclusion sera erronée, à moins que les deux mesures ne soient sur la même échelle de mesure (ayant la même moyenne et la même variance). Puisque nous voulons souvent calculer la corrélation entre deux variables qui ne sont pas mesurées sur la même échelle, il faut trouver une approche plus générale.

La méthode la plus générale et la plus satisfaisante pour décrire la similitude entre deux variables est celle choisie par Pearson. *La corrélation entre deux variables est définie comme étant le degré avec lequel la position relative des observations est la même sur deux variables.* Si nous utilisons cette définition pour le Tableau 6.1, nous voyons qu'il existe effectivement une relation entre la performance aux examens. Par exemple, l'étudiant A obtient la meilleure note aux deux examens, l'étudiant B obtient la note juste en dessous aux deux examens, ainsi de suite jusqu'à l'étudiant E qui obtient la note la plus faible aux examens. Les étudiants maintiennent exactement la même position relative dans chacun des examens.

Nous avons déjà abordé le concept de position relative au chapitre 4. La position d'une observation sur une mesure se définit comme l'écart standardisé qui existe entre la valeur obtenue sur une variable par une observation et la moyenne de cette variable. La valeur étalon Z est justement une manière pratique de calculer cette position. Ainsi, *la corrélation de Pearson mesure le degré de coïncidence entre les valeurs étalons Z, obtenues sur deux mesures*: la corrélation est forte lorsque les valeurs Z obtenues par chaque personne sur les deux variables sont similaires et, dans le cas contraire, la corrélation est plus faible.

Lorsque les valeurs Z obtenues par un ensemble de personnes sur deux variables coïncident, la corrélation est parfaite ($r_{xy} = +1{,}0$): les valeurs Z pour les deux variables sont simultanément positives, négatives ou nulles. Lorsque les valeurs Z des deux variables coïncident, mais qu'elles sont de signes inversés (l'une positive, l'autre négative), la corrélation est parfaite, mais négative ($r_{xy} = -1{,}0$). Lorsque les deux valeurs Z obtenues sont moins semblables (elles coïncident approximativement ou seulement quelquefois),

la corrélation obtenue ne sera pas exactement –1 ni +1, mais elle sera entre ces deux extrêmes. Lorsqu'elles ne coïncident pas du tout, la corrélation est égale à zéro.

Comment calculer la corrélation de Pearson entre deux variables ?

On se souvient que la corrélation se définit par le degré avec lequel la position des observations sur deux variables se maintient. La formule suivante définit formellement la corrélation[1].

$$r_{xy} = \frac{\sum_{i=1}^{N} Z_{Xi} \times Z_{Yi}}{N - 1} \qquad \text{Formule 6.1}$$

où Z_{Xi} et Z_{Yi} correspondent à la position relative de chaque observation sur les variables *X* et *Y* exprimées en valeurs étalons Z, et $N - 1$ est le nombre de sujets moins 1. Nous verrons plus tard la signification du $N - 1$ (voir le chapitre 9).

Les quatre étapes pour obtenir la corrélation de Pearson sont :

1. Convertir chaque valeur en valeur étalon Z.
2. Multiplier les paires de valeurs étalons Z de chaque sujet de l'échantillon.
3. Faire la somme de ces produits.
4. Diviser cette somme par le nombre d'observations moins un.

Le numérateur de la Formule 6.1 donne le degré total de similarité entre les deux mesures. En divisant cette quantité par $N - 1$, on obtient la moyenne de la similarité. *La corrélation est donc un indice de la similarité moyenne dans la position qu'occupent les observations sur les deux variables.*

La corrélation positive parfaite ($r_{xy} = +1{,}00$)

La corrélation positive parfaite indique que les valeurs des deux variables augmentent ou diminuent ensemble pour toutes les observations. Les

1. Il existe plusieurs formules pour calculer la corrélation de Pearson, dont :

$$r_{xy} = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\left[(N\sum X^2) - (\sum X)^2\right]\left[(N\sum Y^2) - (\sum Y)^2\right]}$$

Les amateurs d'algèbre découvriront que toutes ces formules sont identiques.

observations qui sont fortes sur une variable le sont aussi sur l'autre, et celles qui sont faibles sur l'une sont faibles sur l'autre. Puisque la corrélation indique le degré avec lequel les observations maintiennent la même position sur les deux variables, cela implique que les valeurs étalons Z associées à chaque observation seront positives ou négatives sur les deux variables et identiques lorsque la corrélation sera parfaite et positive. Lorsque les valeurs Z_x et Z_y ne sont pas identiques, mais que l'ordre des observations est identique sur les deux variables, les corrélations seront très proches (mais pas nécessairement tout à fait) +1,00.

Le Tableau 6.2 reprend les données du Tableau 6.1 et inclut la valeur étalon Z de chaque observation afin de produire le coefficient de corrélation de Pearson par l'entremise de la Formule 6.1.

La corrélation positive parfaite obtenue au Tableau 6.2 (r_{xy} = +1,00) confirme que la position relative de chaque étudiant demeure exactement la même aux deux examens. Remarquez que la note obtenue par les étudiants D et E est au-dessous de la moyenne pour les deux examens. Mais, puisque le produit de deux quantités négatives est toujours positif, la somme finale sera elle aussi positive. De manière similaire, les étudiants A et B obtiennent tous deux des valeurs Z positives aux deux examens, et le produit de ces deux valeurs sera positif, lui aussi. Dans ce cas, le résultat final sera une corrélation parfaite (r_{xy} = +1,00).

Le graphique de dispersion pour décrire la corrélation

Traçons un graphique qui représente la relation entre la variable X et la variable Y. Ce type de graphique se nomme *graphique de dispersion* ou encore *nuage de points*. L'ordonnée du graphique représente la valeur produite par chacun des sujets sur la variable Y et l'abscisse représente la valeur de ces mêmes sujets sur la variable X. En général, les coordonnées se définissent par la valeur de la variable initiale, mais il est aussi possible de la représenter en valeur étalon Z. Dans le cas présent, les notes à l'examen *X* sont indiquées sur l'abscisse alors que les notes à l'examen *Y* sont placées le long de l'ordonnée. À l'intersection de chaque valeur *X* et de sa valeur *Y* correspondante, nous plaçons une marque qui indique la position de cette observation. Ce point se nomme *la coordonnée* pour cette observa

Tableau 6.2
Corrélation entre les notes obtenues à deux examens par les mêmes étudiants

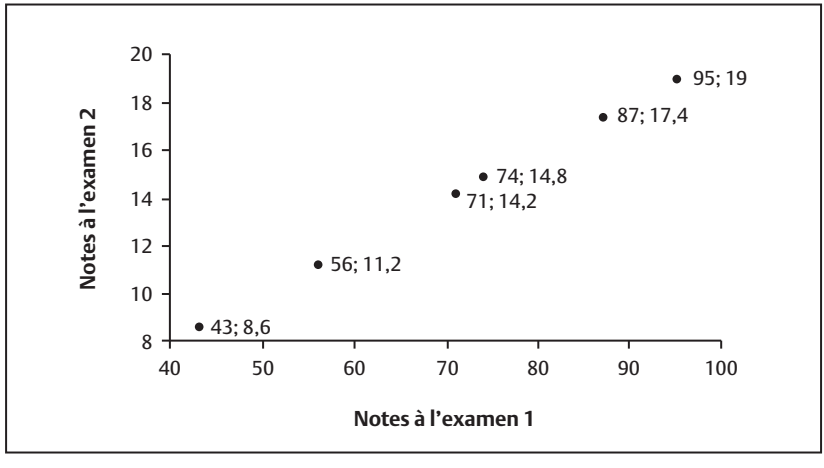
	Examen 1 Note sur 100		*Examen 2 Note sur 20*		
Étudiant	***Score brut X***	Z_X	***Score brut Y***	Z_Y	$Z_{Xi} \times Z_{Yi}$
A	95,0	1,12	19,0	1,12	1,25
B	87,0	0,74	17,4	0,74	0,55
C	74,0	0,14	14,8	0,14	0,02
D	56,0	–0,70	11,2	–0,70	0,49
E	43,0	–1,30	8,6	–1,30	1,70
Somme	355,0		71,0		4,00
N	5		5		5
Résultat	71,0		14,2		1,00
Nom de la statistique	M_X		M_Y		r_{XY}
$r_{xy} = \Sigma(Z_{Xi} \times Z_{Yi}) / (N - 1) = 4 / (5 - 1) = 4 / 4 = 1{,}00$					

tion. Par exemple, la position de l'étudiant E est le point qui se trouve à la coordonnée $\{X, Y\} = \{43{,}0; 8{,}6\}$. La Figure 6.1 indique les coordonnées pour chaque étudiant (habituellement, nous n'indiquons pas les coordonnées des points sur le graphique). Nous répétons cette procédure et, à la fin du processus, la position de toutes les observations sera représentée par cet ensemble de points.

On remarquera que les deux axes du graphique décrivant le nuage de points ne commencent pas à zéro, car personne n'a obtenu une telle note. Les notes les plus basses étant 43,0 pour l'étudiant E à l'examen 1 (X) et 8,6 pour ce même étudiant à l'examen 2 (Y), le graphique commence la numérotation des axes un peu au-dessous des valeurs minimales des données. Dans ce cas, l'abscisse part de la valeur « 40 », et l'ordonnée, de la valeur « 8 ». Cette stratégie produit un graphique plus lisible.

Le graphique de dispersion est utilisé pour représenter visuellement la relation qui existe entre les X et les Y. La Figure 6.1 montre que les étudiants qui tendent à avoir des notes fortes à l'examen X tendent aussi à avoir des notes fortes à l'examen Y et que les performances qui sont faibles sur X sont associées à des performances faibles sur Y. La relation est positive.

FIGURE 6.1 Les coordonnées : la relation entre les notes aux deux examens



La corrélation négative parfaite (r_{xy} = –1,00)

Prenons maintenant la série de données du Tableau 6.3 illustrée à la Figure 6.2. Cette fois, nous voulons calculer la corrélation qui existe entre le nombre de couches de vêtements que cinq personnes portent et la température extérieure. On s'attend à ce que ces cinq personnes portent progressivement *plus* de vêtements au fur et à mesure que la température *baisse :* une température plus élevée devrait donc être associée à *moins* de couches de vêtements. Statistiquement, on s'attend à obtenir une *corrélation négative* entre les deux variables (X est la température extérieure et Y est le nombre de couches de vêtements).

Au Tableau 6.3, nous trouvons que lorsque les valeurs étalons Z_X sont positives pour la température (il fait plus chaud que la moyenne qui est de 10 °C pour nos données), les valeurs étalons (Z_Y) pour le nombre de cou-

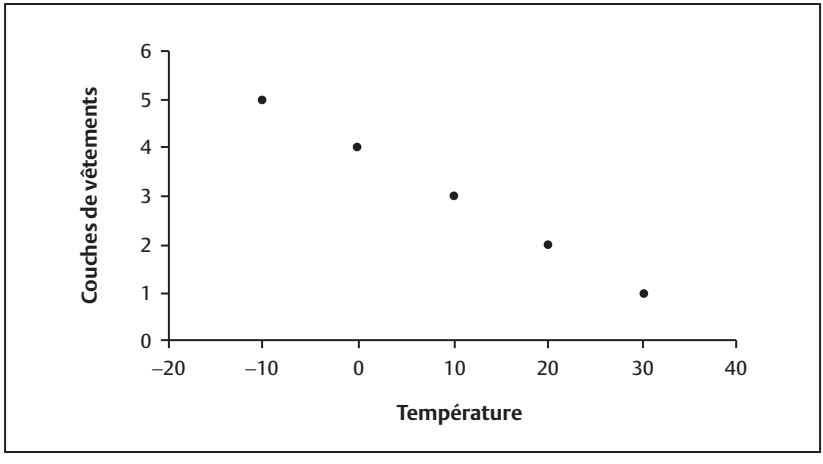
ches de vêtements sont négatives (les personnes portent moins de couches de vêtements que la moyenne, qui est de 3). Les produits $Z_{xi} \times Z_{yi}$ sont tous négatifs, car nous multiplions une valeur Z_X, positive, avec une valeur Z_Y négative, ou vice-versa. La somme de toutes ces valeurs négatives est elle aussi négative (–4). Par conséquent, lorsque nous divisons par N – 1, le calcul indique une corrélation négative (r_{xy} = –1,00).

Tableau 6.3
Corrélation entre la température et le nombre de couches de vêtements portées

	Température en °C		*Nombre de couches de vêtements portées*		
Personne	***Score brut X***	Z_X	***Score brut Y***	Z_Y	$Z_{Xi} \times Z_{Yi}$
A	30	1,26	1	–1,26	–1,60
B	20	0,63	2	–0,63	–0,40
C	10	0,00	3	0,00	0,00
D	0	–0,63	4	+0,63	–0,40
E	–10	–1,26	5	+1,26	–1,60
Somme	50		15		–4,00
N	5		5		5
Résultat	10		3		–1,00
Nom de la statistique	M_X		M_Y		r_{XY}
$r_{xy} = \Sigma(Z_{Xi} \times Z_{Yi}) / N - 1 = -4 / (5 - 1) = -4 / 4 = -1{,}00$					

La corrélation négative indique qu'au fur et à mesure que la température augmente, le nombre de couches de vêtements que l'on porte se réduit, ce qui est raisonnable.

FIGURE 6.2 La relation entre la température et le nombre de couches de vêtements portées



La corrélation nulle (r_{xy} = 0,00)

Les données du Tableau 6.4, illustrées à la Figure 6.3, indiquent le nombre de cigarettes que cinq personnes fument par jour (X) et le nombre de nez (Y) que ces personnes ont ! Nous voyons qu'il n'y a aucune tendance à l'augmentation ou à la réduction des valeurs de Y (nez) au fur et à mesure que les valeurs de X (cigarettes fumées) augmentent. Naturellement, on ne s'attendait pas à détecter une relation entre ces deux variables. Si on calcule la corrélation, on verra qu'elle est égale à zéro : il n'y a aucune relation entre le tabagisme et le nombre de nez. Ce résultat n'est pas une grande surprise, mais on vient de le démontrer statistiquement.

On peut remarquer au Tableau 6.4 que la moyenne pour le nombre de nez est égale à 1 et que toutes les observations portant sur le nombre de nez sont, elles aussi, égales à 1. Par conséquent, toutes les observations se situent exactement à la moyenne (1). La valeur étalon Z pour une observation se trouvant à la moyenne étant 0, toutes les valeurs Z_Y sont égales à 0. Le produit de n'importe quelle valeur par 0 est égal à 0. Donc, pour chaque observation, le numérateur de la Formule 6.1, la quantité $Z_{Xi} \times Z_{Yi}$, est égal à 0. Par conséquent, la somme $\sum(Z_{Xi} \times Z_{Yi})$ est, elle aussi, égale à 0, et en

divisant par $N - 1$, on constate que la corrélation entre le tabagisme et le nombre de nez est $r_{xy} = 0$.

On peut conclure que le tabagisme (même s'il est une mauvaise chose) ne provoque pas la perte du nez. Si on relit cette conclusion, on comprend qu'il s'agit d'une *conclusion causale* (le tabagisme ne *cause* pas la perte du nez). Cette conclusion est tirée d'une corrélation et, dans ce cas, c'est une conclusion valide. Mais cela n'est pas toujours le cas. Nous y reviendrons à la fin de ce chapitre, lorsque nous aborderons la question de la causalité et de la corrélation.

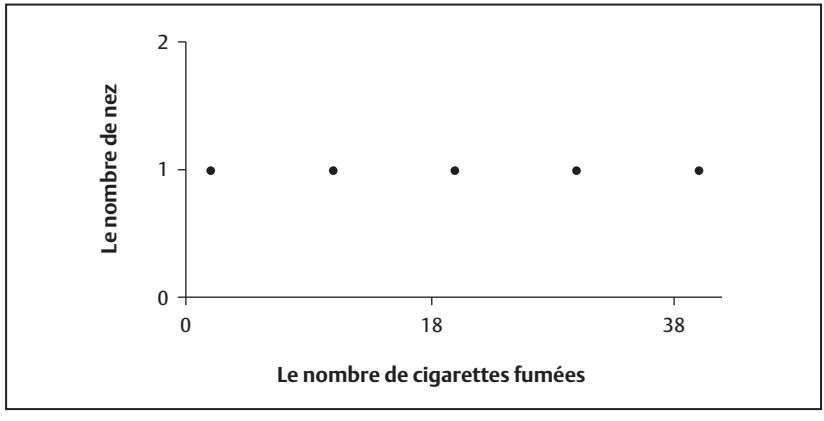
Tableau 6.4
Corrélation entre X (nombre de cigarettes fumées/jour) et Y (nombre de nez de chaque fumeur)

	Nombre de cigarettes fumées/jour		*Nombre de nez*		
Fumeur	***Score brut X***	Z_X	***Score brut Y***	Z_Y	$Z_{Xi} \times Z_{Yi}$
A	40,0	1,26	1,00	0,00	0,00
B	30,0	0,63	1,00	0,00	0,00
C	20,0	0,00	1,00	0,00	0,00
D	10,0	–0,63	1,00	0,00	0,00
E	0,0	–1,26	1,00	0,00	0,00
Somme	100,0		5,00		0,00
N	5		5		5
Résultat	20		1,00		0,00
Nom de la statistique	M_X		M_Y		r_{XY}

$$r_{xy} = \Sigma(Z_{Xi} \times Z_{Yi}) / (N - 1) = 0 / (5 - 1) = 0 / 4 = 0{,}00$$

Quiz rapide 6.2
Selon vous, existe-t-il une relation entre la taille d'une boule de quilles et son poids ? Cette relation est-elle positive ou négative ? Répondez à la même question pour le prix d'un CD et l'argent qu'il vous reste après l'avoir acheté.

FIGURE 6.3 La relation entre le nombre de cigarettes fumées et le nombre de nez



Les corrélations qui ne sont pas parfaites (r_{xy} entre –1,00 et +1,00)

Jusqu'ici, nous avons vu des corrélations parfaites ou nulles (+1, –1 ou 0). Mais en réalité, ces types de corrélations sont plutôt rares. Les corrélations, particulièrement en sciences sociales, tendent à se situer entre ±0,15 et ±0,60, bien qu'elles puissent être plus faibles ou plus fortes dans certains cas. En sciences cognitives ou en sciences économiques, les corrélations sont plus fortes (souvent supérieures à 0,85).

Le Tableau 6.5 présente le salaire et le niveau de scolarité d'un échantillon de 30 personnes. La corrélation entre ces deux mesures est $r_{xy} = + 0, 56$. Le graphique de dispersion qui décrit ces données (Figure 6.4) indique visuellement que les personnes plus scolarisées tendent à obtenir de meilleurs salaires. Ainsi, les personnes qui sont relativement peu scolarisées (la partie inférieure de l'abscisse) tendent à avoir des salaires qui sont plus concentrés vers la partie inférieure de l'ordonnée, et les personnes plus scolarisées (la partie supérieure de l'abscisse) tendent à avoir des salaires plus élevés. On remarque cependant que la corrélation n'est pas parfaite : le salaire n'est pas forcément plus élevé pour toutes les personnes plus scolarisées.

Tableau 6.5
Relation entre salaire et scolarité

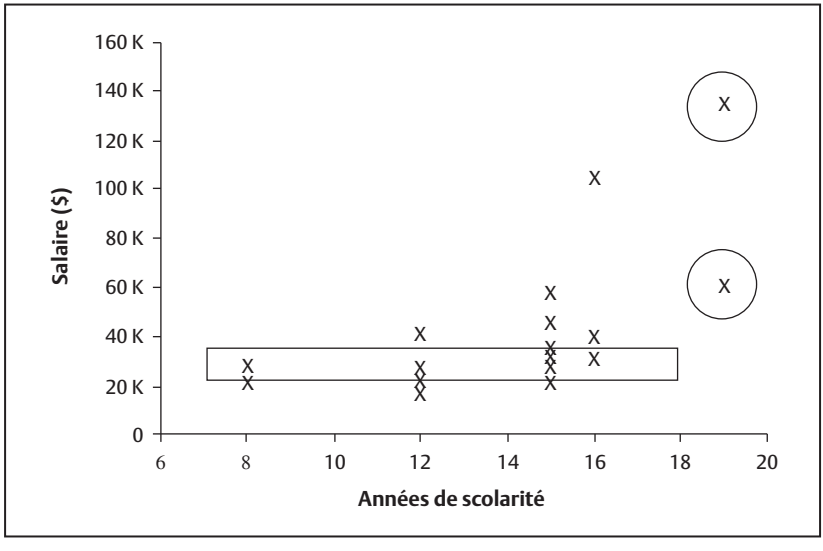
Années de scolarité	*Salaire ($)*	*Années de scolarité*	*Salaire ($)*
8	21 900	15	27 900
8	28 350	15	27 750
12	21 450	15	35 100
12	21 900	15	46 000
12	24 000	15	24 000
12	27 300	15	21 150
12	40 800	15	31 050
12	42 300	15	32 550
12	26 250	15	31 200
12	21 750	16	40 200
12	16 950	16	30 300
15	57 000	16	103 750
15	45 000	16	38 850
15	32 100	19	60 375
15	36 000	19	135 000
$r_{xy} = +0,56$			

Examinons les observations qui sont encerclées à la Figure 6.4. Deux personnes ayant le même niveau de scolarisation (19 années) n'ont pas le même salaire : le salaire de l'une est plus que le double du salaire de l'autre (135 000 et 60 375 $). Les observations encadrées par un rectangle montrent un cas où plusieurs personnes ont le même salaire, bien qu'elles n'aient pas un nombre égal d'années de scolarité. Par exemple, les cinq personnes dont le salaire se situe entre 21 000 et 22 000 $ ont entre 8 et 12 années de scolarité. Nous voyons maintenant ce que la corrélation imparfaite nous dit : il existe effectivement une certaine similarité entre les valeurs Z obtenues entre les deux variables, mais il y a aussi des exceptions.

Quiz rapide 6.3
Supposons que la position de toutes les observations sur la variable X ne se reproduit jamais sur la variable Y. Quelle sera la corrélation entre X et Y ?

La corrélation est un indice de l'ampleur de la relation entre deux variables. Par conséquent, elle permet la comparaison entre les relations. Est-ce que la relation entre X et Y est plus forte que celle qui existe entre A et B ? Par exemple, la corrélation entre la réussite professionnelle (mesurée par le salaire) et le QI pourrait se situer autour de 0,20. La corrélation entre les notes scolaires et le QI pourrait être plus forte, se situant aux alentours de 0,80. Par conséquent, nous pourrions conclure que le QI est plus lié aux notes scolaires qu'à la réussite professionnelle[2]. Ce type d'information est très précieux en recherche comme dans la pratique.

FIGURE 6.4 Le nuage de points de la corrélation salaire-scolarité



2. Il faudra éventuellement faire des tests statistiques additionnels. Ces tests sont esquissés dans les chapitres portant sur l'inférence statistique (chapitres 8 et 9).

Quiz rapide 6.4
La relation entre le nombre d'heures de travail dans une journée et le nombre de minutes de travail dans cette journée est-elle parfaite ? Répondez à la même question pour le nombre d'heures de travail dans une journée et le nombre de dossiers résolus dans cette journée ?

Le coefficient de détermination

Le *coefficient de détermination* est une statistique très simple à calculer et très utile pour l'interprétation des corrélations. Le coefficient de détermination se calcule en mettant le coefficient de corrélation au carré puis en pourcentage. Les valeurs minimale et maximale du coefficient de détermination sont 0 et 100 %. C'est une statistique pratique qui indique, en pourcentage, le degré de relation existant entre deux variables.

$$\text{Coefficient de détermination} = r_{xy}^{\,2} \times 100\,\% \qquad \text{Formule 6.2}$$

Si :

- $r_{xy} = \pm 1$, alors le coefficient de détermination $= 1^2 \times 100\,\% = 100\,\%$;
- $r_{xy} = 0$, alors le coefficient de détermination $= 0^2 \times 100\,\% = 0\,\%$;
- $r_{xy} = \pm 0{,}50$, alors le coefficient de détermination $= 0{,}5^2 \times 100\,\% = 25\,\%$.

On peut remarquer qu'une corrélation de –0,50 ou +0,50 produit le même coefficient de détermination : 25 %. Le coefficient de détermination s'appelle aussi le *pourcentage de variance expliquée* ou le *pourcentage de variance partagée*. Le pourcentage de variance expliquée indique le degré avec lequel la connaissance de la variable X permet de réduire l'incertitude sur la variable Y.

Lorsque la corrélation est parfaite, le coefficient de détermination est de 100 %, et indique que la connaissance de la position relative de chaque observation sur X nous renseigne totalement sur la position relative de chaque observation sur Y. Lorsque la corrélation est égale à 0, le coefficient de détermination sera lui aussi égal à 0 %, et indique que la connaissance de X ne nous apprend rien au sujet de la variable Y.

Le coefficient de détermination est particulièrement utile dans le cas de corrélations imparfaites. Si la relation entre les années de scolarité et le salaire est de 0,56, alors le coefficient de détermination est de $0{,}56^2 \times 100\,\% = 31\,\%$.

Ainsi, la connaissance du niveau de scolarité explique ou réduit l'incertitude au sujet du salaire de 31 %. Ce coefficient nous offre donc une façon pratique d'interpréter l'ampleur de la relation entre les variables. Nous basant sur le coefficient de détermination pour la relation scolarité-salaire, nous pouvons ainsi conclure qu'avoir plus d'années de scolarité est relié à un meilleur salaire, mais que ce n'est pas le seul élément qui « explique » ce salaire.

Le coefficient de non-détermination

Prenons une corrélation de 0,50. Le coefficient de détermination est de 25 %, ce qui veut dire que la variable Y est « expliquée » à 25 % par l'autre variable (X). Mais quel est le niveau de non-relation entre les variables ? Dans ce cas, il existe 75 % de fluctuation dans une variable qui n'est pas lié à l'autre variable, et c'est ce qu'on appelle le *coefficient de non-détermination :*

Coefficient de non-détermination = $(1 - r_{xy}^2) \times 100\,\%$ Formule 6.3

Si :

- $r_{xy} = \pm 1{,}00$, le coefficient de non-détermination = $(1 - 1^2) \times 100\,\% = 0\,\%$;
- $r_{xy} = 0{,}00$, le coefficient de non-détermination = $(1 - 0^2) \times 100\,\% = 100\,\%$;
- $r_{xy} = \pm 0{,}50$, le coefficient de non-détermination = $(1 - 0{,}5^2) \times 100\,\% = 75\,\%$.

Si le coefficient de détermination indique dans quelle mesure la variable X explique la variable Y, le coefficient de non-détermination indique ce que nous n'expliquons pas.

Le coefficient de détermination, de non-détermination et la réduction de l'incertitude relative

Supposons qu'une personne est à l'intérieur d'un contenant scellé et climatisé et que ce contenant est déposé quelque part dans le monde. On demande à cette personne de deviner la température externe en degrés Celsius. Elle n'a aucune base rationnelle pour répondre, le contenant pou-

vant se trouver en Antarctique ou au milieu du Sahara. Nous pouvons alors dire que l'incertitude de cette personne quant à la température externe est au maximum, en l'occurrence à 100 %.

Dans l'espoir de réduire son incertitude (sur la température à l'extérieur du contenant), on lui indique la note obtenue à un examen de statistiques par un étudiant ! Nous savons que la relation entre les températures extérieures et les notes aux examens est $r_{xy} = 0{,}0$ et, par conséquent, que le coefficient de détermination est de 0 % et le coefficient de non-détermination est de 100 %. Où se situe maintenant le degré d'incertitude au sujet de la température ? L'information concernant la note de l'étudiant n'aide aucunement la personne à deviner la température externe. Cette information n'a pas réussi à réduire son incertitude. Le principe est : *lorsque la corrélation est nulle, une variable est incapable de réduire le degré d'incertitude au sujet d'une autre variable.*

Le coefficient de détermination et le concept de la réduction de l'incertitude sont très importants dans plusieurs situations concrètes. Supposons que nous savons qu'il existe une relation négative entre le niveau de soutien familial et le risque de suicide chez les jeunes (plus le soutien familial est fort, moins le risque de suicide est grand). Si nous voulions évaluer le risque de suicide chez une personne, nous pourrions examiner le niveau de soutien qu'elle reçoit ; ainsi, nous aurions une meilleure base pour évaluer son risque de suicide. Si la personne reçoit très peu de soutien, il y a lieu d'être plus inquiet que si le degré de soutien qu'elle reçoit est très fort.

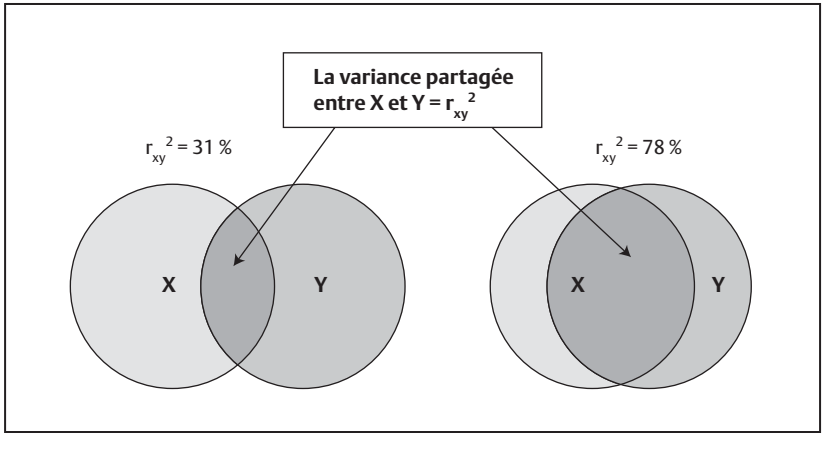
Revenons maintenant à cette personne toujours dans le contenant scellé à qui on demande de deviner la température en degrés Celsius. Mais, cette fois, on lui indique la température externe en Fahrenheit. Elle sait que la corrélation entre les degrés Fahrenheit et les degrés Celsius est parfaite : $r_{FC} = +1{,}0$. Quel serait maintenant son degré d'incertitude quant à la température en degrés Celsius ? La corrélation parfaite produit un coefficient de détermination de 100 % et, par conséquent, le coefficient de non-détermination est de 0 %. Dans ce cas, la connaissance de la température en Fahrenheit réduit *l'incertitude* au sujet de la température en degrés Celsius à 0 %. Cette personne peut maintenant sans erreur indiquer, en degrés Fahrenheit, la température qu'il fait à l'extérieur du contenant. Si on lui dit que la température externe est de 32 °F, elle sait sans le moindre risque d'erreur qu'il

fait 0 en degrés Celsius. Lorsque la corrélation est parfaite, le coefficient de détermination est égal à 100 %, réduisant le coefficient de non-détermination (et le degré d'incertitude) à 0 %. Le principe est donc qu'*au fur et à mesure que la corrélation (et le coefficient de détermination) augmente, l'incertitude se réduit.*

Représentation schématique de la corrélation et du coefficient de détermination

Le coefficient de détermination est un indice de la quantité de variances partagées par deux variables. Quand $r_{xy} = 0$, $r_{xy}^2 = 0\,\%$, nous pouvons dire que X et Y n'ont aucune variance en commun. À l'opposé, lorsque $r_{xy} = \pm 1{,}0$ $r_{xy}^2 = 100\,\%$, cela implique que ce que nous savons de X nous renseigne parfaitement sur Y. La Figure 6.5 schématise ce concept à l'aide d'un *diagramme de Ballantine.* La variance de chaque variable X ou Y prend la forme d'un cercle tandis que le coefficient de détermination est illustré par le degré de chevauchement des cercles. Le degré de chevauchement des deux variables (le coefficient de détermination) est plus fort (78 %) pour les cercles à droite dans la figure que pour ceux à gauche dans la figure (31 %).

FIGURE 6.5 Diagramme de Ballantine représentant schématiquement le pourcentage de variance partagée (r_{xy}^2)



REMARQUES SUPPLÉMENTAIRES

Corrélation et causalité

L'existence d'une corrélation entre X et Y n'indique pas un lien de causalité entre X et Y. Étudions la recherche suivante : tous les trois ans, une immense étude (Programme for International Student Assessment) examine la compétence dans plusieurs matières scolaires d'élèves âgés de 15 ans et résidant dans plus de 40 pays. Dans chaque pays, entre 4 500 et 10 000 élèves passent l'examen. Voici un des résultats obtenus par cette étude en 2003 : il existe une corrélation positive entre la présence d'un lave-vaisselle à la maison (la variable X) et la compétence en lecture, en mathématiques et en sciences (la variable Y). Les élèves qui ont un lave-vaisselle chez eux obtiennent de meilleures notes aux tests standardisés que ceux qui n'en ont pas ! Il existe au moins cinq explications pour ce résultat. Laquelle est exacte ? Peut-on en imaginer d'autres ?

1. La possession d'un lave-vaisselle entraîne la compétence dans ces matières (X cause Y).
2. L'obtention de meilleurs résultats dans ces matières cause l'achat d'un lave-vaisselle (Y cause X).
3. Il n'y a pas de réelle relation entre la présence d'un lave-vaisselle et la compétence des élèves, ce résultat n'étant qu'un accident statistique.
4. Les élèves qui ont un lave-vaisselle n'ont pas besoin de laver la vaisselle et, par conséquent, ils ont plus de temps (variable W) à consacrer à l'étude (X cause W qui, à son tour, cause Y).
5. Les élèves qui ont un lave-vaisselle vivent dans des familles plus riches (variable W) et, parce qu'elles sont plus riches, elles sont plus en mesure d'offrir à leurs enfants une meilleure éducation et de s'acheter un lave-vaisselle. Leur richesse se reflète dans leur performance scolaire et leurs électroménagers (W cause X et Y).

Basées simplement sur la corrélation, toutes ces explications sont possibles. Il est donc impossible d'apporter une conclusion sur la causalité à partir de la seule corrélation.

Cependant, supposons que les chercheurs n'ont trouvé aucune corrélation entre ces deux variables. Dans ce cas, nous pourrions affirmer que le fait de posséder un lave-vaisselle ne cause pas une amélioration des

résultats scolaires. *Ainsi, la présence d'une corrélation n'est pas forcément le signe d'un lien causal, mais l'absence de corrélation confirme l'absence de causalité !*

Corrélation de Pearson et variance des variables

La corrélation entre deux variables sera toujours de zéro lorsque la variance de l'une ou l'autre des variables est égale à zéro. Retournons au Tableau 6.4 et à la Figure 6.3. Toutes les personnes de la banque de données ont exactement la même valeur pour la variable « nombre de nez » et, par conséquent (voir le chapitre 3), la variance du nombre de nez est égale à zéro. Puisque la variance est égale à zéro, chaque personne de la distribution occupe exactement la même position sur la variable « nombre de nez » (c'est-à-dire Z = 0).

Quiz rapide 6.5
Calculez la variance de la variable « nombre de nez » du Tableau 6.5 en vous servant de la formule vue au chapitre 3. Expliquez pourquoi la corrélation entre « nombre de nez » et « tabagisme » est égale à zéro.

La corrélation indique le degré de similitude entre la position relative des observations sur une variable et la position relative de ces mêmes observations sur une autre variable. Au Tableau 6.4, la variable X (nombre de cigarettes fumées) présente de la variance alors que la variable Y (nombre de nez) n'en présente pas. Voyons maintenant si les personnes maintiennent la même position sur les deux variables. La personne A se situe à Z = +1,26 sur la variable X (tabagisme), mais elle se situe à Z = 0 sur la variable Y (nez). Sa position sur la variable X n'est pas maintenue sur la variable Y. La même conclusion s'impose pour presque toutes les observations. Puisque les personnes ne maintiennent pas la même position relative sur les deux variables, la corrélation est zéro. Autrement dit, si une des variables est constante (aucune variance), l'autre variable ne peut rien expliquer, et donc, il n'existe aucune corrélation. On peut aussi en arriver à la même conclusion en appliquant la formule pour la corrélation (Formule 6.1).

Nous pouvons maintenant élaborer un principe général. Plus la variance de l'une ou l'autre des deux variables est petite, plus la corrélation observée sera faible. À la limite, lorsque l'une ou l'autre des variables n'a pas de variance, la corrélation est invariablement égale à zéro.

Quiz rapide 6.6
On veut calculer la corrélation entre la taille des enfants et leur âge. On calcule cette corrélation sur deux groupes d'enfants. Le groupe A : les enfants âgés de 1 à 8 ans. Le groupe B : les enfants âgés de 6 et 7 ans. Pour quel groupe la corrélation a-t-elle le plus de chances d'être grande ?

Corrélation et observations loin de la moyenne

Les observations n'ont pas toutes la même influence sur la corrélation. La corrélation est plus influencée par les observations se trouvant loin de la moyenne que par celles qui lui sont proches. Au Tableau 6.5, nous avons obtenu une corrélation entre salaire et scolarité de $r_{xy} = +0{,}56$ ($r_{xy}^2 = 31\,\%$). Retirons les deux observations encerclées et recalculons la corrélation. Ces deux observations identifient des personnes qui ont une longue scolarité (19 années). Ces deux personnes se situent loin de la moyenne (pour la variable « scolarité »). La corrélation entre le salaire et la scolarité pour les observations restantes est $r_{xy} = 0,40$ ($r_{xy}^2 = 16\,\%$). Le coefficient de détermination est presque moitié moindre. Le retrait des deux seules observations loin de la moyenne a considérablement réduit la corrélation. En l'absence de ces deux observations, la réduction de l'incertitude chez Y (le salaire) à partir de X (la scolarité) est plus faible et il devient beaucoup plus difficile de prédire les salaires à partir du nombre d'années de scolarité. Remettons ces deux observations dans l'échantillon et, cette fois, retirons deux observations qui se trouvent près de la moyenne. La corrélation est maintenant $r_{xy} = +0{,}58$ ($r_{xy}^2 = 33\,\%$). Elle a très peu changé !

En somme, les observations se situant loin de la moyenne ont plus d'influence sur la corrélation que les observations se situant près de la moyenne. Voyons pourquoi. La corrélation se calcule à partir de $\sum(Z_{Xi} \times Z_{Yi}) / N - 1$. Ainsi, plus la quantité $\sum(Z_{Xi} \times Z_{Yi})$ est grande, plus la corrélation sera forte. Or, les valeurs qui se situent plus loin de la moyenne produisent des valeurs étalons Z qui sont plus grandes. Si on les retire, la quantité $\sum(Z_{Xi} \times Z_{Yi})$ sera

nettement plus petite. En conséquence, la corrélation chutera. À l'inverse, si on élimine deux observations proches de la moyenne, leurs valeurs Z étant proches de zéro, ce retrait ne réduira que légèrement $\sum(Z_{Xi} \times Z_{Yi})$ et, par conséquent, la corrélation changera peu.

Quiz rapide 6.7
La corrélation entre X et Y est forte. Supposons que l'on retire une observation qui se situe exactement à la moyenne de la variable X. Qu'adviendra-t-il de la corrélation XY ? Et si la corrélation XY était zéro, qu'arriverait-il si nous retirions une observation qui se trouve à la moyenne de X ?

L'impact des observations loin de la moyenne sur la corrélation n'est rien d'autre qu'un cas particulier du principe précédent selon lequel la corrélation est plus faible lorsque la variance des observations est plus petite. En effet, lorsque nous retirons des observations qui sont loin de la moyenne, les observations qui restent sont plus près les unes les autres. Par conséquent, la variance diminue.

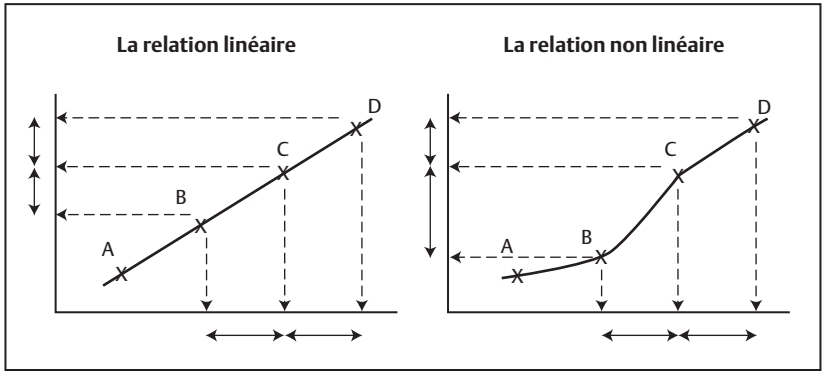
Corrélation de Pearson et relation linéaire

La corrélation de Pearson mesure le degré de linéarité dans la relation entre deux variables. Une *relation linéaire* implique que la taille de l'accroissement ou de la décroissance des valeurs Y est la même pour chaque accroissement ou décroissance de la variable X. La Figure 6.6 clarifie cette idée. Dans le graphique de gauche, nous avons quatre observations. La distance sur l'axe X entre les observations B et C et entre les observations C et D est la même. Voyons maintenant les distances pour ces mêmes observations le long de l'axe Y. Ici encore, la distance B-C est égale à la distance C-D. Chaque accroissement le long de la variable X est accompagné d'un accroissement constant sur la variable Y. La relation est linéaire.

Le graphique de droite de la Figure 6.6 présente, par contraste, une relation non linéaire. Les distances entre les observations sur la variable X sont égales. Cela n'est pas le cas pour les mêmes observations le long de l'axe Y. Ainsi, sur l'axe Y, la distance entre B et C est plus grande que celle entre C et D. Chaque accroissement le long de la l'axe X est accompagné d'un accroissement qui n'est pas constant sur l'axe Y. La relation n'est pas linéaire.

La corrélation de Pearson n'est pas une statistique appropriée lorsque les relations sont non linéaires. Pour cela, il faut faire appel à d'autres types de corrélations (par exemple au ratio de corrélation, la statistique η, êta).

FIGURE 6.6 Représentation graphique d'une relation linéaire et d'une relation non linéaire



Une façon pratique de présenter une corrélation : le tableau des attentes.

La corrélation est un indicateur statistique qui n'est pas toujours facilement compris par les non-statisticiens. Il importe, dans certaines situations, de présenter les résultats d'une analyse de corrélation de manière plus simple. Le tableau des attentes s'avère l'outil idéal pour ce faire.

Examinons la situation suivante : dans un grand centre d'appel, la performance au travail des employés est déterminée par le nombre de clients qu'ils servent dans une journée. Le centre désire améliorer le système qu'il utilise pour faire la sélection des futurs employés, c'est-à-dire choisir ceux qui pourront servir plus de clients..

La vice-présidente du centre demande à un chercheur de développer un nouveau système de sélection des candidats, ce qu'il fait en élaborant un test pour mesurer l'aptitude au travail. La mesure de l'aptitude est la variable X. La performance au travail est la variable Y. Le chercheur émet l'hypothèse que les personnes qui obtiendront les valeurs les plus élevées au test s'avéreront les plus productives au travail. Il entend la vérifier en calculant la corrélation existant entre la mesure de l'aptitude (X) et la performance au travail (Y).

Pour vérifier son hypothèse, il choisit aléatoirement 180 personnes déjà en poste. Les dossiers de la compagnie lui fournissent leur performance au travail : le nombre moyen de clients que chacune a servis, chaque jour, au cours du dernier mois. Cette mesure varie entre 20 et 80. Il administre le test de l'aptitude au travail à ces 180 employés. Il obtient donc, pour chacun, deux informations : sa performance au travail (Y) et sa performance au test d'aptitude (X). Il vérifie l'hypothèse en calculant la corrélation XY et trouve qu'elle est positive et substantielle : $r_{xy} = 0{,}58$. Le chercheur détient maintenant une preuve que la performance au test d'aptitude est positivement liée à la performance au travail. Ainsi, ceux qui démontrent la plus grande aptitude (telle que mesurée par le test) tendent à être plus productifs. On peut aussi affirmer que le fait de connaître l'aptitude au travail (X) réduit l'incertitude quant à l'éventuelle performance au travail (Y).

Techniquement, le chercheur a exécuté une étude de validité concomitante. Pour ce genre d'étude, une corrélation de 0,58 est considérée très substantielle et les psychométriciens diraient que le test est une mesure « valide » de la performance au travail.

Il lui faut maintenant communiquer le résultat de son étude à la vice-présidente du centre. Elle n'est pas statisticienne et une corrélation de 0,58 ne lui dira pas grand-chose. Le chercheur choisit alors de lui présenter la corrélation XY qu'il a obtenue dans un *tableau des attentes.*

Un tableau des attentes est une matrice à double entrée que le chercheur construit de la manière suivante : il divise les employés qui ont participé à son étude en trois groupes de 60 personnes chacun. Dans le groupe 1, qu'il étiquette « Performance faible », il place les personnes qui se situent dans le tiers inférieur de la distribution de la performance au travail. Il place dans le groupe 3, « Performance élevée », les personnes qui se situent dans le tiers supérieur de la performance au travail. Toutes les autres, le tiers de son échantillon qui se situe au milieu de la distribution de la performance au travail, sont placées dans le groupe 2, « Performance moyenne ».

Les notes obtenues au test d'aptitude varient entre 20 et 80. Le chercheur divise alors les performances au test en trois groupes : le groupe 1, « Aptitude faible », inclut les employés qui ont obtenu 39 ou moins au test. Ceux qui obtiennent 60 ou plus forment le groupe 3 : « Aptitude élevée ». Les autres, ceux qui ont obtenu entre 40 et 59, forment le groupe 2 : « Aptitude

moyenne ». Le Tableau 6.7 montre les données observées. Notons que, dans ce tableau, 60 employés se classent dans le groupe « Aptitude faible », et 56 et 64 employés, respectivement, dans les groupes « Aptitude moyenne » et « Aptitude élevée ».

Ensuite, le chercheur identifie pour chaque groupe de performance au test d'aptitude le nombre de personnes qui ont une performance au travail faible, moyenne ou élevée. Nous notons au Tableau 6.7 que, des 60 personnes qui ont obtenu un faible résultat au test d'aptitude, 45, 14 et 1 font respectivement partie des groupes de performance faible, moyenne et élevée (rangée 1 du tableau). Nous pouvons maintenant exprimer ces résultats en pourcentages (indiqués sous une forme probabiliste entre parenthèses dans le tableau). Ainsi, nous voyons que 75 % des personnes qui ont obtenu un faible résultat au test démontrent une faible performance au travail, et que seulement 2 % des personnes qui démontrent un faible niveau d'aptitude présentent un niveau élevé de performance au travail. Environ le quart (23 %) des personnes qui ont obtenu un faible résultat au test fournissent une performance moyenne au travail. En interprétant ces pourcentages en termes probabilistes, nous pouvons conclure que celles qui ont obtenu un résultat faible au test ont une très faible probabilité (p = 0,02) de fournir une forte performance au travail, une probabilité intermédiaire (p = 0,23) d'être moyennement productives et une très forte probabilité de fournir une piètre performance au travail (p = 0,75).

Tableau 6.7
Le tableau des attentes

	Performance au travail (Y)			
Aptitude (X)	***Grp 1 (Faible)***	***Grp 2 (Moyenne)***	***Grp 3 (Élevée)***	***Total***
Faible	45 (0,75)	14 (0,23)	1 (0,02)	60
Moyen	13 (0,23)	29 (0,52)	14 (0,25)	56
Élevé	2 (0,03)	17 (0,27)	45 (0,70)	64
TOTAL	60	60	60	180

Nous procédons à ces analyses pour chaque rangée du tableau des attentes. Prenons la troisième rangée de données du Tableau 6.7 par exemple : des 64 personnes qui ont démontré une forte performance au test, 3 % (2/64) sont peu productives, 27 % (17/64) sont moyennement productives et 70 % (45/64) sont très productives. En exprimant ces pourcentages en termes probabilistes, nous pouvons conclure que les personnes qui réussissent très bien le test (aptitude élevée ; 60 et plus) présentent une très forte probabilité ($p = 0{,}70$) d'être des employés très productifs (groupe 3) et une très faible probabilité ($p = 0{,}03$) de fournir une piètre performance au travail.

Si la vice-présidente décide d'administrer ce test d'aptitude aux postulants, nous pourrons constater, en consultant le tableau des attentes, qu'il serait préférable de ne pas embaucher le candidat qui obtiendra un score faible (< 40) au test car il présentera une faible probabilité de fournir une prestation de travail exceptionnelle ($p = 0{,}02$) et une très forte probabilité ($p = 0{,}75$) de ne pas être performant. Mais s'il obtenait plus de 59, sa probabilité de devenir un employé très productif serait très forte ($p = 70$) et il serait alors pertinent de l'embaucher.

De fait, le tableau des attentes ne sert qu'à reproduire, en termes qu'il est plus facile de comprendre et de mettre en pratique, l'information déjà établie par la corrélation : plus grande est l'aptitude d'une personne, plus élevée sera sa performance au travail.

Une question pourrait maintenant vous venir en tête : si le tableau des attentes est une façon pratique et simple de montrer la corrélation entre deux variables, pourquoi avons-nous calculé la corrélation (et vous avoir fait étudier un chapitre complet sur le sujet) ? La réponse nous ramène à la discussion du chapitre 1 portant sur les échelles de mesures. La mesure de l'aptitude et celle de la performance au travail du Tableau 6.7 sont des échelles à intervalles. Le tableau des attentes a traduit ces variables en échelles catégorielles (nominales).

Comme nous l'avons vu au chapitre 1, la conversion d'une échelle à intervalles en une échelle nominale réduit la précision des données. Ainsi, la catégorie « Aptitude faible » englobe, à la fois, la personne qui a obtenu 20 au test et celle qui a obtenu 39, et considère que cette dernière a fourni une performance très différente d'une autre personne qui aurait obtenu 40, seulement un point de plus. Ainsi, la catégorisation occasionne une perte

d'information importante. Dans le chapitre suivant, nous allons étudier une autre technique, la régression simple, qui nous permet de faire le même genre de prédiction sans convertir les données en variables nominales. Mais, pour cela, il vous faudra apprendre et comprendre d'autres techniques statistiques !

SOMMAIRE DU CHAPITRE

La corrélation de Pearson est un indice statistique qui indique le degré de similitude entre la position des observations sur une variable et la position de ces mêmes observations sur une deuxième variable. Elle se limite à indiquer le degré de relation linéaire qui existe entre deux variables qui sont mesurées avec des échelles à intervalles ou des échelles de rapport. La corrélation prend des valeurs allant de 0 à ±1,0. La relation est parfaite lorsqu'elle est égale à ±1,0 et elle se réduit au fur et à mesure qu'elle se rapproche de 0. Le signe de la corrélation indique la direction de la corrélation. La corrélation et ses coefficients de détermination et de non-détermination sont utilisés pour interpréter dans quelle mesure la connaissance d'une variable réduit l'incertitude face à une deuxième variable. La corrélation de Pearson est influencée par les valeurs se situant loin de la moyenne, et est relativement peu influencée par les observations se situant près d'elle. Enfin, la présence d'une corrélation n'indique pas nécessairement la présence d'un lien causal entre les variables. Mais l'absence de corrélation indique une absence de causalité.

La corrélation fait partie de statistiques descriptives très utilisées en sciences humaines et en sciences sociales, principalement parce qu'elle indique dans quelle mesure la connaissance d'une variable nous renseigne au sujet d'une seconde variable. Enfin, il est possible de présenter la corrélation entre deux variables sous une forme plus simple, le tableau des attentes.

EXERCICES DE COMPRÉHENSION

1. Nous calculons la corrélation entre deux variables X et Y. La variable X est une constante. La corrélation sera alors de _____ .
   a) +1,0
   b) –1,0
   c) 0,0
   d) n'importe quelle valeur entre –1 et +1
2. Les personnes qui se situent au-dessus de la moyenne sur la variable X se situent au-dessus de la moyenne sur la variable Y. Nous voyons aussi que toutes les personnes qui se situent au-dessous de la moyenne sur X se situent au-dessous de la moyenne sur Y.
   La corrélation entre X et Y sera _____.
   a) positive
   b) négative
   c) aux alentours de zéro
   d) impossible à déterminer avec les informations fournies
3. Nous trouvons une corrélation de zéro entre X et Y. Pourquoi ?
   a) La variable X ou la variable Y est une constante.
   b) La position relative des observations sur X ne correspond en rien à leur position sur Y.
   c) La relation n'est pas linéaire.
   d) Toutes ces réponses peuvent être justes.
4. La corrélation entre le nombre d'enfants par famille et la richesse des parents est fortement négative. Dans un parc, nous observons deux familles ; la famille A a 6 enfants, alors que la famille B n'en a qu'un seul. Il est probable que _____.
   a) la famille A soit plus riche que la famille B
   b) la famille B soit plus riche que la famille A
   c) la famille A soit aussi riche que la famille B
   d) Toutes ces réponses sont également probables.
5. Nous remarquons une corrélation positive très élevée entre le nombre de voitures dans les villes et le nombre de citoyens de ces villes qui sont atteints de troubles respiratoires. Laquelle de ces affirmations est vraie ?

a) Les gens qui ont des troubles respiratoires achètent plus de voitures.
b) Les voitures étant une source de pollution, elles causent beaucoup de troubles respiratoires.
c) Les personnes qui ont des voitures font moins d'activité physique, ce qui leur occasionne des troubles respiratoires.
d) Toutes ces réponses sont possibles.

6. Nous étudions la relation entre le stress et la performance au travail. Nous observons que les personnes qui sont très peu stressées performent très mal, mais au fur et à mesure que leur degré de stress augmente, leur performance s'améliore jusqu'à un certain point. Par contre, à partir du moment où leur stress dépasse ce point, leur performance se dégrade rapidement. La relation entre stress et performance est _____, et la corrélation de Pearson sera _____.

a) linéaire ; positive
b) linéaire ; positive
c) non linéaire ; proche de zéro
d) non linéaire ; soit positive, soit négative, mais pas zéro

7. Pour le même groupe d'enfants, nous mesurons le quotient intellectuel aussi bien que la performance scolaire. Nous exprimons les valeurs pour ces deux variables en valeurs étalons Z. Pour chacun des élèves, nous calculons la différence entre la valeur Z de son QI et la valeur Z de sa performance scolaire. Pour chacun des élèves, cette différence est égale à zéro. Nous calculons la corrélation entre les deux variables, QI et succès scolaire. La corrélation r_{xy} = _____.

a) +1
b) -1
c) 0
d) n'importe quelle valeur entre –1 et +1

8. Nous créons un diagramme de dispersion pour la relation entre les variables X et Y. Une personne se trouve à la coordonnée (100 ; 3,7). Cette personne a obtenu la valeur _____ pour X et la valeur _____ pour Y.

9. La corrélation entre X et Y est de 0,60. En connaissant X, nous pouvons réduire l'incertitude sur la variable Y de _____ %.

Réponses

1. c
2. a
3. d
4. b
5. d
6. c
7. a
8. 100 et 3,7
9. 36

Page laissée blanche

CHAPITRE 7
LA RÉGRESSION LINÉAIRE SIMPLE

Page laissée blanche

CHAPITRE 7

LA RÉGRESSION LINÉAIRE SIMPLE

La corrélation est un indice de la relation générale qui existe entre deux variables. Son calcul indique dans quelle mesure la connaissance d'une variable réduit l'incertitude que nous avons au sujet d'une deuxième variable. Nous nous tournons maintenant vers une application pratique de la corrélation : la *régression linéaire simple*.

La régression simple est une technique statistique qui se sert de la corrélation entre X et Y pour prédire ou estimer la position inconnue d'une observation spécifique sur la variable *Y* à partir de la connaissance que nous avons de sa position sur la variable *X*. Cette technique est fort utile par exemple lorsqu'il s'agit de prédire la performance au travail d'une personne à partir de notre connaissance de ses expériences antérieures, la température du lendemain à partir de notre connaissance des courants d'air ou les probabilités de réussite d'un étudiant à son examen dans une matière précise à partir de sa réussite scolaire générale.

Supposons qu'on sache que la note moyenne obtenue à un cours de statistiques est de 70 %, en se basant sur les résultats obtenus dans ce cours les années précédentes. On veut prédire la note que deux étudiants, Jean et Jeanne, obtiendront dans ce cours et ce, avant même qu'ils ne le suivent. En l'absence d'autres informations, la meilleure estimation qu'on a de la note de Jean et de celle de Jeanne est la moyenne généralement attribuée dans ce cours, en l'occurrence 70 % (la moyenne étant la meilleure estimation de chacune des valeurs d'une distribution ; voir le chapitre 3). En nous basant sur cette moyenne générale au cours (70 %), nous prédisons que Jean et Jeanne obtiendront la même note, soit la moyenne historique,

70 %. Supposons maintenant que nous détenons une information supplémentaire : les étudiants qui ont une moyenne générale élevée ont tendance à obtenir des notes supérieures à la moyenne dans le cours de statistique et ceux qui ont une moyenne générale faible obtiennent généralement des notes faibles dans ce cours. Autrement dit, il existe une corrélation positive entre la moyenne générale et la note dans le cours de statistiques.

Jeanne a une très forte moyenne générale alors que Jean a une moyenne générale très faible. Allons-nous prédire la même note de 70 % (la moyenne historique) pour les deux étudiants ? Sachant que les étudiants qui sont plus forts en général, comme l'est Jeanne, tendent à mieux réussir leur cours de statistiques, on aurait raison d'anticiper (de prédire) que cette étudiante obtiendra plus que 70 % dans son cours. Quant à Jean, qui obtient des notes générales beaucoup plus faibles, nous pouvons prédire qu'il obtiendra probablement une note plus faible que 70 %.

Au lieu de prédire la même note (70 %) pour les deux étudiants, la prédiction est maintenant différenciée et plus précise. Nous nous sommes servis de la relation historique qui existe entre une première variable (la moyenne générale, la variable X) et une deuxième variable (la note dans le cours de statistiques, la variable Y) afin de faire une prédiction de la variable Y, alors que nous connaissons la position de ces personnes seulement sur la variable X. La régression linéaire est la technique statistique qui permet de faire ce genre de prédictions.

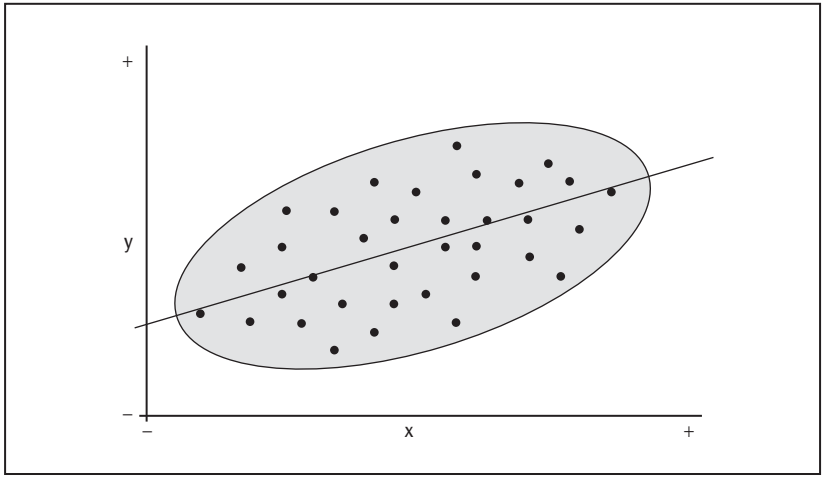
LE GRAPHIQUE DE DISPERSION ET LA DROITE DE RÉGRESSION

On se souvient, tel qu'il a été vu à partir du chapitre 6, que le graphique de dispersion — ou le nuage de points — présente pour chaque observation sa position simultanée sur deux variables, c'est-à-dire ses coordonnées. La Figure 7.1 présente un tel graphique de dispersion avec des données fictives. On y remarque que la position des points révèle une tendance dans les observations : celles qui sont plus fortes sur X (plus proches du côté positif de l'abscisse) tendent aussi à être plus fortes sur Y (plus proches du côté positif de l'ordonnée). Ainsi (voir chapitre 6), lorsque les valeurs tendent à être similaires (plus positives ou plus négatives) sur les deux variables, la

corrélation entre les deux variables sera positive. Dans le cas de la Figure 7.1, la corrélation de Pearson est, par conséquent, positive. Elle est $r_{xy} = +0{,}60$.

Traçons une ligne droite qui représente cette tendance. Cette ligne, indiquée en noir, est *la droite de régression*. La droite de régression décrit la tendance générale indiquée par le graphique de dispersion. Puisque la corrélation est la tendance générale, nous disons alors que la droite de régression est une représentation graphique de la corrélation qui existe entre X et Y. On remarque la position de cette droite de régression par rapport au nuage de points. Elle est très importante et nous allons y revenir.

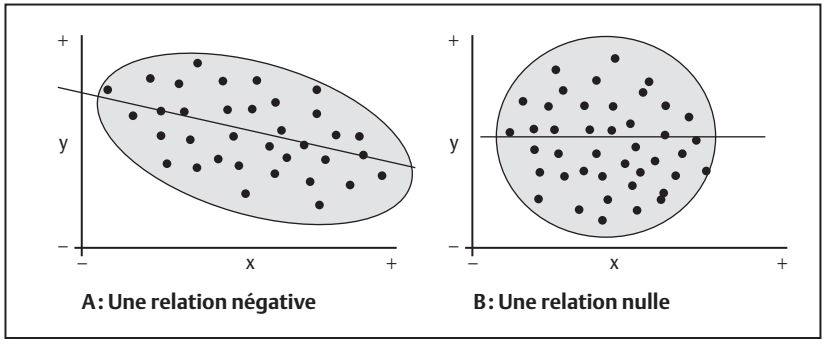
FIGURE 7.1 Graphique de dispersion représenté par un ovale et la corrélation positive par une droite de régression



Étudions maintenant la Figure 7.2. Les graphiques montrent les nuages de points et leurs droites de régression. Le Graphique A montre une relation négative ($r_{xy} = -0{,}60$) alors que le Graphique B montre une relation nulle ($r_{xy} = 0$).

- Graphique A : les valeurs fortes sur X tendent à être associées avec des valeurs faibles sur Y et les valeurs faibles sur X tendent à être associées avec des valeurs fortes sur Y. La droite de régression montre cette tendance négative. La corrélation XY est négative (environ $r_{xy} = -0{,}60$).

FIGURE 7.2 Nuages de points et droites de régression pour deux relations



- Graphique B : les valeurs fortes sur la variable X ne semblent pas être systématiquement associées ni avec des valeurs fortes ni avec des valeurs faibles sur Y ; cela semble aussi être le cas pour les valeurs faibles sur X. La droite de régression ne montre ni une tendance négative ni une tendance positive. La corrélation XY est proche de zéro.

En comparant les Figures 7.1 et 7.2, on remarque l'angle entre la droite de régression et l'abscisse. On appelle cet angle *la pente*. Lorsque la corrélation s'approche de zéro, la droite de régression s'approche d'une ligne horizontale, parallèle à l'abscisse : sa pente, par rapport à l'abscisse, est égale à zéro. Mais lorsque la corrélation augmente, la pente s'éloigne de l'horizontale. L'angle de la droite de régression correspond à la magnitude de la corrélation.

La droite de régression indique aussi et simultanément la direction de la relation. Avec une relation négative, la droite de régression tombe avec un accroissement des valeurs sur X tandis qu'avec une relation positive, la droite de régression augmente avec un accroissement des valeurs sur X.

Quelques conventions

En régression linéaire, on utilise une variable pour en prédire une autre. Par convention, la variable que l'on veut prédire est la variable dépendante qui est généralement identifiée par la lettre *Y*. La variable utilisée pour faire

cette prédiction, la variable indépendante, est généralement identifiée par la lettre *X*. Lorsqu'on veut prédire la distance parcourue en kilomètres à partir de la distance parcourue en milles, la variable dépendante *Y* est le nombre de kilomètres parcourus et la variable indépendante *X* est le nombre de milles parcourus. Si l'on veut faire l'inverse, la distance en kilomètres devient la variable indépendante *X* et la distance en milles devient la variable dépendante *Y*. Ainsi, la notion de variable indépendante ou dépendante est totalement déterminée par l'analyste.

Quiz rapide 7.1
Vous désirez prédire le degré de pollution dans les villes nord-américaines en fonction du nombre de voitures enregistrées dans chaque ville. Quelle est la variable indépendante et quelle est la variable dépendante ? Et si vous vouliez prédire le nombre de voitures à partir du degré de pollution, quelle serait chacune de ces deux variables ?

Les statistiques de la régression linéaire

Le Tableau 7.1 présente plusieurs températures mesurées en Celsius (variable *X*) et en Fahrenheit (variable *Y*). La corrélation entre les températures mesurées en Fahrenheit et en Celsius (r_{xy} = 1,0) est parfaite et positive, et le coefficient de détermination est de 100 %. Si l'on a comme information la température en Celsius, on est en mesure d'estimer la température en Fahrenheit avec une précision totale parce que la corrélation entre ces deux échelles de température est parfaite.

Tableau 7.1
Températures en Celsius et en Fahrenheit

Celsius	–40	–30	–20	–10	0	10	20	30	40
Fahrenheit	–40	–22	–4	14	32	50	68	86	104

La Figure 7.3 décrit le nuage de points pour la relation entre Celsius et Fahrenheit ainsi que sa droite de régression. À présent, regardons le point situé le long de l'abscisse qui décrit 0 °C. On peut tracer une ligne verticale qui part de 0 °Celsius sur l'abscisse et qui se prolonge jusqu'à la droite

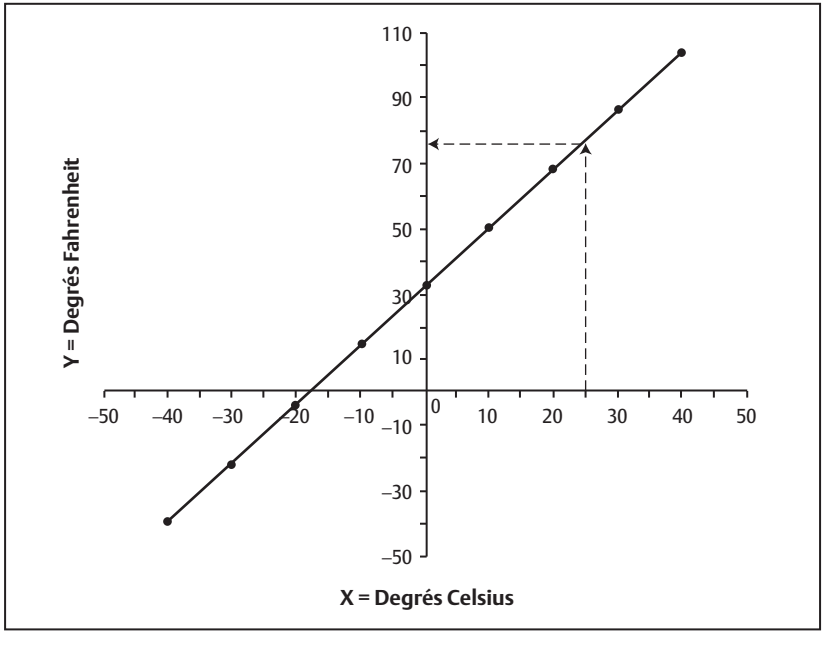
de régression. À partir du point où cette ligne verticale coupe la droite de régression, on trace une ligne horizontale que l'on prolonge jusqu'à l'ordonnée, qui représente les températures en Fahrenheit. La température en Fahrenheit indiquée par cette ligne verticale est 32. Cela nous indique qu'une valeur de 0 °C correspond à 32 °F. On s'est donc servis de la droite de régression pour faire une « prédiction » de la température (inconnue) en Fahrenheit à partir d'une température connue (dans ce cas 0 °C). En consultant les données du Tableau 7.1, on voit qu'effectivement 0 °C correspond à 32 °F.

Quiz rapide 7.2
Faites une autre prédiction en utilisant la droite de régression de la Figure 7.3. Prédisez la température en Fahrenheit lorsqu'il fait 30 °C. Vérifiez le résultat que vous obtenez au Tableau 7.1.

Supposons que l'on désire prédire la température en Fahrenheit pour 25 °C. Cette température n'est pas dans la banque de données (voir Tableau 7.1). On utilise donc la Figure 7.3 pour estimer la réponse. Nous trouvons 25 °C le long de l'abscisse et nous traçons la ligne verticale jusqu'à la droite de régression. À partir de l'intersection de cette ligne et de la droite de régression, nous traçons une ligne horizontale et nous lisons le point où l'ordonnée est coupée. Cette valeur est 77 °F. Ainsi, nous avons « prédit » une valeur inconnue (77 °F) à partir d'une valeur connue (25 °C) grâce à notre connaissance de la corrélation générale entre degrés Celsius et degrés Fahrenheit et en faisant appel à la droite de régression. Nous venons de mettre en pratique la régression linéaire.

Comme on peut l'avoir remarqué en travaillant avec la Figure 7.3, la validité de la prédiction dépend totalement de la droite de régression. Cette ligne doit être à la bonne place. Mettons-la à la mauvaise place : on décale la droite de régression de la Figure 7.3 en la faisant glisser un ou deux centimètres plus bas de manière à ce qu'elle coupe l'ordonnée à la valeur de –10 plutôt qu'à la valeur de +32. On remarque alors que les prédictions deviennent complètement fausses. Par exemple, pour une température de 0 °C, nous prédisons une température de –10 °F, ce qui est faux.

FIGURE 7.3 Droite de régression : température en degrés Fahrenheit en fonction de la température en degrés Celsius (r_{xy} = +1,0)



Quiz rapide 7.3
Utilisez la Figure 7.3 pour déterminer quelle est la température en Celsius pour une température de –30 °F. Si votre réponse n'est pas celle inscrite au Tableau 7.1, pouvez-vous en expliquer la raison ?

Donc, un des défis consiste ici à nous assurer que la droite de régression coupe l'ordonnée à la bonne place. Le point d'intersection entre la droite de régression et l'ordonnée se nomme *ordonnée à l'origine*. Comme nous le verrons bientôt, nous aurons besoin d'une formule statistique pour définir correctement cette valeur. Pour l'instant, il est suffisant de noter que l'ordonnée à l'origine est définie par la valeur que la variable dépendante Y atteint lorsque la valeur de la variable dépendante X est de 0. Dans la Figure 7.3, l'ordonnée à l'origine est +32, car il s'agit de la température (Y) en Fahrenheit lorsque la température en Celsius (X) est égale à 0. Mais ce n'est pas tout. Il faut aussi déterminer correctement la pente de la droite de régression pour produire des prédictions justes.

Lorsque la corrélation XY est nulle, la droite de régression est parallèle à l'abscisse. Au fur et à mesure que la corrélation XY s'éloigne de zéro, l'angle que la droite de régression fait avec l'abscisse augmente. Lorsque la corrélation est parfaite, la droite de régression coupe l'abscisse à un angle de 45 degrés. La pente de la droite de régression prend le nom particulier de *coefficient de régression non standardisé* ou, plus simplement, de *coefficient de régression* que, par convention, nous représentons par la lettre *b*. Le coefficient de régression et la corrélation bivariée (voir le chapitre 6) sont étroitement liés, les deux représentant le degré et la direction de la relation entre les variables X et Y. Ainsi, le positionnement de la droite de régression exige que deux calculs soient exécutés : un pour déterminer l'ordonnée à l'origine, l'autre pour déterminer le coefficient de régression, ce dernier se calculant à partir de la corrélation qui existe entre X et Y.

Revenons à la Figure 7.2. On remarque que la pente de la droite de régression est 0 pour la relation nulle (Graphique B), ce qui indique que son coefficient de régression est b = 0. Le coefficient de régression b est plus grand que 0 dans le Graphique A parce que la corrélation est plus grande que 0. Au fur et à mesure que la corrélation se réduit, la pente de la droite de régression et sa valeur numérique (le coefficient b) se réduisent pour atteindre 0 lorsque la corrélation est nulle. Le coefficient de régression prend toujours le signe de la corrélation. Lorsque la corrélation est négative, le coefficient de régression est négatif aussi. Nous pouvons donc prévoir que lorsque la corrélation entre X et Y est égale à 0, le coefficient de régression b sera lui aussi égal à 0. Mais, comme nous le verrons bientôt, la valeur maximale du coefficient de régression b n'est pas 1,0, comme c'est le cas pour la corrélation. La taille maximale du coefficient dépend de l'échelle de mesure des variables originales.

Déterminer la position de la droite de régression

Le problème central en régression consiste à déterminer correctement la position de la droite de régression. Une droite de régression est déterminée par deux éléments : son ordonnée à l'origine et son coefficient de régression. Il nous faudra donc calculer ces deux statistiques. Par convention, les statisticiens identifient l'ordonnée à l'origine par le symbole *a* (parfois aussi

β_0) et ils utilisent le symbole *b* (ou β – la lettre grecque bêta) pour identifier le coefficient de régression[1].

Comment quantifier les coefficients b et a ? C'est ce que nous allons voir en présentant les formules de calcul dans un premier temps puis en les expliquant afin de saisir leur signification.

Le calcul du coefficient de régression b

La Formule 7.1 indique comment calculer la statistique du coefficient de régression linéaire b. Ce coefficient de régression reflète la relation qui existe entre X et Y.

$$b = r_{xy} \times \frac{s_Y}{s_X} \qquad \text{Formule 7.1}$$

où b est le coefficient de régression, r_{xy} est la corrélation entre *X* et *Y*, s_Y est l'écart-type de la variable *Y* et s_X est l'écart-type de la variable *X*.

Pour les données du Tableau 7.1, la corrélation entre la variable *X* (degrés Celsius) et la variable *Y* (degrés Fahrenheit) est r_{xy} = 1,0. L'écart-type pour la variable *X* est 27,4 et 49,3 pour la variable *Y*.

Le calcul du coefficient b donne

$$\begin{aligned} b &= r_{xy} \times \frac{s_Y}{s_X} \\ &= +1{,}0 \times 49{,}3 \,/\, 27{,}4 \\ &= +1{,}0 \times 1{,}8 \\ &= +1{,}8 \end{aligned}$$

Le coefficient de régression b indique dans quelle mesure les valeurs de la variable Y changent en fonction de chaque changement de valeurs chez X. Pour les données du Tableau 7.1, nous avons trouvé b=+1,8, indiquant que chaque augmentation de 1 °C équivaut à une augmentation de 1,8 °F. On peut remarquer que le coefficient de régression est toujours du même signe que le coefficient de corrélation. Si le coefficient b avait été –1,8, nous aurions conclu que chaque augmentation de 1 °C correspondrait à une réduction de 1,8 °F.

1. À la fin du chapitre, nous verrons la différence entre le coefficient b et le coefficient β (bêta).

La Formule 7.1 permet deux constats supplémentaires :

a) Lorsque la corrélation est égale à zéro, le coefficient b est lui aussi égal à zéro.
b) Lorsque les écarts types de *Y* et de *X* sont égaux, le coefficient de régression b se réduit au coefficient de corrélation.

Mais, avant de produire une prédiction valide de la valeur Y à partir de la valeur X, il faut prendre en considération l'ordonnée à l'origine. Pour l'instant, explorons le coefficient de régression b.

L'explication du coefficient de régression b

Le coefficient de régression se calcule à partir de deux éléments : la corrélation (r_{xy}) et le rapport entre les écarts types des deux variables (s_Y/s_X).

1) La corrélation indique le degré avec lequel les valeurs de X et celles de Y correspondent lorsque les deux valeurs sont exprimées en valeurs étalons Z. Puisque les valeurs étalons Z sont des valeurs standardisées, la corrélation est une statistique qui exprime la relation entre les variables *X* et *Y* en valeurs standardisées. Lorsque la corrélation est parfaite (par exemple +1), chaque valeur étalon Z_X correspond exactement à la valeur étalon Z_Y pour chaque observation. Lorsque la corrélation est égale à 0,0, chaque valeur Z_x peut correspondre à n'importe quelle valeur Z_y. La droite de régression indique la relation entre X et Y. Donc, la droite de régression doit prendre en considération la corrélation X et Y.

Puisque la corrélation est construite avec des valeurs standardisées, les conclusions auxquelles elle conduit ne peuvent être que des conclusions en valeurs standardisées. Mais en régression linéaire, nous voulons prédire la valeur de la variable Y (à partir de X) en valeur brute et non pas en valeur standardisée. Si un Californien (qui ne comprend pas les degrés Celsius) demande la température qu'il fait à Montréal (20 °C), il ne serait pas très utile de lui dire qu'il fait Z = +0,73 ou b = 1,8 ! On voudra lui répondre en degrés Fahrenheit, l'échelle qu'il comprend. Pour en arriver à la réponse, il faut calculer l'équivalent en Fahrenheit de la température en Celsius. C'est pour résoudre ce problème que la deuxième partie (s_Y/s_X) de la formule pour le coefficient de régression b existe.

2) Le rapport s_Y/s_X est une correction arithmétique qui permet d'exprimer la corrélation en valeurs brutes (non standardisées). Les valeurs brutes correspondent aux chiffres qui sont utilisés pour mesurer les variables X et Y.

L'écart-type est construit à partir des valeurs brutes d'une variable. Le rapport entre les deux écarts types (s_Y/s_X) reflète donc le rapport entre les valeurs brutes pour les deux variables, X et Y. La taille de l'écart-type est directement tributaire de la taille des chiffres qui sont utilisés pour mesurer les variables. Par exemple, si on mesure un salaire (qui peut varier entre 0 et 200 000 $), l'écart-type prendra une valeur numérique chiffrée en milliers. Mais si on mesure des années de scolarité (qui varient entre 0 et 24), l'écart-type prendra des valeurs numériques chiffrées en dizaines. Si nous calculons l'écart-type des valeurs X (10, 20, 30), on verra que cet écart-type est 10 fois plus grand que l'écart-type lorsque X se mesure avec les chiffres 1, 2, 3.

Supposons la corrélation r_{xy} =+1,0, que la variable *X* prend les valeurs 1, 2, 3 et que les valeurs correspondantes pour *Y* sont 2, 4 et 6. Nous voyons alors que chaque changement de 1 unité sur *X* (de 1 à 2 et de 2 à 3) correspond à un changement de 2 unités sur *Y* (de 2 à 4 et de 4 à 6). Le coefficient de régression b le reflète et prend la valeur b = 2 : chaque changement de 1 unité sur X correspond à un changement de 2 unités sur Y. Si les changements sur *X* sont de 1 unité (1 à 2, 2 à 3) et les changements correspondants sur *Y* sont de 10 unités (de 10 à 20 et de 20 à 30), le coefficient b est de 10. Mais si on suppose qu'il n'existe pas de variance sur la variable Y : tous obtiennent la même valeur sur Y (la corrélation $r_{xy} = 0$, voir le chapitre 6), les changements sur X ne sont associés à *aucun* changement sur Y. Le coefficient b est alors égal à zéro. Ainsi, le rapport (s_Y/s_X) est une correction qui permet d'exprimer les valeurs de la corrélation en valeurs non standardisées.

Le calcul de l'ordonnée à l'origine

L'ordonnée à l'origine est utilisée, en régression, pour déterminer le point exact où la droite de régression coupe l'ordonnée. Il se définit par la Formule 7.2.

$$a = M_Y - b \times M_X \qquad \text{Formule 7.2}$$

où *a* est l'ordonnée à l'origine, M_Y est la moyenne de la variable *Y*, M_X est la moyenne de la variable *X* et b est le coefficient de régression.

À partir des données du Tableau 7.1, nous calculons la moyenne de Y (la température en Fahrenheit : $M_Y = 32$), la moyenne de X (la température moyenne en Celsius : $M_X = 0$) et le coefficient b que nous avons déjà calculé, b = 1,8. Nous utilisons maintenant la Formule 7.2 afin de calculer l'ordonnée à l'origine et nous trouvons a = 32.

$$\begin{aligned} a &= M_Y - b \times M_X \\ &= 32 - 1{,}8 \times 0 \\ &= 32 \end{aligned}$$

Ainsi nous voyons que la droite de régression « coupe » l'ordonnée à la valeur +32.

L'ordonnée à l'origine peut prendre des valeurs positives, négatives ou nulles.

- Une valeur positive implique que, lorsque X est à zéro, la valeur de Y est plus grande que zéro.
- Une valeur négative implique que, lorsque X est à zéro, la valeur de Y est plus petite que zéro (c'est-à-dire qu'elle prend un signe négatif).
- Une valeur nulle implique que, lorsque X est à zéro, la valeur de Y est elle aussi égale à zéro.

L'explication de l'ordonnée à l'origine et sa relation avec b.

L'ordonnée à l'origine est la valeur de *Y* lorsque la droite de régression coupe l'ordonnée. Si on étudie la Figure 7.3, on voit que la droite de régression indique la valeur de 32 sur l'ordonnée (la température en degrés Fahrenheit) lorsque la valeur de l'abscisse (la température en degrés Celsius) est égale à zéro. L'ordonnée à l'origine est donc 32.

Lorsque le coefficient de régression est égal à zéro (ce qui implique qu'il n'y a pas de corrélation entre *X* et *Y*), l'ordonnée à l'origine est égale à la moyenne de *Y* : la connaissance de *X* ne réduit pas l'incertitude de *Y* (voir le chapitre 6). On peut noter que lorsque la corrélation XY est zéro (ce qui produira b = 0), la meilleure estimation qu'on ait de *Y* est la moyenne de sa distribution (voir le chapitre 3 qui explique pourquoi, en l'absence de toute autre information, la moyenne est la meilleure estimation de n'importe

quelle valeur d'une distribution). La droite de régression, lorsque b est égal à zéro, est une ligne horizontale qui coupe l'ordonnée (la variable Y) à sa moyenne. Ainsi, pour chaque valeur de *X*, la valeur prédite de *Y* sera toujours la même valeur de *Y*, en l'occurrence la moyenne de *Y*.

L'équation de régression linéaire

Même si on travaille avec des nuages de points et la droite de régression, et qu'il est possible de prédire les valeurs *Y* pour chaque valeur de *X*, cela n'est pas particulièrement pratique. En répondant au Quiz rapide 7.2, on remarquera sans doute la difficulté de trouver la réponse exacte. Puisque maintenant cette droite de régression est définie mathématiquement, pourquoi ne pas se servir directement d'une équation, sans passer par la tâche fastidieuse de construire un diagramme de dispersion ? Pour cela, il faudra faire appel à l'équation utilisée pour construire une ligne droite,

$$Y = a + (b \times X) \qquad \text{Formule 7.3}$$

où *Y* est la valeur de la variable dépendante que nous voulons estimer à partir de X, qui est la valeur (connue) de la variable indépendante, et a et b sont respectivement l'ordonnée à l'origine et le coefficient de régression.

Nous pouvons, grâce à la Formule 7.3, prédire n'importe quelle valeur dépendante Y en fonction de n'importe quelle valeur indépendante X. Le Tableau 7.2 indique les valeurs en Fahrenheit à partir des valeurs en Celsius du Tableau 7.1. La procédure de calcul est simple. Nous savons déjà — parce que nous les avons calculés — que les coefficients a et b sont respectivement +32 et +1,8.

Par exemple, quelle est la température en Fahrenheit (Y) lorsqu'il fait 100 °C (X) ? $Y = 32 + 1{,}8 \times 100 = 212$. Lorsqu'il fait 100 °C, il fait 212 °F.

Quiz rapide 7.4
Quelle est la température en Fahrenheit lorsqu'il fait 22 °C ?

Tableau 7.2
Équation de régression pour prédire la température en Fahrenheit pour une température mesurée en Celsius

Variable indépendante X	*Équation de régression* $Y = 32 + 1{,}8 \times X$	*Valeurs prédites de la variable dépendante Y*
Celsius		*Fahrenheit*
–40	$= 32 + 1{,}8 \times -40$	–40
–30	$= 32 + 1{,}8 \times -30$	–22
–20	$= 32 + 1{,}8 \times -20$	–4
–10	$= 32 + 1{,}8 \times -10$	14
0	$= 32 + 1{,}8 \times -00$	32
10	$= 32 + 1{,}8 \times 10$	50
20	$= 32 + 1{,}8 \times 20$	68
30	$= 32 + 1{,}8 \times 30$	86
40	$= 32 + 1{,}8 \times 40$	104

L'erreur de prédiction en régression linéaire

Jusqu'à présent, nous avons expliqué la régression linéaire en utilisant l'exemple de la corrélation entre les échelles de température en Fahrenheit et en Celsius, parce qu'elles sont en parfaite corrélation, ce qui facilite la compréhension. Mais les corrélations parfaites sont très rares en réalité. Il est donc temps de passer au concept de la régression lorsque la corrélation n'est pas parfaite.

Dans le cas des corrélations imparfaites, il faut introduire le concept de l'*erreur de prédiction,* que l'on appelle aussi *l'erreur d'estimation*. L'erreur d'estimation est utilisée pour calculer une statistique qui porte le nom d'*erreur type d'estimation*. Les formules vues précédemment font une prédiction de la valeur probable de Y pour une valeur de X donnée, et l'erreur type d'estimation indique le degré d'erreur possible pour cette prédiction.

L'erreur de prédiction

Lorsque nous faisons une régression, nous avons un ensemble de valeurs pour X et Y. Nous avons la corrélation entre ces deux séries de valeurs plus les informations descriptives des deux variables (leurs écarts types et leurs moyennes). Nous utilisons ces informations pour définir la droite de régression (les coefficients a et b). En appliquant la Formule 7.3, nous obtenons les valeurs prédites de Y pour chaque valeur de X. En conséquence, nous avons pour chaque X deux informations : la véritable valeur Y qui lui correspond ainsi que la valeur prédite, que nous noterons maintenant $\hat{Y}$. Nous pouvons aussi calculer la différence entre Y et $\hat{Y}$, Cette différence s'appelle *l'erreur de prédiction*.

Au Tableau 7.2, la colonne à gauche contient un ensemble de valeurs X (les températures en degrés Celsius), la colonne centrale applique l'équation de régression et la colonne à droite donne le résultat du calcul, c'est-à-dire les valeurs $\hat{Y}$: les valeurs prédites (de la température en degrés Fahrenheit) pour chacune des valeurs X. La corrélation entre X et Y est 1,00 dans ce cas. Maintenant, si on compare les valeurs de $\hat{Y}$ du Tableau 7.2 aux valeurs Y initiales inscrites au Tableau 7.1, on voit que ces valeurs sont identiques, ce qui indique que la prédiction est parfaite, ne produisant aucune erreur. Cette prédiction est parfaite parce que la corrélation entre ces deux variables est parfaite.

Lorsque la corrélation entre deux variables n'est pas égale à 1,00, les prédictions ne sont pas parfaites. Nous faisons des erreurs de prédiction. Lorsque nous prédisons que l'étudiant qui consacre trois heures d'étude pour chaque heure de cours obtient 90 % à son examen de statistiques, alors, qu'en réalité, cet étudiant obtient 100 %, nous avons fait une erreur de prédiction. Appelons e la quantité d'erreurs pour une prédiction. Il s'agit de la différence entre la valeur prédite $\hat{Y}$ pour une observation et sa véritable valeur Y.

$$e = (\hat{Y} - Y) \qquad \text{Formule 7.4}$$

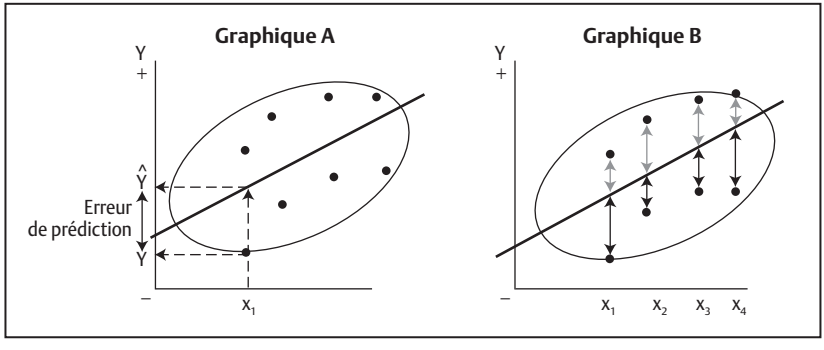
où e est l'erreur de prédiction, Y est la valeur véritable obtenue par l'observation et $\hat{Y}$ est la valeur prédite pour cette même observation sur la variable Y.

On remarque que e peut prendre des valeurs positives, négatives ou nulles. Lorsque la prédiction est parfaitement juste, e est égale à zéro. Lorsque

e est positive, nous avons fait une erreur de *surestimation* (nous avons prédit une valeur pour Y qui est supérieure à sa vraie valeur). Lorsque *e* est négative, nous avons fait une erreur de *sous-estimation* (nous avons prédit une valeur pour Y qui est inférieure à sa valeur réelle).

Le Graphique A de la Figure 7.4 présente une partie du nuage de points extraite de la Figure 7.1. Supposons que nous désirons prédire la valeur de *Y* à partir d'une observation ayant comme *X* la valeur X_1. Nous traçons une ligne verticale de X_1 jusqu'à la droite de régression, puis une ligne horizontale jusqu'à l'ordonnée, ce qui donne la valeur prédite $\hat{Y}_1$ correspondant à X_1. Or, le Y réel, dans ce cas, est plus bas. Si nous traçons une ligne horizontale de cette observation vers l'ordonnée, nous trouvons le Y_1 réel. Nous avons fait, pour cette première observation, une erreur de prédiction que nous notons e_1 et qui n'est que la différence entre la valeur réelle Y_1 et la valeur prédite $\hat{Y}_1$.

FIGURE 7.4



Quiz rapide 7.5

À la Figure 7.4, nous avons une valeur prédite ($\hat{Y}$) et une valeur observée (Y) pour l'observation qui se situe à X1 sur la variable X. La valeur prédite est-elle surestimée, sous-estimée ou exacte ?

Généralisons la situation. On voit dans le Graphique B de la Figure 7.4 des erreurs de prédiction qui sont plus ou moins grandes pour toutes les observations. Pour faciliter la lecture de ce graphique, les erreurs de sous-estimation sont décrites avec des flèches pâles et les erreurs de surestimation, avec des flèches foncées.

Puisque nous voulons les prédictions les plus exactes possible, nous voulons une droite de régression qui minimise l'erreur moyenne de prédiction. Nous avons vu au chapitre 3 que la moyenne est la valeur qui décrit chaque observation de la distribution avec une erreur moyenne minimale (c'est-à-dire zéro, car la somme des écarts entre la moyenne et les observations est toujours égale à zéro). Dans le cas de la régression linéaire, la droite de régression est exactement à la bonne place lorsqu'elle fait autant de surestimations que de sous-estimations des valeurs prédites. En d'autres termes, la droite de régression est à la bonne place lorsque la somme des surestimations est égale à la somme des sous-estimations. Nous savons (voir le chapitre 3) que la moyenne est calculée correctement lorsque la somme des écarts positifs et égale à la somme des écarts négatifs. La pente de la régression est par conséquent à la bonne place lorsqu'elle est à la moyenne des erreurs d'estimation. Cependant, comme on l'a vu avec la moyenne, ce n'est pas parce que la droite de régression fait les meilleures prédictions possibles que celles-ci sont excellentes. Il faut trouver une façon d'estimer la taille de ces erreurs de prédiction.

L'erreur type d'estimation

Nous pouvons calculer l'erreur faite pour chaque prédiction afin d'en calculer la moyenne en utilisant la Formule 7.5 que l'on connaît déjà (il s'agit de la formule habituelle pour le calcul d'une moyenne).

$$\sum_{i=1}^{N} e_i / N \qquad \text{Formule 7.5}$$

où $\sum_{i=1}^{N} e_i$ est la somme des erreurs de prédiction pour chaque observation Y_i.

Or, cette procédure pose un problème. On se souvient que la droite de régression a été créée de manière à ce que la somme des sous-estimations égale la somme des surestimations : elle se situe à la moyenne des erreurs. Donc, lorsqu'on fait la somme des erreurs de prédiction, on a autant de valeurs positives (surestimation) que de valeurs négatives (sous-estimation),

et c'est ce qui va créer un problème. La quantité $\sum_{i=1}^{N} e_i / N$ sera invariablement égale à zéro !

Pour solutionner ce problème, on utilise la stratégie à laquelle on a déjà eu recours pour calculer la variance. Ainsi, chacune des différences entre la véritable valeur de Y et sa valeur prédite sera mise au carré, puis l'on prendra la moyenne de ces erreurs de prédiction au carré. L'erreur de prédiction moyenne au carré sera toujours plus grande que zéro (sauf lorsque la prédiction est parfaite). Cette procédure donnera la variance des erreurs. Enfin, en calculant la racine carrée de la variance des erreurs, on obtient l'écart-type des erreurs.

Pour éviter la confusion, on donne un nom particulier à l'écart-type des erreurs de prédiction, l'*erreur type d'estimation*, dont le symbole statistique est s_e.

Nous savons déjà comment calculer l'écart-type des observations autour de la moyenne ($s = \sqrt{[\sum(Xi\text{-}M)^2/N-1]}$). L'erreur type d'estimation (l'écart-type des erreurs) se calcule de la même façon : on calcule la différence entre chaque erreur et la moyenne des erreurs que nous mettons au carré ; nous faisons la somme de ces quantités ; puis nous divisons cette quantité par les degrés de liberté (N-1), et enfin, nous tirons la racine carrée du résultat. La Formule 7.6a explicite le concept.

$$s_e = \sqrt{\frac{\sum_{i=1}^{N} (e_i - M_e)^2}{N-1}} \qquad \text{Formule 7.6a}$$

La clé, ici, consiste à bien comprendre que l'erreur ($e = \hat{Y} - Y$) est une véritable variable et, comme telle, il est facile de calculer son écart-type. Mais on se souvient que l'erreur moyenne (M_e) est toujours égale à zéro, ce qui permet de simplifier la formule et de produire la Formule 7.6b qui donne le calcul de l'erreur type d'estimation.

$$s_e = \sqrt{\frac{\sum_{i=1}^{N} (e_i - 0)^2}{N-1}} = \sqrt{\frac{\sum_{i=1}^{N} e^2}{N-1}} \qquad \text{Formule 7.6b}$$

où s_e est l'erreur type d'estimation (faite par la régression), e_i est l'erreur ($\hat{Y}_i - Y_i$) de prédiction pour chaque observation i et N est le nombre d'observations.

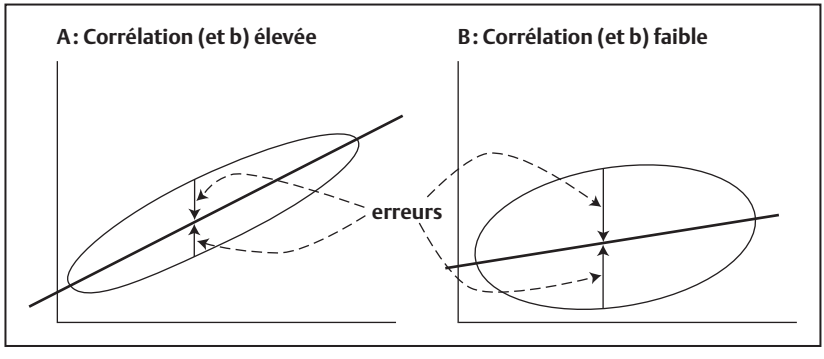
Les étapes à suivre pour calculer l'erreur type d'estimation sont les suivantes :

1. On calcule, pour chaque prédiction, l'erreur qu'elle fait ($e = \hat{Y}_i - Y_i$).
2. On ne peut pas prendre la somme de ces différences, car elle sera égale à zéro. Alors on utilise la technique habituelle pour se débarrasser du problème, on met chaque erreur e au carré [$e^2 = (\hat{Y}_i - Y_i)^2$].
3. On fait la somme de ces différences au carré ($e^2 = [\sum(\hat{Y}_i - Y_i]^2$).
4. On calcule la moyenne de cette somme en divisant, comme pour l'écart-type, par N – 1 : ($s_e^2 = [\sum(\hat{Y}_i - Y_i]^2$ / N–1). On a maintenant calculé la variance des erreurs d'estimation.
5. Enfin, on prend la racine carrée de cette sommation et on obtient l'erreur type d'estimation : ($s_e = \sqrt{s_e^2}$).

Pour chaque prédiction, nous pouvons maintenant établir la valeur prédite et, par le bais de l'erreur type, nous pouvons également connaître l'erreur typique de cette prédiction. On va bientôt voir l'utilité de l'erreur type d'estimation lorsqu'il s'agit de tirer des conclusions pratiques. Avant d'y venir, il convient d'admettre que cette technique pour le calcul de l'erreur type d'estimation peut être plutôt fastidieuse. Il vaudrait mieux trouver un procédé plus simple.

Il existe une relation importante entre l'erreur d'estimation et la corrélation. Lorsque la corrélation est élevée, les erreurs d'estimation sont plus petites que lorsque la corrélation est faible. À la Figure 7.5, le diagramme de dispersion A est tracé à partir d'une corrélation élevée (ce qui produira un coefficient de régression élevé). Par contre, le diagramme de dispersion B représente une corrélation (et un coefficient de régression) plus faible. La taille des erreurs de prédiction (donc la différence entre la valeur véritable Y_i et la valeur prédite $\hat{Y}_i$) est représentée, dans chaque graphique, par la longueur des flèches. Il est donc clair qu'une relation plus faible entre la variable indépendante X et la variable dépendante Y engendre des erreurs de prédiction plus grandes que lorsque la relation XY est plus élevée.

FIGURE 7.5 Relation entre la corrélation et l'erreur d'estimation



La Formule 7.7 présente une approche plus simple pour calculer l'erreur type d'estimation. Elle mise sur le fait que la taille de la corrélation et la taille des erreurs de prédiction sont en étroite relation. La Formule 7.7 est une approximation, mais son résultat est très proche de celui que nous pourrions obtenir en utilisant le procédé plus complexe décrit ci-dessus :

$$s_e = s_Y \sqrt{1 - r^2_{xy}} \qquad \text{Formule 7.7}$$

où s_e est l'erreur type d'estimation, s_Y est l'écart-type de la variable dépendante et r^2_{xy} est la corrélation au carré (le coefficient de détermination) pour la corrélation entre *X* et *Y*.

Il est facile de comprendre logiquement pourquoi la Formule 7.7 produit une estimation fort précise de l'erreur type d'estimation : lorsque la corrélation est parfaite (le coefficient de détermination est égal à 1), la prédiction de *Y* à partir de *X* ne peut faire aucune erreur. Lorsque aucune erreur n'est possible, l'erreur type d'estimation doit nécessairement être zéro. Calculons l'erreur type d'estimation lorsque $r_{xy} = 1{,}0$ et $s_y = 10$. Nous connaissons déjà la réponse. Puisque la corrélation est parfaite, la prédiction sera parfaite, et l'erreur typique — l'erreur type d'estimation — devra obligatoirement être égale à zéro, si notre Formule 7.7 est la bonne.

$$\begin{aligned} s_e &= s_Y \sqrt{1 - r^2_{xy}} \\ &= 10 \sqrt{1 - 1^2} \\ &= 10\ (0) = 0 \end{aligned}$$

Lorsque la corrélation entre X et Y n'est pas parfaite, la possibilité existe de faire des erreurs de prédiction. La Figure 7.6 présente ce type de situations où nous voyons plusieurs véritables valeurs Y associées à la même valeur X_4 de la variable indépendante. La ligne pointillée de la Figure 7.6 indique la valeur $\hat{Y}_4$, qui est celle prédite pour toutes les observations qui se situent à la valeur X_4, ce qui implique que nous allons faire des erreurs de prédictions. Présumons que ces erreurs sont distribuées normalement[2]. L'erreur type d'estimation est l'écart-type de cette distribution. Nous savons, en nous référant aux caractéristiques de la distribution normale (voir le chapitre 5) qu'environ 68 % des observations d'une distribution normale se situent entre −1 et +1 écart-type de la moyenne. Par conséquent, 68 % des erreurs de prédictions associées à X_4 se situent entre ±1 erreur type d'estimation de la moyenne de cette distribution, qui, elle, est $\hat{Y}_4$. Nous pouvons alors conclure que la meilleure estimation que nous ayons de la valeur X_4 est $\hat{Y}_4$, mais qu'il y a 68 % des chances que la véritable valeur qui correspond à X_4 se situe à plus ou moins une erreur type d'estimation de la valeur prédite $\hat{Y}_4$.

Par exemple, si la valeur prédite est $\hat{Y}_i = 10$ et que l'erreur type d'estimation est égale à 1, il y a 68 % de chances que la vraie valeur Y_i se situe en réalité entre 9 et 11. Ainsi, l'erreur type d'estimation fournit une fourchette de valeurs où la valeur réelle de Y_i correspondant à la valeur X_i a 68 % de chances de se trouver.

L'utilité de l'erreur type d'estimation

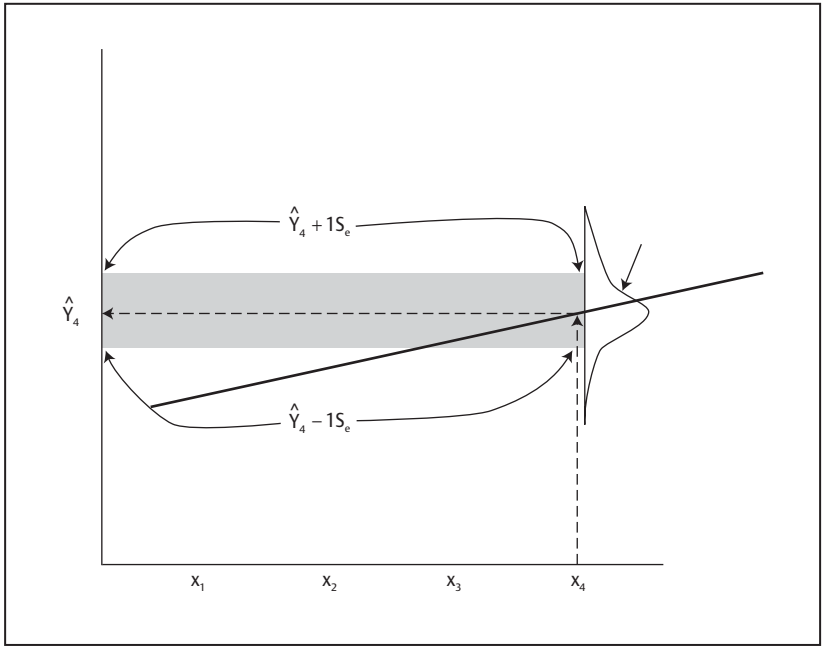
Présumons que les valeurs *X* et Y mises en relation sont distribuées normalement. En conséquence, les prédictions faites par la régression sont, elles aussi, distribuées normalement. Et puisque les prédictions sont normalement distribuées, les erreurs de prédiction le seront aussi.

Pour établir cette fourchette de valeurs, il faut suivre les étapes suivantes.

1. On calcule une valeur prédite $\hat{Y}$ pour chaque valeur X qui nous intéresse (Formule 7.3).
2. On calcule l'erreur type d'estimation de la régression (Formule 7.6 ou 7.7).
3. On ajoute l'erreur type d'estimation à la valeur prédite. Cela indique la valeur maximale de la fourchette de valeurs.

2. Les « analyses des résiduelles » sont des procédures statistiques qui servent à vérifier cette présomption. Ces procédures sont décrites dans des manuels plus avancés.

FIGURE 7.6 Représentation et utilisation de l'erreur type d'estimation pour les relations XY qui ne sont pas parfaites



4. On soustrait l'erreur type d'estimation de la valeur prédite. Cela indique la valeur minimale de la fourchette de valeurs.
5. Il y 68 % de chances que la valeur de Y pour chaque observation X se trouve entre les deux extrémités de cette fourchette de valeurs.

EXEMPLE DE PRÉDICTION DE LA NOTE À UN EXAMEN FINAL

Dans un cours, les étudiants ont deux examens : un examen partiel et un examen final. L'examen partiel a lieu et il est corrigé avant la date limite d'abandon des cours. La note de passage est de 50 %. En général les étudiants qui pensent ne pas réussir le cours préfèrent l'abandonner plutôt que d'avoir un échec inscrit à leur bulletin.

Il arrive chaque année qu'à la suite des résultats à l'examen partiel, au moins un étudiant consulte son professeur pour savoir s'il devrait abandonner le cours. Ce dernier, à partir de la note de l'étudiant à l'examen

partiel, estime sa note probable à l'examen final. Techniquement, le professeur prédit la note finale de l'étudiant (qui est inconnue) à partir de sa note partielle (qui, elle, est connue). On se trouve face à un problème pour lequel la régression linéaire fournit la solution.

La corrélation entre les notes à l'examen partiel et à l'examen final telle qu'elle est établie à partir des résultats obtenus au cours des cinq dernières années est de 0,825, soit une très forte corrélation. On peut présumer que cette relation se maintiendra[3]. On exécute donc une régression linéaire à partir de ces données. Les résultats de cette analyse sont présentés aux Tableaux 7.3 à 7.5. Les calculs sont exécutés par le logiciel SPSS.

D'abord, on calcule les statistiques descriptives des variables, les moyennes et les écarts types (Tableau 7.3), et la corrélation entre les résultats aux examens partiel et final observée dans les années antérieures (Tableau 7.4). À partir de cette corrélation et de l'écart-type de la variable (Y), on estime l'erreur type d'estimation (Tableau 7.4).

Tableau 7.3
Statistiques descriptives pour les deux examens

	Moyenne	***Écart-type***	***N***
Examen final (*Y*)	66,9	13,6	131
Examen partiel (*X*)	57,6	9,74	131

Tableau 7.4
Corrélation, coefficient de détermination et erreur type d'estimation pour la relation entre l'examen partiel et l'examen final

r	***r²***	***Erreur type d'estimation évaluée avec la Formule 7.7***
0,825	0,681	7,7

3. Si nous ne pouvons pas faire cette présomption, la régression linéaire n'est d'aucune utilité.

Au Tableau 7.5, on utilise les informations provenant des Tableaux 7.3 et 7.4 pour calculer le coefficient de régression aussi bien que l'ordonnée à l'origine.

Tableau 7.5
Coefficients pour la régression du Tableau 7.3

Ordonnée à l'origine *a*	0,55
Coefficient de régression *b*	1,15

On a déjà calculé l'ordonnée à l'origine (a = 0,548) ainsi que le coefficient de régression (b = 1,15).

Trois étudiants désirent faire une estimation de leur note à l'examen final en fonction de leur note à l'examen partiel. Les étudiants A, B et C ont respectivement obtenus 40, 60 et 80 à l'examen partiel. Armé de ces statistiques, on fera appel à l'équation de la régression linéaire ($Y = a + b \times X$, Formule 7.3) afin d'estimer la note de chacun à l'examen final. Par exemple, pour l'étudiant A qui a obtenu 40 à son examen partiel, nous prédisons une note de 46,55 à l'examen final : $Yi = 0,55 + (1,15 \times 40) = 46,55 = 46,6$.

Le Tableau 7.6 montre la prédiction de la note finale pour ces trois étudiants.

Tableau 7.6
Prédiction de la note à l'examen final ($\hat{Y}$) pour trois étudiants à partir de leur note à l'examen partiel (*X*)

						Fourchette de valeurs (erreur type d'estimation)	
Étudiant	***Note à l'examen partiel X***	***a***	***b***	s_e	***Note prédite*** $\hat{Y} = a + b \times X$	***Note minimale prédite (68 %)***	***Note maximale prédite (68 %)***
A	40	0,55	1,15	7,7	46,6	38,9	54,4
B	60	0,55	1,15	7,7	69,7	62	77,4
C	80	0,55	1,15	7,7	92,7	85	100,4

Mais ce n'est pas tout. On veut aussi connaître la fourchette de valeurs à l'intérieur de laquelle la note prédite pour chaque étudiant pourrait se trouver. Pour réaliser ce projet, on calcule l'erreur type d'estimation ; dans ce cas, avec la Formule approximative 7.7 : $s_e = s_Y \sqrt{1 - r^2_{xy}} = 13{,}6\ (1 - 0{,}681^2) = 7{,}69 = 7{,}7$. En additionnant l'erreur type d'estimation à la note prédite pour l'étudiant A, on obtient 46,66 + 7,7 = 54,36. En soustrayant l'erreur type d'estimation de la valeur prédite, on obtient 38,96. Ainsi, on peut prédire que la note à l'examen final pour l'étudiant A sera 46,6, mais qu'il y a de fortes possibilités pour qu'elle se trouve entre 39 et 54.

À partir de ces résultats, voici quelques conseils à donner aux trois étudiants.

Pour l'étudiant A : on peut lui prédire une performance de 46,6 à l'examen final, ce qui lui vaudra un échec. Mais en considérant l'erreur type d'estimation, on dira (en arrondissant) qu'il y a deux chances sur trois pour que sa note soit aussi faible que 39 mais pas plus forte que 54. S'il choisit de rester dans le cours, cet étudiant risque l'échec s'il ne change rien à ses habitudes d'étude ou s'il ne rattrape pas la matière qu'il n'a pas encore comprise. Pour les étudiants B et C, la régression linéaire prédit qu'ils vont, tous deux, obtenir plus de 50 % à l'examen final (69,66 et 92,66 respectivement). En prenant en considération l'erreur type d'estimation, nous suggérons aux étudiants B et C de rester dans le cours : ils ont tous deux de très bonnes chances de le réussir.

Quiz rapide 7.6
Supposons que les étudiants A, B et C choisissent d'abandonner le cours si leur note finale risque d'être inférieure à 75 %. Quelle serait votre recommandation pour ces étudiants ?

La différence entre le coefficient b et le coefficient β

Le coefficient β a la même signification et il est calculé de la même façon que le coefficient b. La différence entre eux est que le coefficient β se calcule lorsque les variables X et Y sont standardisées en valeurs étalons Z. Reprenons la Formule 7.1 qui est utilisée pour calculer le coefficient b. Il s'agit de la corrélation multipliée par le rapport entre les écarts types de la

variable dépendante et de la variable indépendante. On se souviendra que la moyenne et l'écart-type de toutes les distributions de valeurs étalons Z sont 0 et 1 respectivement. Par conséquent, à la Formule 7.1, $s_X = 1 = s_Y$, et le rapport s_Y/s_X sera lui aussi égal à 1,0. Le calcul du coefficient b exige la multiplication de la corrélation par le rapport entre les deux écarts types (X et Y). Puisque le rapport s_Y/s_X est invariablement égal à 1,0 lorsque les distributions X et Y sont standardisées, le coefficient de régression β se réduit obligatoirement au coefficient de corrélation, ce que la Formule 7.8 nous indique. Puisque les variables X et Y sont standardisées, le coefficient de régression β le sera aussi et c'est pour cette raison qu'on lui donne le nom de *coefficient de régression standardisé*. Ainsi, dans le cas de la régression linéaire simple, le coefficient β est invariablement identique à la corrélation.

$$b = r_{xy} \times \frac{s_Y}{s_X} = \beta = r_{xy} \text{ lorsque } s_y = s_x \qquad \text{Formule 7.8}$$

L'ordonnée à l'origine pour la régression standardisée

L'ordonnée à l'origine, lorsque nous travaillons avec la régression linéaire standardisée, est invariablement zéro. Dans ce cas, nous utilisons le symbole β_0 pour le différencier de a, le symbole que nous utilisons pour la régression non standardisée. Étudiez la formule pour le calcul de l'ordonnée à l'origine

$$a = M_Y - b \times M_X$$

Puisque nous travaillons avec des données standardisées, nous savons que la moyenne des deux variables sera zéro. Mettons ces chiffres dans la formule.

$$\beta_0 = 0 - \beta \times 0$$

L'ordonnée à l'origine standardisée est invariablement zéro.

La régression simple et la régression multiple

Ce chapitre traite exclusivement de la régression linéaire simple. Elle prend le nom de régression simple parce qu'elle permet la prédiction d'une valeur

dépendante à partir *d'une seule* variable indépendante. Il est possible de faire la prédiction d'une variable dépendante à partir de plusieurs variables indépendantes. Par exemple, nous pourrions prédire la note finale à un examen en prenant en considération simultanément le degré d'intérêt de l'étudiant pour la matière, le nombre d'heures d'étude qu'il y consacre et l'approche pédagogique du professeur. Cette forme de régression est connue sous le nom de *régression linéaire multiple*. Celle-ci fait appel aux mêmes concepts que ceux qui ont été abordés dans ce chapitre, mais les formulations sont plus complexes, impliquant des calculs additionnels. Les textes plus avancés traitent de la régression multiple.

SOMMAIRE DU CHAPITRE

La régression linéaire est la statistique que l'on utilise le plus souvent dans les domaines appliqués. C'est une procédure statistique qui permet de prédire une valeur précise (sur une variable *Y*) à condition que l'on connaisse sa valeur sur une deuxième variable (*X*) et la relation générale (la corrélation) entre les deux variables *X* et *Y*. Pour prédire cette valeur, on calcule deux statistiques : le coefficient de régression b et l'ordonné à l'origine a. Ces deux coefficients se calculent à partir de la corrélation, des moyennes de *X* et de *Y* et des écarts types de ces deux variables. Le résultat est une prédiction $\hat{Y}$, la valeur probable que la personne obtiendra sur la variable dépendante Y, à partir de sa performance sur la variable indépendante X. Une fois cette valeur prédite, on peut calculer l'erreur type d'estimation. La combinaison de la valeur prédite et de l'erreur type d'estimation permet d'estimer une valeur inconnue avec un degré de certitude déterminé.

EXERCICES DE COMPRÉHENSION

1. La corrélation entre deux variables (X et Y) est égale à +1,0 alors que la corrélation entre deux autres variables (A et B) est égale à 0,0. Si nous utilisons X pour prédire Y et si nous utilisons A pour prédire B, le coefficient de régression standardisé (β) pour X sera _____ et le coefficient de régression standardisé pour A sera _____.

a) entre –1 et +1 ; entre –1 et +1
b) entre –1 et +1 ; 1,0
c) + 1 ; 0
d) + ou –1 ; + ou –1,0

2. Nous faisons une régression pour prédire Y à partir de X. Le coefficient de régression standardisé est égal à +1,0. La position de Julie sur la variable X est égale à Z = + 1,2. Quelle sera la position de Julie sur la variable Y ?
a) +Z = 1,0
b) Z = +1,2
c) Z entre 0 et +1,0
d) Impossible à déterminer puisque nous ne connaissons pas l'ordonnée à l'origine.

3. L'équation de régression pour prédire Y à partir de X nous donne le résultat suivant : Y = 12 + 2,4X. La valeur de l'ordonnée à l'origine est _____ ; le coefficient b est _____.
a) 12 ; 2,4
b) 2,4 ; 12
c) 12 ou 2,4 ; 12 ou 2,4
d) impossible à déterminer car nous ne connaissons pas l'écart type

4. On nous apprend que le lien entre les variables *Y* et *X* est : $Y = 3 + 49X$. Une personne obtient 1 à la variable X. Nous prédisons alors que la valeur Y pour cette personne sera _____.

5. En analysant le graphique de dispersion pour une relation entre X et Y, nous voyons que toutes les coordonnées du graphique se trouvent exactement sur la droite de régression. Il est alors certain que le coefficient β standardisé _____.
a) peut prendre n'importe quelle valeur entre –1 et +1
b) est obligatoirement +1
c) est obligatoirement –1
d) est obligatoirement +1 ou –1

6. La corrélation entre X et Y est égale à zéro, et nous construisons une équation de régression pour cette relation. Quelle sera la valeur standardisée prédite Y_p, pour chacune de ces trois valeurs standardisées de X : $Z_{X1} = -1$; $Z_{X2} = 0$; $Z_{X3} = +2$?

a) $Z_{Y1} = -1$; $Z_{Y2} = 0$; $Z_{Y3} = +2$
b) $Z_{Y1} = 0$; $Z_{Y2} = 0$; $Z_{Y3} = 0$
c) $Z_{Y1} = -1$ à $+1$; $Z_{Y2} = -1$ à $+1$; $Z_{Y3} = -1$ à $+1$
d) La corrélation XY étant zéro, il est impossible de construire une régression.

7. Dans cette distribution, l'écart qui existe entre chaque coordonnée XY et la droite de régression est invariablement égal à zéro.
a) Par conséquent, chaque valeur de Y sera parfaitement prédite par la valeur de X qui lui est associée.
b) Par conséquent, chaque valeur de Y ne pourra pas être prédite par la valeur de X qui lui est associée.
c) Par conséquent, l'erreur type d'estimation sera moins grande que 1,0.
d) Toutes ces réponses sont fausses.

8. Nous calculons l'écart qui existe entre chaque coordonnée et la droite de régression. Nous faisons la somme de ces écarts, prenant bien en considération le signe (positif ou négatif) de chaque différence. La somme de ces écarts _____.
a) sera égale à la moyenne des écarts types de X et de Y
b) sera égale au coefficient de régression b
c) sera égale à l'erreur d'échantillonnage
d) sera égale à zéro

9. À la suite de cette régression, nous voyons que le coefficient standardisé β est égal à 1,0. Quel sera le coefficient non standardisé b ?
a) 1,0
b) +1 ou –1
c) 0
d) Impossible à déterminer avec les informations fournies.

Réponses

1. c
2. b
3. a
4. 52
5. d
6. b
7. a
8. d
9. d

CHAPITRE 8
LES CONCEPTS DE L'INFÉRENCE STATISTIQUE

Page laissée blanche

CHAPITRE 8

LES CONCEPTS DE L'INFÉRENCE STATISTIQUE

Ce chapitre présente les notions fondamentales de l'inférence statistique, c'est-à-dire l'ensemble des règles qui permettent l'interprétation que l'on peut faire et les conclusions qui peuvent être tirées des résultats d'une enquête ou d'une recherche, résultats qui sont analysés statistiquement. Ces règles sont organisées en *tests statistiques* qui vont servir à donner des réponses affirmatives ou négatives à des questions formulées sous forme d'hypothèses. Ces tests statistiques peuvent aider à comprendre un phénomène ou à prendre une décision. Par exemple, pour savoir si une nouvelle thérapie est meilleure que l'ancienne (oui/non), ou si le fonctionnement cérébral des criminels diffère de celui des non-criminels (oui/non).

Les concepts que nous abordons maintenant sont comme des legos. Une fois emboîtés, ils donnent les tests statistiques que nous verrons dans les autres chapitres. Nous allons ici étudier les principes qui sous-tendent le concept d'une inférence statistique. Tout d'abord, voyons les différences entre le concept de l'échantillon et le concept de la population.

L'ÉCHANTILLON ET LA POPULATION : LES DEUX CONCEPTS FONDAMENTAUX DE L'INFÉRENCE

Rien n'est plus important pour l'inférence statistique que les concepts d'*échantillon* et de *population*. Le texte ci-dessous présente la conclusion

tirée de l'étude d'un échantillon d'élèves et qui portait sur la population d'élèves québécois.

L'Enquête internationale sur les mathématiques et les sciences

On mène périodiquement une vaste étude internationale afin d'évaluer le degré de compétence des écoliers en mathématiques et en sciences. Des milliers de jeunes enfants dans plus de 40 pays vont ainsi passer le même examen. Et on a constaté ceci : « [E]n maths, les élèves québécois de quatrième année du primaire ont obtenu 550 points en 1995. Huit ans plus tard, leur score était de 506 points. La chute : 44 points. » À partir de ce constat, les chercheurs peuvent affirmer ceci : « On peut conclure sans grand risque de se tromper que les résultats de 2003 sont à la baisse » ; c'est-à-dire que les élèves québécois sont moins forts dans ces matières qu'auparavant.

Mais les écoliers de quatrième année du Québec (la population) n'ont pas tous fait partie de l'étude. Seule une partie des enfants, un *échantillon,* y a participé. Ainsi, on a tiré une conclusion au sujet de la *population* (« tous » les élèves de quatrième année du Québec) à partir des informations provenant d'un échantillon (seulement un groupe d'élèves). Cela est l'essence même de l'inférence statistique.

Source : Adaptation autorisée d'un article de Marie Allard, *La Presse*, 9 décembre 2005.

La population

En statistique, le terme « population » est utilisé dans un sens général, et pas seulement démographique. L*e concept de population fait référence à toutes les observations possibles au sujet d'un phénomène.* Voici quelques illustrations de ce concept :

- Le tour de taille de tous les Américains vivant à New York.
- Le revenu de tous les Italiens.
- Le salaire de tous les joueurs de la Ligue nationale de hockey.
- Les notes obtenues dans un cours de statistique par tous les étudiants qui y sont inscrits.
- L'attitude de tous les Canadiens envers la légalisation du cannabis.

Lorsqu'on s'intéresse à une population, nous parlons de *toutes* les observations qui existent au sujet d'un phénomène (l'obésité des New-Yorkais, le revenu des Italiens, etc.). Par conséquent, cela implique souvent un nombre immense d'observations. Si on s'intéresse aux attitudes politiques des pauvres en Europe, nous parlons des attitudes de millions d'individus.

Les populations sont souvent de taille infiniment grande (par exemple la population des électrons ou la population des étoiles dans le ciel), mais cela n'est pas nécessairement le cas. Si nous voulons connaître l'attitude envers les statistiques des 150 étudiants qui suivent un cours en particulier, la population est alors de 150 personnes. Si nous nous intéressons au salaire des joueurs de la LNH en 2002-2003, la population est composée de 679 athlètes. Mais si nous nous intéressons au salaire des joueurs de hockey professionnels *en général*, ce nombre sera beaucoup plus grand, car il existe plusieurs ligues professionnelles de hockey en Amérique du Nord, en Europe et en Asie. Le mot clé est *tout*. Lorsque nous avons accès à l'ensemble des informations, nous travaillons directement avec une population. La population réfère donc à toutes les observations possibles qui existent au sujet d'un phénomène.

Dans la grande majorité des cas, nous voulons tirer des conclusions au sujet de la population. Par exemple, la compétence en sciences des enfants québécois s'est-elle détériorée entre 1995 et 2003 ? Le problème est que nous n'avons généralement pas accès à toute la population et cela à cause de contraintes pratiques : il est généralement trop coûteux et trop compliqué de mesurer une population entière[1]. Par conséquent, nos observations proviennent d'un échantillon, un sous-ensemble de tous les membres de la population, mais les conclusions que l'on en tire s'appliquent, elles, à l'ensemble de la population et, dans l'exemple proposé, à tous les écoliers québécois de quatrième année.

L'échantillon

L'échantillon fait référence à un sous-ensemble des observations extrait d'une population. Un échantillon contient moins d'observations qu'une

1. Les recensements que les gouvernements font périodiquement — en 2006 pour le Canada — constituent une exception. Lors d'un recensement, toute la population d'un pays fournit des informations. Bien que dans plusieurs pays, y compris le Canada, il soit illégal de ne pas répondre aux questions du recensement, il reste néanmoins que ces études n'incluent pas véritablement 100 % des membres de la population : certaines personnes sont absentes du pays au moment de l'étude, d'autres sont malades, d'autres encore n'ont pas de domicile fixe, etc. Mais puisque la grande majorité des habitants y répond, les recensements sont généralement considérés comme incluant toute la population.

population (sinon l'échantillon serait la population), et souvent même considérablement moins. Le Tableau 8.1 illustre la distinction entre population et échantillon pour une diversité de phénomènes.

Tableau 8.1
Exemples des concepts de populations et d'échantillons

Population	*Échantillon*
L'attitude de tous les Canadiens envers la légalisation du cannabis.	L'attitude de 1 000 Canadiens envers la légalisation du cannabis.
Le tour de taille de tous les Américains à New York.	Le tour de taille de 500 New-Yorkais.
Le revenu de tous les Italiens.	Le revenu de 200 Italiens.
Le salaire payé à tous les joueurs de la LNH.	Le salaire des joueurs de la LNH lors de la saison 2002-2003.
Les notes obtenues par tous les étudiants inscrits à ce cours de statistique.	Les notes obtenues par 15 étudiants inscrits à ce cours de statistique.
Tous les écoliers suédois ayant des difficultés en lecture.	Un groupe d'écoliers ayant des difficultés en lecture dans une école suédoise.

LA POPULATION, L'ÉCHANTILLON ET L'INFÉRENCE

Dans presque toutes les situations, nous cherchons à apprendre quelque chose ou à tirer une conclusion au sujet d'une population. Ainsi, les politiciens désirent connaître leurs chances d'être élus ou la popularité de leurs programmes électoraux. Les sociologues désirent examiner l'attitude des immigrants envers l'intégration sociale. Les grands magasins désirent savoir si les produits placés à proximité des caisses enregistreuses se vendent mieux. Les compagnies pharmaceutiques désirent savoir si leurs nouveaux médicaments sont efficaces. Dans tous les cas, les conclusions importantes sont celles qui se rapportent à la population en général (les électeurs, les immigrants, les ventes, l'efficacité des médicaments).

En mesurant la population (tous les électeurs ou tous les immigrants, par exemple), nous pourrions alors tirer nos conclusions. Lorsque les populations sont relativement petites (tous les étudiants d'une classe ou le salaire

de chaque joueur de hockey d'une ligue en particulier pour une année précise, par exemple), cette solution est tout à fait envisageable. Cependant, on comprendra facilement que cette solution n'est pas praticable dans la majorité des situations. Il serait beaucoup trop coûteux et peu pratique de sonder tous les électeurs ou d'interviewer tous les immigrants.

Dans de telles situations, nous faisons appel à un échantillon. Ainsi, un groupe relativement restreint d'observations est extrait de cette population et les mesures sont prises exclusivement sur celui-ci. Les résultats que nous obtenons à partir de cet échantillon sont ensuite appliqués à la population, c'est-à-dire que les résultats obtenus dans l'échantillon sont utilisés pour réaliser une *inférence* au sujet de la population. Ainsi, nous nous servons d'une information connue (les informations produites par l'échantillon) afin de tirer une conclusion sur quelque chose d'inconnu (les informations qui décrivent la population). Le processus d'inférence sert donc à tirer une conclusion générale (la population) qui, elle, n'est pas mesurée, à partir d'une information précise (l'échantillon) que nous avons effectivement mesurée.

La distinction entre la population et l'échantillon ne dépend pas du nombre d'observations : si on étudie toutes les personnes atteintes d'une maladie rare, cette population pourrait n'être composée que de 200 personnes. En revanche, un sondage sur les intentions de vote inclut habituellement 1 000 personnes. Ce n'est pas parce que la population n'est composée que de 200 personnes qu'il s'agit d'un échantillon et ce n'est pas parce qu'on a interviewé 1 000 électeurs que ces derniers constituent pour autant une population.

Ainsi, si on veut examiner le niveau de stress des vendeurs dans une compagnie déterminée afin de tirer une conclusion au sujet de la personnalité des vendeurs en général, les 1 000 vendeurs de cette étude forment un *échantillon* de vendeurs tiré de la *population* de vendeurs. Mais si on ne veut décrire que ces 1 000 vendeurs, ces 1 000 vendeurs représentent alors une population, et non un échantillon, de vendeurs.

Lorsqu'une compagnie pharmaceutique fait une étude pour évaluer l'efficacité d'un nouveau médicament, elle administre ce médicament à un échantillon de patients et compare le taux de guérison de cet échantillon à un autre échantillon de patients qui, lui, n'a pas reçu le médicament ou

qui reçoit un placebo. Si on constate des effets bénéfiques sur le premier groupe, ce résultat devient intéressant pour la compagnie, car il lui permet potentiellement de tirer une conclusion générale : le médicament a des effets bénéfiques sur la *population* de patients.

Quiz rapide 8.1
Vous avez à votre disposition les données du recensement de Statistique Canada réalisé en 2006. Vous devez déterminer le salaire médian payé au Canada cette même année. Travaillez-vous en utilisant un grand échantillon ou une population ? Justifiez votre réponse.

L'ÉCHANTILLON REPRÉSENTATIF ET L'ÉCHANTILLON ALÉATOIRE

Pour parvenir à une conclusion au sujet de la population, il est essentiel de bien choisir l'échantillon. Si l'on désire analyser l'attitude des écoliers envers l'école, il est évident que l'échantillon doit être composé d'écoliers. Mais il faut aussi remplir une autre condition : l'échantillon doit être un miroir fidèle de la population. Lorsque la constitution de l'échantillon ressemble beaucoup à la population, on dit que cet échantillon est *représentatif* de la population. Le texte suivant décrit un processus d'échantillonnage qui, lui, n'est pas représentatif.

Une étrange histoire d'échantillonnage : l'escroquerie Bre-X

Au milieu des années 1990, une compagnie minière canadienne, Bre-X, explore une région de l'Indonésie dans l'espoir d'y trouver des dépôts d'or. La compagnie prétend qu'elle a découvert un site prometteur et elle demande aux investisseurs de l'appuyer dans ses efforts d'extraction et de commercialisation du minerai.

Bre-X présente aux investisseurs des échantillons de terre provenant de cette région. Ils ont été extraits aléatoirement du site, affirme-t-elle. Les résultats obtenus par l'analyse des échantillons sont utiles puisqu'ils permettent d'inférer la concentration d'or qui existe dans la population, en l'occurrence le site découvert par Bre-X. Les chimistes et ingénieurs miniers évaluent donc la concentration d'or à l'intérieur de ces échantillons.

L'analyse de ces échantillons révélant une forte concentration d'or, les chimistes infèrent alors que le site (la population), représenté par les échantillons, doit, lui aussi, contenir de l'or. En fait, la concentration d'or dans les échantillons est telle qu'ils concluent que le site découvert par Bre-X est une immense mine d'or, peut-être même la plus riche du monde. Du jour au lendemain, les actions en bourse de Bre-X grimpent de 2 à 238 $ l'action. La personne qui investit 10 000 $ un jour devient millionnaire le lendemain.

Hélas, des milliers de petits investisseurs ont cru en vain au miracle. En réalité, le site ne contenait pas plus d'or qu'un jardin montréalais. La compagnie avait triché en ajoutant volontairement de l'or dans les échantillons. La concentration d'or que les échantillons contenaient n'était donc pas du tout représentative de la population, c'est-à-dire de la concentration d'or existant dans le site. Par conséquent, la conclusion au sujet du site indonésien était fausse.

Cette escroquerie mène à deux constats :

1) L'analyse statistique d'un échantillon n'est utile que lorsque nous voulons tirer une conclusion au sujet d'une population.
2) L'analyse d'un échantillon nous renseigne sur la population que si l'échantillon représente adéquatement la population.

Dans l'exemple coloré suivant, nous voulons étudier l'attitude des femmes à l'égard des salons de coiffure. Nous pensons que la couleur de cheveux des femmes peut avoir un impact sur cette attitude. Si, dans cette population, 30 % des femmes sont blondes, 60 % brunes et 10 % rousses, un échantillon représentatif serait composé de proportions identiques de blondes, de brunes et de rousses (30 %, 60 %, et 10 % respectivement).

On remarque que ce contrôle dans la constitution d'un échantillon ne peut fonctionner que si l'on connaît la distribution de cette caractéristique (la couleur des cheveux) dans la population. Dans la majorité des cas, la distribution dans la population est inconnue ou n'est que très approximativement connue. Pour combler cette lacune, il est habituel de procéder autrement. On extrait de la population un *échantillon aléatoire*. Si nous choisissons un échantillon véritablement aléatoire de femmes pour notre étude, il sera composé naturellement d'environ 30 %, 60 % et 10 % de femmes respectivement blondes, brunes et rousses. Un échantillon aléatoirement choisi sera une représentation fidèle de la population de laquelle il est extrait.

L'échantillon aléatoire : les deux principes fondamentaux

Dans leur application, les techniques requises pour produire des échantillons aléatoires peuvent être fort complexes (on trouvera une explication de celles-ci dans des ouvrages spécialisés). Cependant, ces techniques reposent sur deux principes relativement simples qui produiront, s'ils sont respectés, une sélection aléatoire des échantillons.

Le critère de la chance égale

Le critère de la chance égale est respecté lorsque *chaque membre de la population de laquelle l'échantillon est tiré a une chance égale d'être choisi.* Si on exécute un sondage sur l'attitude des étudiants d'université envers le gouvernement, on met le nom de tous les étudiants dans un chapeau et on tire au hasard 1 000 personnes. L'échantillon sera alors aléatoirement choisi, car chaque étudiant universitaire a une chance égale d'être choisi. Supposons qu'on utilise une autre technique : des intervieweurs se placent à l'entrée de la cafétéria et posent leurs questions aux 1 000 premiers étudiants qui s'y présentent. Les chances d'être choisi ne sont plus égales, car les étudiants qui ne mangent pas à la cafétéria et ceux qui se présentent à la cafétéria après le départ des intervieweurs n'ont aucune chance d'être choisis. Cet échantillon ne sera pas représentatif de tous les étudiants, mais seulement de ceux qui mangent à la cafétéria et qui mangent plus tôt que les autres.

L'élection présidentielle aux États-Unis en 1948

Lors de la campagne électorale de 1948, Harris Truman briguait les suffrages pour le renouvellement de son mandat présidentiel contre son adversaire Thomas Dewey. Sondage après sondage, Truman était donné perdant, et cela, par une marge considérable. À l'époque, les journaux du matin devaient être imprimés la veille et le dépouillement des suffrages était très lent. Au lendemain de l'élection, plusieurs grands quotidiens américains, confiants dans les résultats indiqués par les sondages et incapables d'attendre le résultat officiel, annonçaient à la une de leur édition matinale l'écrasante victoire de Dewey. Mais, à leur stupéfaction et leur grande honte, Truman était le vainqueur ! L'inférence qui avait été faite à partir des sondages était tout simplement erronée.

En 1949, le Social Science Research Council réalisa une étude pour comprendre pourquoi les sondages avaient été dans l'erreur. Parmi les problèmes identifiés, les procédures de sélection des échantillons ont été mises en cause. Entre autres, les sondages avaient été souvent réalisés au téléphone. À l'époque, presque tous les Américains urbains avaient le téléphone, mais cela n'était pas vrai dans les milieux ruraux. Or, Harry Truman était beaucoup plus populaire en milieu rural que son adversaire. Ces échantillons « aléatoirement choisis » violaient le concept de la « chance égale » en excluant de nombreux électeurs ruraux qui n'avaient pas le téléphone et qui étaient des électeurs qui appuyaient Truman. Les chances que les électeurs soient inclus dans les sondages n'étant pas égales, l'inférence à la population fut erronée, à la grande joie de Truman.

Il est clair que la sélection des échantillons est essentielle à la validité des inférences que les sondages permettent. Mais il serait faux de conclure, à partir de cette anecdote, que les sondages ne veulent rien dire. Empiriquement, même lorsqu'il existe certaines divergences dans les résultats des sondages, en général ceux-ci prédisent fort bien l'issue des élections.

Le critère de l'indépendance des réponses

Le principe de l'indépendance implique que la réponse fournie par une personne (une observation) n'est pas influencée par la réponse fournie par une autre. Voici deux exemples illustrant ce principe et dans lesquels l'indépendance des réponses n'est pas maintenue :

- Lorsqu'un dictateur demande un vote de soutien à main levée, il peut être dangereux de dévoiler son opinion lorsque celle-ci est minoritaire. Les votes ne sont donc pas indépendants, car le vote d'une personne est influencé par celui des autres. L'échantillon de votes (l'opinion exprimée) fournit une estimation biaisée de la population (l'opinion réelle). Dans ce cas, la vraie attitude des électeurs ne sera pas bien représentée et l'inférence vers la population sera erronée.
- Si on voulait examiner le temps d'écoute de la télévision des enfants, on pourrait choisir un échantillon composé d'enfants provenant de la même famille. Or, cet échantillon viole le concept de l'indépendance, car le temps d'écoute d'un enfant sera influencé par le temps d'écoute des autres enfants de sa famille.

Quiz rapide 8.2
Expliquez en quoi l'échantillon tiré par Bre-X viole le critère de la chance égale, celui de l'indépendance ou ces deux critères.

Quiz rapide 8.3
On désire étudier l'écoute de la télévision des familles avec enfants. Quel va être maintenant le sujet d'étude ? Est-ce que la procédure d'échantillonnage décrite ci-dessus va à l'encontre du principe de l'indépendance dans ce cas-ci ?

STATISTIQUES ET PARAMÈTRES

Supposons que nous avons à notre disposition la taille de toutes les femmes du Canada (la population). Nous savons, par ailleurs, que certaines femmes sont plus grandes que d'autres. En supposant que la distribution de la taille est normalement répartie, nous pouvons décrire cette population de tailles en calculant sa moyenne et son écart-type. Lorsqu'on travaille avec des populations, ces informations prennent le nom de *paramètres* auxquels,

par convention, on attribue des lettres de l'alphabet grec. La moyenne est identifiée par μ (mu), l'écart-type par σ (sigma). De même, la corrélation est décrite par le symbole ρ (rho).

Lorsqu'on travaille avec un échantillon, on a aussi la distribution de l'échantillon et, comme pour toutes les distributions, celle-ci peut être décrite par sa moyenne et son écart-type. *Les descripteurs des échantillons prennent le nom de statistiques*. Comme on l'a sans doute remarqué dans les chapitres antérieurs, elles sont identifiées, par convention, par des lettres de l'alphabet latin (M, s, r_{xy}). Le Tableau 8.2 présente les noms et les symboles qui décrivent les caractéristiques des populations et des échantillons.

Tableau 8.2
Descripteurs des populations et des échantillons

	Caractéristiques de la population	***Caractéristiques de l'échantillon***
	Paramètres	**Statistiques**
Moyenne	μ (mu)	M
Écart-type	σ (*sigma*)	s
Corrélation	ρ_{xy} (rho)	r_{xy}

Quiz rapide 8.4
Le test de QI développé par Weschler est tel que le QI moyen est de 100. Est-ce une statistique ou un paramètre ? Doit-on écrire $M_{QI} = 100$ ou $\mu_{QI} = 100$?

En bref, les paramètres font référence à la description des populations et les statistiques font référence à la description des échantillons. Ainsi, les paramètres décrivent ce qui est vrai, alors que les statistiques produisent, à partir d'échantillons, la meilleure estimation de la même réalité.

Nous abordons maintenant le lien entre les statistiques et les paramètres, et ces liens vont servir à tirer des conclusions. *Tout ce que nous allons maintenant étudier est vrai si les échantillons sont aléatoires et s'ils sont tirés de populations normalement distribuées.* Sinon, rien n'est nécessairement vrai. Heureusement, la normalité est une présomption acceptable dans la

majorité des situations et, tant que les deux critères pour leur sélection sont respectés scrupuleusement, les échantillons sont aléatoires[2].

La relation entre les statistiques et les paramètres

Aux chapitres 3 à 5, nous avons vu que nous pouvons décrire une distribution normale si on connaît sa moyenne et son écart-type[3]. Par conséquent, la description de n'importe quelle population normalement distribuée implique une connaissance de ces paramètres (μ, σ). Nous avons également vu que les échantillons sont utiles lorsqu'ils représentent la population de laquelle ils sont extraits. Si l'échantillon est représentatif de la population, cela équivaut à dire que les *statistiques* qui décrivent l'échantillon décrivent aussi les *paramètres* de la population. En pratique, cela veut dire qu'à partir de la description que nous avons de l'échantillon, il est possible de faire une description de la population. En jargon statistique, cela implique que :

La meilleure estimation de μ est M. Formule 8.1a

La meilleure estimation de σ est s. Formule 8.1b

Ces égalités représentent un axiome fondamental pour l'inférence statistique : *la meilleure estimation des paramètres d'une population normalement distribuée est les statistiques de l'échantillon aléatoirement tiré de cette population.* Puisque nous voulons toujours inférer quelque chose au sujet de la population (les paramètres), cet axiome revient à confirmer qu'à partir des statistiques, nous pouvons inférer les paramètres de la population.

Le calcul des paramètres de la population

Le Tableau 8.3 décrit les formules statistiques qui définissent le calcul des paramètres et des statistiques (la moyenne et la dispersion). En pratique,

2. Jusqu'à présent, nous avons fait abstraction de N, le nombre d'observations. En général, plus un échantillon contient d'observations, plus il sera en mesure de représenter adéquatement la population de laquelle il est extrait, à condition, bien entendu, que soient respectés les deux critères pour la sélection aléatoire.
3. Nous présumons, ne l'oublions pas, la normalité de la population. Par conséquent, les paramètres de l'asymétrie ou de la modalité ne sont pas pertinents : la distribution est, par définition, unimodale et symétrique.

il est rarement possible de connaître les paramètres d'une population, car on a rarement accès aux informations provenant d'une population entière. Les calculs présentés ici sont donc essentiellement abstraits. On compare dans ce tableau les deux jeux de formules.

La formule pour calculer la moyenne μ est identique à celle que l'on utilise pour calculer la moyenne de l'échantillon (M), la somme des observations divisée par le nombre d'observations. Mais l'écart-type de la population (σ) est calculé à partir d'une formule qui diffère légèrement de celle utilisée pour calculer la statistique correspondante (s).

Tableau 8.3
Formules de calcul des paramètres et des statistiques

Formules : paramètres	***Formules : statistiques***
$\mu = \sum_{i=1}^{N} X_i / N$	$M = \sum_{i=1}^{N} X_i / N$
$\sigma = \sqrt{\sum_{i=1}^{n} (X_i - \mu)^2 / N}$	$s = \sqrt{\sum_{i=1}^{n} (X_i - M)^2 / (N-1)}$

Le calcul du paramètre de dispersion (l'écart-type σ) exige que la somme des écarts à la moyenne au carré ($[Xi - \mu]^2$) soit divisée par N, le nombre d'observations. Pour la statistique équivalente (l'écart-type s), nous divisons la somme des carrés par N – 1, le nombre de *degrés de liberté*.

Le concept de degré de liberté

Le concept de degré de liberté est cependant parfois difficile à comprendre. Les deux explications suivantes pourraient être utiles pour surmonter cet obstacle.

Explication A. La variance (ou l'écart-type) d'un échantillon est invariablement calculée en divisant la somme des carrés par les degrés de liberté, dans ce cas N – 1. En effet, calculer la variance en divisant par N aura

tendance à donner une estimation de la variance de la population σ trop petite puisque, probablement, certains scores extrêmes ne seront pas dans l'échantillon (alors qu'ils sont dans la population). On dit qu'en divisant par *N,* le calcul de la variance est biaisé, car il donne une réponse généralement trop petite. Pour éviter cela, on agrandit légèrement l'écart-type de l'échantillon, ce qui se fait en réduisant son diviseur. Au lieu de N, on utilise N – 1. Mais pourquoi retirer une observation (N – 1) ? Pourquoi pas deux (N – 2) ou plus (N – 3) ? L'explication B nous donne la réponse.

Explication B. L'écart-type décrit la différence moyenne entre chaque observation et la moyenne de la distribution. L'échantillon extrait de la population la représente bien à condition qu'il soit aléatoirement tiré. La sélection aléatoire implique l'indépendance des observations. Aucune observation ne doit être influencée par une autre, elles doivent toutes être indépendantes. Donc, l'écart-type de l'échantillon représente bien l'écart-type de la population si toutes les différences qui proviennent de l'échantillon sont indépendantes. Or, lorsqu'on calcule la variance des échantillons, *une des différences n'est jamais indépendante* !

Les données du Tableau 8.4 présentent le problème. On y trouve la moyenne d'un échantillon composé de trois observations. La moyenne de ces trois valeurs est M = 2. Le tableau indique la valeur obtenue pour les observations A et B (1 et 2 respectivement), mais la valeur de l'observation C n'est pas indiquée. Qu'est-ce que cette dernière valeur doit *nécessairement* être ?

Si vous êtes capable de déduire la valeur de l'observation C à partir des autres informations disponibles (la moyenne plus les deux autres valeurs dans l'échantillon), cela implique que cette dernière information (C) n'est pas indépendante : sa valeur est déterminée par les autres informations. N'étant pas indépendante, cette observation ne respecte pas l'un des deux principes fondamentaux pour la sélection aléatoire des échantillons. Avant de lire le prochain paragraphe, le lecteur devrait tenter de déduire la valeur que doit prendre l'observation C du Tableau 8.4.

Pour trouver cette valeur, il faut calculer les écarts entre chaque observation disponible et la moyenne des observations. La troisième colonne du tableau montre que les écarts entre les deux premières observations et la moyenne sont respectivement de –1 et 0. Au chapitre 3, nous avons vu que

la somme des écarts entre la moyenne et les valeurs est toujours égale à zéro. La somme des deux premiers écarts étant –1 + 0 = –1, il faut que l'écart de l'observation C soit égal à + 1 (– 1 + 0 + 1 = 0). Si nous ajoutons + 1 à la moyenne, nous obtenons 3, ce qui est, dans ce cas, la valeur que l'observation C prend obligatoirement. Puisque nous avons été capables de déduire la valeur manquante pour l'observation C à partir des autres observations et de la moyenne, cette observation ne peut prendre n'importe quelle valeur et, par conséquent, elle n'est pas indépendante. Cette dernière observation C ne respecte pas le critère de la sélection aléatoire.

Lorsque nous calculons la variance de l'échantillon, nous divisons la somme des écarts au carré par N – 1 afin d'éliminer statistiquement l'influence d'une observation qui n'est pas aléatoire. La correction doit se faire en réduisant la taille de l'échantillon N par une seule observation : N – 1 (et non pas N – 2, ou N – 3) car seule *une* observation est non indépendante. Cette correction — la division par le degré de liberté — maintient la caractéristique complètement aléatoire de l'échantillon, ce qui est essentiel puisque l'inférence à la population n'est valide que lorsque toutes les informations provenant d'un échantillon sont aléatoirement extraites de la population. Voilà pourquoi nous disons que la formule de calcul de la variance (et écart-type) de l'échantillon produit une estimation non biaisée du paramètre de la population (σ).

Tableau 8.4
Concept de degré de liberté

Observation	*Valeur obtenue*	*Écart relatif à la moyenne*
A	1	1 – 2 = –1
B	2	2 – 2 = 0
C	?	? – 2 = ?
Moyenne pour les 3 observations	**2**	–

Quiz rapide 8.5
Calculez la moyenne des valeurs du Tableau 8.4 en définissant « 4 » comme la valeur de l'observation C. La moyenne trouvée est-elle de 2 ? Refaites le calcul, mais cette fois, utilisez la valeur 3 pour l'observation C. La moyenne ainsi calculée est-elle juste ?

Le calcul de la variance d'un échantillon avec ou sans correction pour les degrés de liberté ne fait pas une grande différence lorsqu'on travaille avec de grands échantillons. Mais lorsqu'on le fait avec de petits échantillons, la différence peut être très appréciable. Cela est particulièrement important pour les champs disciplinaires contraints de travailler avec de petits échantillons. Par exemple, les recherches en neuropsychologie ou celles qui expérimentent sur des singes sont généralement limitées à de très petits échantillons et, dans ce cas, la correction pour le degré de liberté est essentielle.

Quiz rapide 8.6
Choisissez la bonne formule de la variance pour les deux cas suivants. Cas 1 : Vous désirez déterminer la variance des notes en statistiques pour votre classe. Cas 2 : Vous désirez déterminer la variance des notes en statistiques pour tous les étudiants de l'université à partir de celles obtenues dans votre cours.

LA THÉORIE, L'HYPOTHÈSE ET LA VÉRIFICATION DE L'HYPOTHÈSE NULLE

La théorie, l'hypothèse et la vérification de l'hypothèse nulle forment le trépied sur lequel repose la méthode scientifique.

- Une *théorie* est une explication de la réalité. Par exemple, la théorie de l'anxiété explique le phénomène « problème de lecture chez les enfants ». Cette représentation de la réalité pouvant être juste ou fausse, il est nécessaire de la vérifier empiriquement.
- La vérification empirique de la théorie exige la mise en place de deux hypothèses, l'*hypothèse* et son inverse, l'*hypothèse nulle*.
- L'hypothèse est une conséquence observable qui découle de la théorie et qui devrait être vraie si la théorie est juste.
- L'hypothèse nulle est l'inverse de l'hypothèse.

Faire la lecture aux petits chiens

Un certain nombre d'enfants éprouvent des difficultés en lecture. Une *théorie* explique que ces enfants évitent la lecture parce que cette activité leur cause de l'anxiété. Inspirée par cette théorie, une psychologue scolaire postule *l'hypothèse* selon laquelle une intervention qui réduirait l'anxiété aurait comme effet d'encourager et d'améliorer la lecture chez cette population d'écoliers. Elle désire *vérifier* son hypothèse. Elle choisit aléatoirement deux *échantillons* de cette population d'enfants. Lors des périodes scolaires consacrées à la lecture, un des deux échantillons, le groupe *expérimental,* va lire un conte pour enfants à un petit chien ! L'autre, *l'échantillon témoin*, suit le programme habituel. À la fin de l'année, le niveau de lecture *moyen* atteint par l'échantillon expérimental est *comparé* à celui du groupe témoin. L'hypothèse est-elle *confirmée ?* L'intervention devrait-elle être *généralisée* à toutes les écoles ?

La terre est-elle ronde ?

Il y a de cela plusieurs milliers d'années, les philosophes grecs ont formulé la théorie selon laquelle la terre était ronde plutôt que plate. De cette théorie découlait l'hypothèse suivante : si la terre est ronde, en observant l'horizon au-dessus de la mer, celui-ci devrait apparaître courbé plutôt que droit. L'hypothèse nulle, dans ce cas, est que l'horizon n'est pas courbé.

Si la théorie est juste, il s'ensuit que l'hypothèse (une prédiction qui découle de la théorie) sera empiriquement vérifiée. Par exemple, l'hypothèse selon laquelle la lecture aux petits chiens améliore la lecture sera vraie si le groupe d'enfants de l'échantillon expérimental obtient de meilleurs résultats en lecture que les enfants du groupe témoin. Si les résultats obtenus dans les deux groupes sont égaux, l'hypothèse n'est pas confirmée, et cela jette un doute sur la théorie qui l'a inspirée. On attribue généralement à l'hypothèse le symbole « H ». Si nous avons plusieurs hypothèses, nous les distinguons avec des numéros (H_1, H_2, etc.). Formulons l'hypothèse pour l'étude portant sur la lecture faite aux petits chiens.

H : Les enfants qui font la lecture aux petits chiens améliorent leur niveau de lecture plus que les enfants qui ne font pas ce type de lecture.

La *vérification* de l'hypothèse est un ensemble de règles qui établissent les conditions sous lesquelles on peut tester l'hypothèse. La façon de procéder est d'établir une *hypothèse nulle*. L'hypothèse nulle est (1) l'inverse de l'hypothèse ; et (2) représente une situation hypothétique précise qui peut être dès lors testée de façon précise. Par exemple, si on pose l'hypothèse que les hommes et les femmes sont de taille différente, alors l'hypothèse nulle

postule que les hommes et les femmes sont de même taille. Cette hypothèse nulle propose une égalité de taille entre les membres des deux sexes. Cette situation précise est facile à tester : il suffit de tirer aléatoirement de la population un échantillon de femmes et un échantillon d'hommes, de les mesurer tous, de calculer la taille moyenne de chaque groupe et de vérifier si ces deux moyennes sont égales ou non.

La notion d'hypothèse est un des piliers de la méthode scientifique. La *vérification* de l'hypothèse est une structure de règles qui établissent les conditions qui doivent être vraies pour « accepter » ou « rejeter » l'hypothèse. *Formellement, les procédures statistiques ne sont pas capables d'indiquer si une hypothèse est « vraie ». En revanche, elles sont tout à fait capables d'indiquer si une hypothèse est « fausse ».* Les statistiques ne nous permettent pas « d'accepter » une hypothèse, mais elles nous permettent de la « rejeter ». Comment alors « confirmer » une hypothèse ? La méthode scientifique propose de jumeler à l'hypothèse H une hypothèse rivale, *l'hypothèse nulle*, H_0 — qui est son inverse.

- Si H prédit qu'il y a une différence ou une corrélation, H_0 prédit toujours qu'il n'y a pas de différence ou de corrélation.
- Si nous rejetons l'hypothèse nulle (H_0 est fausse), son hypothèse inverse (H) doit être vraie (il y a une différence ou une corrélation).
- Si nous ne rejetons pas l'hypothèse nulle (H_0 n'est pas fausse), cela ne voudrait pas nécessairement dire que H est fausse. Nous sommes limités à dire que nous ne pouvons pas accepter H.

Le langage utilisé pour faire une distinction entre le rejet et le non-rejet de l'hypothèse et de l'hypothèse nulle est certes un peu opaque, mais il va au cœur de l'inférence, et avec un peu d'application en étudiant les pages suivantes, on peut le maîtriser.

Appliquons ces règles à l'étude portant sur les troubles de lecture des enfants. Les Formules 8.2a et 8.2b indiquent le jeu d'hypothèses que la méthode scientifique exige. Puisque cette expérience est faite dans le but de tirer une conclusion au sujet de la population (tous les enfants qui ont des troubles de lecture), il faudrait faire une comparaison entre la moyenne des deux populations, μ_E et μ_T. Si notre théorie est juste, la population d'enfants bénéficiant de cette intervention devrait être différente (en ce qui concerne la compétence en lecture) de celle des enfants qui n'en bénéficient pas. Les

Formules 8.2a et 8.2b décrivent symboliquement l'hypothèse et l'hypothèse nulle :

$$H : \mu_E \neq \mu_T \qquad \text{Formule 8.2a}$$

$$H_0 : \mu_E = \mu_T \qquad \text{Formule 8.2b}$$

où H et H_0 sont respectivement l'hypothèse et l'hypothèse nulle et μ_E et μ_T sont les moyennes des populations des enfants qui font (μ_E) ou ne font pas (μ_T) la lecture aux petits chiens.

Quiz rapide 8.7
Vous observez plusieurs voitures qui roulent sur l'autoroute à 100 km/h. Or, plus loin sur la route, il y a des gravats sur la chaussée. Lorsque les chauffeurs les verront, que croyez-vous qu'ils feront ? Quelle est votre hypothèse ? Votre hypothèse nulle ? Pouvez-vous les écrire avec des symboles ?

Cependant, nous ne pouvons pas calculer la moyenne de la population puisque nous n'avons pas accès à la population d'observations. En revanche, nous savons que la meilleure estimation de la moyenne de la population est la moyenne de l'échantillon. Si la théorie est juste et si les échantillons sont tirés aléatoirement, les enfants qui participent à l'expérience (le groupe expérimental, noté E) n'obtiendront pas le même résultat en lecture que le groupe qui n'y participe pas (le groupe témoin, noté T). La moyenne étant la meilleure estimation de performance en lecture de chaque distribution, nous devons alors comparer la moyenne en lecture obtenue par chaque échantillon d'enfants (M_E et M_T).

- Lorsque les moyennes des deux groupes sont très dissemblables, nous rejetons l'hypothèse nulle (elle est fausse), ce qui nous contraint à accepter son opposée : H. Celle-ci est vraie, ce qui appuie la théorie et renforce ainsi notre confiance en sa véracité.
- Lorsque les moyennes des deux groupes sont les mêmes, nous ne pouvons pas rejeter l'hypothèse nulle (nous ne pouvons pas conclure qu'elle est fausse). Puisque nous ne pouvons rejeter H_0, nous ne pouvons pas accepter H. *Cependant, nous n'avons pas démontré que H est fausse, seulement qu'il n'y a pas de preuves qu'elle soit vraie.*

Tout cela mène aux deux conclusions suivantes :

- Si H_0 est rejetée (fausse), H est nécessairement vraie.
- Si H_0 n'est pas rejetée (n'est pas fausse), la preuve que H est vraie n'est pas établie, mais H n'est pas nécessairement fausse.

Les exemples suivants illustrent cette importante subtilité.

Quiz rapide 8.8
Le philosophe Montesquieu formula la théorie suivante : les conditions climatiques ayant un impact sur le tempérament des humains, les habitants des pays nordiques sont moins émotifs que ceux des pays plus chauds. Élaborez une hypothèse et une hypothèse nulle empiriquement vérifiables qui découleraient de cette théorie.

Exemples d'hypothèses et d'hypothèses nulles

Les illustrations suivantes serviront à bien saisir les nuances importantes entre l'hypothèse et l'hypothèse nulle ainsi que les conclusions auxquelles elles mènent. Les licornes sont ces chevaux mythologiques qui portent une corne au milieu du front. On aimerait prouver que les licornes existent. On établit donc un jeu d'hypothèses comprenant l'hypothèse (H) et son opposée, l'hypothèse nulle (H_0).

H : Les licornes existent, c'est-à-dire que le nombre de licornes $\neq 0$.

H_0 : Les licornes n'existent pas, c'est-à-dire que le nombre de licornes $= 0$.

On lance une expédition pour trouver des licornes en fouillant toutes les capitales européennes, les savanes africaines et les forêts tropicales. En vain. Peut-on affirmer que les licornes n'existent pas ? Peut-être existent-elles en Arctique ou peut-être sont-elles plus capables de se cacher que vous ne l'êtes de les trouver ? On ne peut pas accepter H (elles existent), mais on ne peut pas plus accepter H_0 (elles n'existent pas). On peut remarquer la subtilité : notre intention était de prouver que les licornes existent. Or, les données ne confirment pas leur existence, mais elles ne démontrent pas qu'elles n'existent pas : c'est le *statu quo ! Dans ce cas, nous n'avons aucune preuve de l'existence des licornes, mais on ne peut pas conclure que les licornes n'existent pas.*

Utilisons nos symboles pour formaliser notre quête. Dans notre échantillon, nous n'avons pas trouvé de licornes. Nous concluons alors :

Non-rejet de H_0 : il n'y a pas de preuves que les licornes existent (H n'est pas prouvé).

Il ne reste qu'à demander une nouvelle subvention au gouvernement afin de poursuivre nos recherches sur l'existence des licornes…

Supposons que l'on trouve une licorne (elle se cachait en Provence). L'hypothèse nulle (H_0) est maintenant rejetée, car au moins une licorne existe. Par conséquent, on doit accepter H et conclure que les licornes existent. Utilisons nos symboles pour tirer cette conclusion :

Rejet de H_0 : les licornes existent !

Prenons un deuxième exemple : l'évaluation de l'intervention portant sur les difficultés de lecture des enfants.

L'hypothèse nulle stipule que la lecture aux petits chiens n'améliore pas la compétence en lecture des enfants. Dans ce cas, elle prévoit que la moyenne pour le groupe témoin et le groupe expérimental est la même : $H_0 : M_E = M_T$ (ce qui implique $\mu_E = \mu_T$).

L'hypothèse (H), quant à elle, avance que la lecture aux petits chiens améliore la compétence en lecture des enfants. C'est-à-dire que la compétence en lecture moyenne des deux groupes ne sera pas la même : $H : M_E \neq M_T$ (ce qui implique que $\mu_E \neq \mu_T$).

Il ne reste qu'à examiner les résultats de l'expérience. Si la moyenne en lecture obtenue par les enfants est sensiblement égale pour les deux groupes, on ne peut pas rejeter H_0. Puisqu'on ne rejette pas H_0, on ne peut pas accepter H et conclure que la lecture aux petits chiens favorise la compétence en lecture. Mais on ne peut pas conclure qu'elle ne la favorise pas. En effet, la théorie nous indique que le stress est la cause des problèmes de lecture chez les enfants, et l'hypothèse propose que la lecture aux petits chiens réduit le stress, cette réduction de stress améliorant la lecture. Puisque nous ne pouvons pas rejeter l'hypothèse nulle, nous ne pouvons pas conclure que la théorie est juste. Mais nous ne pouvons pas assurément conclure, sur la seule base de cette étude, que la théorie est fausse. Peut-

être que d'autres techniques de réduction du stress pourraient favoriser la compétence en lecture, ou peut-être que la lecture aux grands chiens plutôt qu'aux petits chiens aurait plus d'effets.

Ainsi, l'hypothèse nulle ne peut jamais être démontrée. Cette règle logique est applicable à toutes les formulations d'hypothèses. L'hypothèse peut cependant être appuyée de deux façons : a) si les enfants qui reçoivent l'intervention deviennent meilleurs que les enfants qui ne la reçoivent pas ; ou b) si les enfants qui ne reçoivent pas l'intervention sont supérieurs à ceux qui la reçoivent ! Cette distinction entre les hypothèses est abordée dans le prochain chapitre dans la section portant sur les hypothèses *unicaudales* vs les hypothèses *bicaudales*.

Le texte suivant illustre ces concepts en présentant un exemple politico-militaire réel.

L'hypothèse nulle et la guerre en Irak

L'histoire de la guerre en Irak offre une illustration frappante que l'hypothèse nulle ne peut jamais être prouvée.

En 2003, l'armée américaine envahit l'Irak et justifie son attaque en affirmant que ce pays possède des armes de destruction massive (ADM). Les inspecteurs de l'ONU affirment le contraire : L'Iraq ne possède pas d'ADM. Qui a raison ? Pouvons-nous confirmer l'hypothèse américaine (H) et prouver que celle de l'ONU, l'hypothèse nulle (H_0), est fausse ?

Nous avons deux hypothèses rivales :

H_0 : Le nombre d'ADM en Iraq = 0 (hypothèse nulle de l'ONU).

H : Le nombre d'ADM en Iraq $\neq$ 0 (hypothèse américaine).

À la suite de l'invasion, les troupes américaines lancent des fouilles, mais ne trouvent aucune ADM. Puisque nous ne pouvons pas rejeter H_0, nous ne pouvons pas accepter H, l'hypothèse des Américains. Mais l'ONU ne peut pas plus affirmer que H_0 est vraie et conclure qu'il n'y a pas d'ADM en Irak. Après tout, il est possible qu'il y en ait en Iraq, mais qu'elles n'aient pas encore été découvertes.

L'ONU peut, par contre, affirmer qu'il n'existe pas de preuve voulant que l'hypothèse de l'armée américaine soit vraie. Il est faux de conclure qu'il n'existe pas d'ADM en Irak, mais il est juste de conclure qu'il n'y a aucune preuve de leur existence.

Supposons que l'on découvre une seule ADM dans ce pays. On pourrait alors rejeter H0 et, ce faisant, on serait contraints d'accepter H. Ainsi, on rejette ou on ne rejette pas H0. Mais on ne peut jamais l'accepter !

La fluctuation dans les échantillons aléatoires

Supposons que la note moyenne obtenue par les étudiants à un examen est de 70 %; supposons également que l'on prend aléatoirement cinq étudiants du cours et que l'on calcule leur moyenne à l'examen. Est-ce que la moyenne obtenue par ces étudiants sera de 70 % ? Il est fort probable que non. Leur note moyenne sera au moins un peu différente. Supposons maintenant que l'on choisit (aléatoirement) cinq autres étudiants et que nous calculons la note moyenne obtenue par ce deuxième échantillon de cinq personnes. Cette moyenne sera-t-elle égale à 70 % ? Encore une fois, il y a fort à parier qu'elle ne sera pas exactement de 70 %. Deux échantillons tirés de la même population peuvent avoir des moyennes différentes. Chaque échantillon provient d'une population. Dans la population, les observations individuelles se répartissent à travers les valeurs de la variable. Sous la présomption d'une distribution normale, par exemple, la plupart des observations seront proches de la moyenne (μ) alors que d'autres observations, certes plus rares, mais néanmoins présentes, se situeront plus loin de la moyenne. Puisque les échantillons sont aléatoirement extraits de la population, il est quasi certain que presque tous contiendront une proportion au moins légèrement différente d'observations proches ou éloignées de la moyenne. Par conséquent, les échantillons seront tous au moins un peu différents les uns des autres. Puisque la moyenne est tributaire des observations que l'échantillon contient, il est tout aussi quasi certain que la moyenne de plusieurs échantillons extraits de la même population sera minimalement quelque peu différente.

Dans ces conditions, obtenir deux échantillons ayant très précisément la même moyenne est virtuellement impossible. Cette fluctuation naturelle dans les échantillons aléatoires s'appelle l'*erreur d'échantillonnage*. Une analogie avec une pièce de monnaie permet de mieux saisir le problème.

Nous savons que lorsque nous jouons à pile ou face, nous avons autant de chances d'obtenir face que pile et ce, à n'importe quel lancer. Nous pouvons exprimer ce constat en disant que dans la population, les piles et les faces sont également fréquentes, ce qui implique que la probabilité d'obtenir un lancer face est égale à celle d'obtenir un lancer pile, et que la probabilité de chacun est de 0,50. Nous concluons que la moyenne de la

population de faces est $\mu_{face} = 0{,}50$. Imaginons la toute première fois qu'une extraterrestre joue à pile ou face. Elle désire estimer si les piles et les faces sont également probables. Elle pose son problème sous la forme d'hypothèses :

$H_0 : \mu_f = \mu_p$ (piles et faces dans la population sont également fréquentes).
$H : \mu_f \neq \mu_p$ (piles et faces dans la population ne sont pas également fréquentes).

Elle conçoit une expérience qui lui permettra de trancher. Elle constitue un premier échantillon d'observations en lançant une pièce de monnaie dix fois. Elle note les résultats obtenus en codant pile = 0 et face = 1. Elle calcule le nombre moyen de fois ou la pièce retombe sur face dans cet échantillon de 10 lancers. Si la moitié des lancers donne face, la moyenne sera 0,50. Si 30 % des lancers donnent des faces, la moyenne sera 0,30.

La première ligne du Tableau 8.5 présente ses résultats. Son premier échantillon possède un nombre égal de lancers piles et faces ($M_f = M_p = 0{,}50$). L'extraterrestre conclut alors au non-rejet de H_0 : vraisemblablement, elle n'aurait pas de bonnes raisons de croire que la population de piles et de faces n'est pas égale à 0,50.

Pour avoir plus de certitude, elle répète l'expérience neuf autres fois et calcule la moyenne de faces pour chacune des neuf expériences, chacune étant composée de 10 lancers. Si elle trouve que la moitié des lancers donne des faces, la moyenne de faces sera égale à 0,5 et, par conséquent, elle ne pourra pas rejeter H_0. Si elle trouve une moyenne autre que 0,5, elle pourra alors rejeter H_0.

À sa grande surprise, son deuxième échantillon (ligne 2 au Tableau 8.5) la force à réviser sa conclusion. Ici, $M_f \neq M_p$. Elle se doit de tirer la conclusion inverse et de rejeter H_0 ($\mu_f \neq 0{,}50$). L'examen des autres expériences ne fait qu'augmenter la confusion. Parce que les échantillons ne produisent pas tous les mêmes résultats, ils produisent des conclusions différentes, rejet ou non de H_0 (on remarque les résultats diamétralement opposés des échantillons 9 et 10).

Tableau 8.5
Moyenne de faces dans 10 échantillons tirés aléatoirement d'une population de résultats à pile ou face

Échantillon	*Nombre de faces sur 10 lancers*	*Proportion de faces*	*Décision*
1	5	5/10 = 0,5	non-rejet de H_0
2	3	3/10 = 0,3	rejet de H_0
3	6	6/10 = 0,6	rejet de H_0
4	8	8/10 = 0,8	rejet de H_0
5	4	4/10 = 0,4	rejet de H_0
6	5	5/10 = 0,5	non-rejet de H_0
7	4	4/10 = 0,4	rejet de H_0
8	6	6/10 = 0,6	rejet de H_0
9	0	0/10 = 0,0	rejet de H_0
10	10	10/10 = 1,0	rejet de H_0
Total pour les 10 échantillons	51	51/100 = 0,51	rejet de H_0

Quelle est alors l'estimation la plus raisonnable de la proportion de piles et de faces dans la population ? Celle qui provient de l'échantillon 2 ? 3 ? 7 ? etc. ? La confusion de l'extraterrestre vient du fait qu'elle s'attend à retrouver dans l'échantillon très précisément ce qu'elle suppose dans la population. Elle ignore que les différents échantillons sont assujettis à l'erreur d'échantillonnage. Les moyennes des échantillons diffèrent plus ou moins les unes des autres et la plupart ne tomberont pas exactement sur la moyenne de la population. L'erreur d'échantillonnage est inévitable, car la moyenne de l'échantillon n'est qu'une estimation de la moyenne de la population.

À la dernière minute, l'extraterrestre se souvient d'un principe important : *les échantillons plus grands produisent une estimation plus précise de la moyenne de la population.* Elle calcule la moyenne du nombre de faces basée sur les 100 lancers. Elle obtient $M_f = 0,51$. Elle estime alors que 0,51 est une excellente estimation de la moyenne réelle. La conclusion à rete-

nir est qu'accroître la taille de l'échantillon augmente le degré de précision dans l'estimation que les statistiques font des paramètres. À la limite, un échantillon de taille égale à la taille de la population est un estimateur parfait.

En utilisant la moyenne basée sur la totalité des échantillons ($M_f = 0{,}51$ au Tableau 8.5), l'extraterrestre est tentée de rejeter H_0 qui prévoit que $\mu_f = 0{,}50$. Mais elle a retenu sa leçon : la différence entre la proportion de faces (0,51) et la proportion de piles (0,49) ne serait-elle pas attribuable à l'erreur d'échantillonnage ? Supposons qu'elle décide néanmoins de rejeter l'hypothèse nulle : sa décision pourrait être justifiée, car la proportion de faces qu'elle a obtenue, même après 100 lancers, n'est pas exactement 0,50, comme le stipule l'hypothèse nulle. Mais on sait que la proportion de piles et de faces est en réalité égale. En rejetant H_0, notre extraterrestre va dans ce cas commettre une erreur d'inférence.

Les erreurs d'inférence

Les joueurs de pile ou face savent que la vraie moyenne de faces dans la population est en réalité de 0,50. Pourtant, cette valeur ne se retrouve pas dans 80 % des échantillons du Tableau 8.5 ! À chaque fois que l'extraterrestre conclut au rejet de H_0, elle fait une *erreur d'inférence :* elle conclut à partir d'un échantillon qu'il existe une différence entre le nombre de piles et le nombre de faces alors qu'en réalité, il n'en existe pas dans la population. Cette erreur prend le nom d'*erreur de type I* (ou d'*erreur alpha*).

L'erreur de type I (erreur alpha) consiste à conclure à partir des échantillons qu'il existe une différence dans la population alors qu'il n'en existe pas.

Mais supposons que la pièce de monnaie est truquée : la vraie moyenne de faces pour cette pièce étant $\mu = 0{,}40$. À partir de deux échantillons (1 et 6 au Tableau 8.5), nous aurions conclu que le nombre de piles et de faces est égal, alors qu'en réalité, il ne l'est pas. Dans ce cas, nous faisons une erreur d'inférence qui se nomme *erreur de type II* ou *erreur bêta*.

L'erreur de type II (erreur bêta) consiste à conclure à partir des échantillons qu'il n'existe pas de différence dans la population alors qu'il en existe une.

Au Tableau 8.6, nous voyons qu'à partir des résultats obtenus dans l'échantillon, on peut soit conclure au rejet de H_0, soit conclure à son non-rejet. Si l'on rejette H_0 et qu'il existe une différence dans la population, notre conclusion est juste. Si l'on ne peut pas rejeter H_0 et qu'il n'y a pas de différence dans la population, notre conclusion est juste aussi. Mais si nous rejetons H_0 et qu'il n'y a pas de différence dans la population, notre conclusion est fausse : il s'agit d'une erreur de type I (alpha). Si nous concluons qu'il n'y a pas de différence, alors qu'elle existe en réalité dans la population, nous faisons une erreur de type II (bêta).

Tableau 8.6
Erreurs d'inférence de type I (α) et de type II (β)

La conclusion consiste à	***La différence réelle dans la population***	
	existe ($\mu_1 \neq \mu_2$).	**n'existe pas ($\mu_1 = \mu_2$)**
rejeter H_0 (car $M_1 \neq M_2$)	Conclusion juste	Erreur de type I (erreur alpha)
ne pas rejeter H_0 (car $M_1 = M_2$)	Erreur de type II (erreur bêta)	Conclusion juste

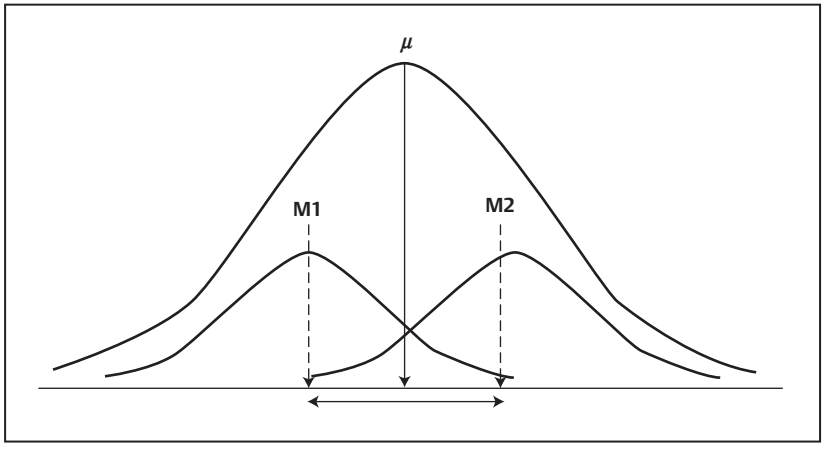
Quiz rapide 8.9
Un chercheur s'intéresse à la relation entre le tabagisme et le cancer. Il injecte de la fumée de cigarette dans la cage de trois rats (groupe avec traitement). À trois autres rats aléatoirement choisis, il n'injecte pas de fumée (groupe sans traitement). Trois semaines plus tard, il exécute des biopsies sur les rats et observe qu'aucun rat n'est atteint de cancer. Il ne rejette pas l'hypothèse nulle et conclut que la fumée de cigarette ne cause pas le cancer. Quelles étaient ses hypothèses ? Êtes-vous en accord ou en désaccord avec cette conclusion ? Pourquoi ?

Une ou plusieurs populations ?

La Figure 8.1 représente le polygone des effectifs de deux échantillons indépendants et aléatoires provenant, présumons-le, de la même population. En sélectionnant un échantillon au hasard, les unités d'observations qui composent l'échantillon peuvent provenir de n'importe quelle partie de la distribution, y compris de ses extrêmes. Par pur hasard, certains échan-

tillons contiendront plus d'observations se trouvant aux extrêmes de la distribution et certains autres moins, ce qui causera une différence entre leurs moyennes. La Figure 8.1 montre une population dont la moyenne est μ (indiquée par la flèche centrale). De cette population, deux échantillons (M_1 et M_2) sont aléatoirement tirés. La moyenne de chacun de ces deux échantillons est indiquée par une flèche en pointillé. La double flèche de la Figure 8.1 représente la différence entre les moyennes des deux échantillons.

FIGURE 8.1 Moyennes de deux échantillons (M1 et M2) extraits de la même population dont la moyenne est μ



Lorsque nous tirons un échantillon aléatoire d'une distribution normale et que nous calculons sa moyenne, nous obtenons du même coup « la meilleure estimation » de la moyenne de la population. Un deuxième échantillon aléatoire extrait de cette même population produira lui aussi une moyenne qui sera « la meilleure estimation » de la population. Mais il est certain que cette deuxième moyenne ne sera pas identique à celle du premier échantillon (Figure 8.1). Puisque nous présumons qu'il n'existe qu'une seule moyenne dans la population[4], quelle est la « bonne » estimation de la moyenne de la population : celle qui provient du premier ou

4. Il ne faut pas oublier que nous présumons que la distribution de la population est normale. Les distributions normales sont unimodales, ce qui implique que chaque distribution n'aura qu'une seule moyenne.

du deuxième échantillon ? Comment interpréter cette différence entre les moyennes des échantillons ? Deux interprétations sont plausibles.

Interprétation 1. Il y a une erreur dans l'un ou l'autre des échantillons, ou dans les deux lorsque l'on fait une estimation de la moyenne de la population μ. Dans ce cas, la différence entre les moyennes des échantillons est attribuable à l'erreur d'échantillonnage.

Interprétation 2. Il n'y a *pas* d'erreur d'estimation de la moyenne μ dans les deux échantillons. Puisque chaque population ne peut avoir plus d'une moyenne, il faut conclure que les échantillons proviennent de populations différentes !

Afin de trancher entre ces deux interprétations, nous allons faire appel au concept de *l'erreur type de la moyenne*. Ce concept est directement tributaire du concept de l'erreur d'échantillonnage : puisque les échantillons extraits de la même population ne sont quasi jamais identiques, il s'ensuit que les moyennes de ces échantillons ne le seront pas non plus. La fluctuation naturelle des moyennes des échantillons, provoquée par l'erreur d'échantillonnage, se calculera éventuellement dans une nouvelle statistique, l'erreur type de la moyenne. L'erreur type de la moyenne est donc la différence « typique » que l'on trouve dans les moyennes de plusieurs échantillons tirés de la même population. Le prochain chapitre expliquera la mécanique de cette valeur, mais pour l'instant, tenons-nous-en à ses principes.

Supposons que nous avons le QI de deux échantillons de jeunes, un groupe de jeunes qui portent des chandails verts (V), l'autre des chandails bleus (B). Nous désirons savoir si ceux qui portent des chandails verts ont des niveaux de QI différents de ceux qui portent des chandails bleus. Établissons d'abord nos hypothèses

H_0 : Le QI moyen des deux populations est égal ($\mu_v = \mu_b$).
H : Le QI moyen des deux populations n'est pas égal ($\mu_v \neq \mu_b$).

Si nous concluons H, cela voudra dire que la population de QI pour les « verts » n'est pas la même que la population de QI des « bleus ». Nous avons deux populations distinctes.

Si nous concluons H_0, nous n'avons pas de raison de croire que les jeunes vêtus de vert n'appartiennent pas à la même population de QI que ceux qui portent des chandails bleus.

Nous trouvons les résultats suivants : $M_v = 105$, $M_b = 95$. La différence est de 10 points de QI. Maintenant, supposons que l'on a l'information suivante : il arrive fréquemment que les moyennes de deux échantillons de QI tirés de la *même* population diffèrent par 10 points de QI. Autrement dit, la différence typique entre deux échantillons extraits de la même population est de 10. Puisque la différence que vous avez observée entre les deux moyennes est égale à la différence que l'on observe typiquement entre deux moyennes tirées de la même population, nous n'avons pas de base raisonnable pour conclure que la différence observée révèle l'existence de deux populations. Par conséquent, il faut conclure à l'absence de preuves voulant que les verts et les bleus proviennent de populations différentes, et on ne peut pas rejeter H_0. Ce faisant, il n'est pas possible d'accepter H.

Mais supposons que la différence typique entre deux échantillons — l'erreur d'échantillonnage — est de 2. Maintenant, la différence observée entre les deux moyennes (10) est bien plus grande que la différence « typique », ce qui permet de rejeter H_0 et donc de conclure que les deux échantillons proviennent de populations différentes.

Cette comparaison de la taille de la différence entre les moyennes des échantillons et l'erreur type de la moyenne forme la base de nombreux tests statistiques, et c'est à partir de cette comparaison qu'il sera possible de trancher entre l'hypothèse et l'hypothèse nulle. La mécanique de cette comparaison est l'objet du chapitre 9.

Les hommes viennent de Mars, les femmes viennent de Vénus

John Gray, l'auteur de *Les hommes viennent de Mars, les femmes viennent de Vénus*, ne se doutait pas que son best-seller servirait un jour d'exemple dans un ouvrage de statistique ! Dans son livre, John Gray illustre les nombreuses différences qui existent entre les hommes et les femmes dans leur façon de voir la vie. En langage statistique, il nous dit que les hommes et les femmes forment deux populations (l'une de Mars, l'autre de Vénus). Vérifions son hypothèse en choisissant une variable (disons la tolérance

au désordre, par exemple) sur laquelle les Martiens et les Vénusiennes devraient différer si la théorie de Gray est juste.

Les hypothèses sont :

H : La tolérance des hommes et des femmes face au désordre n'est pas la même. Ils font partie de deux populations différentes (Mars et Vénus).

H_0 : La tolérance des hommes et des femmes face au désordre est la même. Ils font partie de la même population (les Terriens ?).

Nous choisissons aléatoirement un groupe d'hommes et un groupe de femmes et nous plaçons chaque membre de chaque groupe dans une salle en désordre. Après une heure, nous demandons à chaque personne de décrire son expérience. Au cours d'une entrevue de 5 minutes, nous comptons pour chaque personne le nombre de fois qu'elle parle du désordre dans la salle.

Les hypothèses sont donc :

H : Les hommes et les femmes ne mentionnent pas le désordre également ($\mu_h \neq \mu_f$).

H_0 : Les hommes et les femmes mentionnent le désordre également ($\mu_h = \mu_f$).

Les données de notre expérience sont les suivantes : en moyenne, les hommes font 4 fois mention du désordre ($M_h = 4$), alors qu'en moyenne, les femmes en font mention 8 fois ($M_F = 8$). La différence entre ces deux moyennes est de 4 mentions de désordre. Supposons que la différence typique est de 5, cela veut dire que deux échantillons provenant de la même population peuvent avoir des moyennes qui diffèrent par 5. Autrement dit, il est possible de trouver une différence de 5 points entre deux échantillons de femmes ou entre deux échantillons d'hommes. Puisque nous avons trouvé une différence de 4 et que la différence typique est de 5, nous ne pouvons pas rejeter l'hypothèse nulle et conclure que nous avons deux populations. Nous n'avons pas de réelles évidences que les hommes viennent de Mars et les femmes de Vénus. Dans ce cas, nous ne pouvons pas rejeter H_0.

Supposons que la différence obtenue entre les hommes et les femmes est de 10 points. Nous voyons maintenant que la différence typique (5 points) est beaucoup plus petite que la différence observée (10 points). Nous pouvons désormais rejeter H_0 et, par conséquent, nous acceptons H. Eh oui, les hommes et les femmes viennent de deux planètes différentes ou, statistiquement parlant, ils appartiennent à des populations différentes !

SOMMAIRE DU CHAPITRE

L'analyse statistique sert à tirer une conclusion au sujet d'une population à partir d'un échantillon qui en est aléatoirement extrait. La population représente toutes les valeurs qui existent sur une variable, alors que l'échantillon fait référence à un sous-ensemble de cette information. Lorsque les échantillons sont aléatoirement tirés d'une population, ils ont de bonnes chances de la représenter adéquatement. Un échantillon est aléatoire lorsque chaque individu de la population détient une chance égale de faire partie de l'échantillon et lorsque la réponse de chacun n'est pas influencée par la réponse des autres. Les statistiques décrivant ces échantillons fournissent une estimation des paramètres de la population et, de ce fait, en donnent une description. Nous élaborons une hypothèse que nous comparons à son opposée, l'hypothèse nulle. La vérification des hypothèses implique la collecte d'informations auprès d'échantillons et le test de l'hypothèse consiste à comparer les échantillons entre eux. Mais la simple comparaison entre les moyennes des échantillons n'est pas directement interprétable, car il existe une fluctuation naturelle entre tous les échantillons, qu'ils soient ou non extraits de la même population. Cette fluctuation naturelle, l'erreur d'échantillonnage, est essentielle pour l'interprétation des résultats. Enfin, lorsque nous tirons des conclusions au sujet des échantillons, celles-ci peuvent parfois être erronées, ce qui nous amène à tirer des conclusions au sujet des échantillons qui ne sont pas juste non plus. Ces erreurs d'inférence portent le nom d'erreur de type I et d'erreur de type II.

EXERCICES DE COMPRÉHENSION

1. Vous avez à votre disposition l'âge et la taille des étudiants qui se trouvent dans votre cours. Vous désirez tirer des conclusions au sujet de leur âge et de leur taille. Dans ce cas,vous seriez en train de décrire _____.
   a) un échantillon
   b) une population
   c) une population inférée à partir d'un échantillon
   d) un échantillon inféré à partir d'une population
2. Vous postulez que les étudiants sont de plus grands consommateurs de bière que les étudiantes. Quelle serait la formulation de l'hypothèse nulle ?
3. La taille de l'échantillon est exactement de la taille de la population. L'erreur type d'échantillonnage sera alors _____.
   a) exactement égale à l'écart-type de la distribution
   b) approximativement égale à l'écart-type de la distribution
   c) exactement égale à zéro
   d) approximativement égale à zéro
4. L'échantillon A est extrait d'une population ayant une très petite variance. L'échantillon B est de la même taille que l'échantillon A, et il est extrait de la même population. L'erreur d'échantillonnage dans ce cas sera probablement _____.
   a) est probablement petite
   b) est probablement grande
   c) peut être grande ou petite
   d) est impossible à déduire avec les informations disponibles
5. La différence entre les moyennes sur la variable X produites par deux échantillons est plus petite que l'erreur type de la moyenne.
   a) Il est certain alors que nous pouvons rejeter H_0.
   b) Il est certain alors que nous ne pouvons pas rejeter H_0.
   c) Il est certain alors que les deux échantillons proviennent obligatoirement d'une seule population.
   d) Selon les tests statistiques, toutes ces réponses peuvent être justes.

6. Nous avons un échantillon (E) composé de 100 individus et nous avons une population (P) composée de 100 individus. Nous désirons calculer la variance de chacune de ces distributions. Le dénominateur de la formule de calcul sera _____ pour la distribution E et il sera _____ pour la distribution P.

a) 99 ; 99
b) 100 ; 100
c) 100 ; 99
d) 99 ; 100

7. Nous concluons au rejet de H_0 à partir de nos échantillons. Malheureusement notre conclusion est erronée. Par conséquent, nous venons de faire une erreur d'inférence de type ____.

a) I
b) II
c) I, si le nombre d'observations est petit
d) II, si le nombre d'observations est grand

8. En nous basant sur les échantillons, nous concluons qu'ils ne proviennent pas de populations différentes. Malheureusement, notre conclusion est erronée. Par conséquent, nous venons de faire une erreur d'inférence de type ____.

a) I
b) II
c) I, si le nombre d'observations est petit
d) II, si le nombre d'observations est grand

9. La différence entre les moyennes de deux échantillons est deux fois plus grande que l'erreur type de la moyenne. Nous pouvons alors conclure qu'il _____.

a) est certain que les deux échantillons proviennent de la même population
b) est certain que les deux échantillons proviennent de populations différentes
c) y a de bonnes chances que les deux échantillons proviennent de la même population
d) y a de bonnes chances que les deux échantillons proviennent de populations différentes

Réponses

1. b
2. Les étudiants et les étudiantes font une consommation égale de bière.
3. c : Puisque l'échantillon contient toutes les personnes de la population, l'échantillon et la population sont identiques. Il ne peut pas exister, dans ce cas, une erreur d'échantillonnage.
4. a : La variance de la population étant petite, la différence entre les observations de cette population est petite. Les deux échantillons seront alors composés d'observations très similaires, créant une petite erreur d'échantillonnage.
5. b
6. d
7. a
8. b
9. d : Nous ne pouvons pas choisir b parce qu'une erreur d'inférence, dans ce cas de type I, est toujours possible.

CHAPITRE 9
LA MÉCANIQUE DE L'INFÉRENCE STATISTIQUE

CHAPITRE 9

LA MÉCANIQUE DE L'INFÉRENCE STATISTIQUE

La vérification d'une hypothèse implique qu'on l'oppose à une hypothèse nulle. Nous devons décider si les échantillons proviennent ou ne proviennent pas de la même population. Dans ce chapitre, nous présentons les procédures et les conventions qui permettent de rejeter ou non l'hypothèse nulle. Nous ne rejetons pas l'hypothèse nulle lorsque les échantillons obtiennent les mêmes moyennes, et, dans le cas inverse, nous la rejetons. Mais nous avons vu au chapitre précédent que les échantillons, même lorsqu'ils sont extraits de la même population, n'ont pas exactement la même moyenne. Il existe une variation naturelle dans la composition des échantillons, cette variation étant attribuable à l'aléa. Ainsi, une simple différence entre les échantillons ne peut pas être interprétée directement pour choisir l'une ou l'autre des hypothèses, H ou H_0, puisque l'aléa pourrait en être responsable. Par conséquent, il devient impératif de quantifier cette variation naturelle, c'est-à-dire la différence typique à laquelle nous pouvons nous attendre entre deux échantillons lorsque les deux sont extraits de la même population ou entre la moyenne de la population et la moyenne d'un échantillon. Nous allons rejeter l'hypothèse nulle lorsque la différence observée entre les échantillons est « nettement plus grande » que cette différence « typique » entre les échantillons ou entre l'unique échantillon et la moyenne de la population. Pour cela, il nous faudra un critère qui nous aidera à distinguer une différence « nettement plus grande » d'une différence « typique ».

Est-ce que le niveau de toxines dans les rivières québécoises dépasse les normes ? Pour répondre à cette question, nous devons mesurer le niveau de toxines dans un échantillon de rivières que nous allons comparer avec la norme, qui, dans ce cas représente la population. Ici, un seul échantillon est requis car nous connaissons la moyenne dans la population (la norme). Par contre, dans le cas suivant, il faut constituer deux échantillons : est-ce que les rivières québécoises sont plus polluées que les rivières ontariennes ? Maintenant, nous devons cueillir deux échantillons de rivières, un provenant de la population de rivières québécoises, l'autre provenant de la population de rivières ontariennes. La question statistique, dans ce dernier cas, revient à déterminer si les deux échantillons de rivières (québécoise et ontarienne) ont une forte ou une faible chance de provenir de la même population de pollution.

Dans ce chapitre, nous allons voir la procédure statistique qui permet de calculer deux statistiques importantes : l'*erreur type de la moyenne* et l'*intervalle de confiance autour de la moyenne*. La confrontation des hypothèses H et H_0 — et le concept de la *signification statistique* — découle de ces considérations. La maîtrise des éléments discutés dans ce chapitre est déterminante pour la maîtrise des chapitres subséquents et elle exige la compréhension des chapitres antérieurs, en particulier le chapitre 8 ainsi que le chapitre 5 qui porte sur la distribution normale.

QUAND LES ÉCHANTILLONS ALÉATOIRES NE SONT PAS IDENTIQUES : L'ERREUR D'ÉCHANTILLONNAGE

Imaginons une population d'observations distribuées normalement. Nous savons (voir le chapitre 5) que la majorité des observations (environ 68 %) se trouvent près de la moyenne de la population ($\mu \pm 1$ écart-type) et qu'environ 32 % des observations se trouvent plus loin. Par exemple, dans une population normale ayant 100 et 15 respectivement comme moyenne et comme écart-type, environ 68 % des observations se situent entre 85 et 115 et environ 32 % des observations sont inférieures à 85 et supérieures à 115.

Tirons de cette population plusieurs échantillons de taille identique. Tous ces échantillons étant extraits de la même population, nous nous attendons à ce que chacun soit composé de 68 % d'observations relative-

ment proches de la moyenne de la population (entre 85 et 115) et de 32 % d'observations se situant plus loin (moins que 85 et plus de 115). Mais puisque la sélection des échantillons est aléatoire, nous ne pouvons pas garantir que ces proportions se maintiendront rigoureusement pour tous les échantillons. Certains échantillons contiendront une proportion plus grande d'observations plus éloignées ou plus proches de la moyenne que d'autres échantillons. Cette variation naturelle dans la composition exacte des observations contenues dans les échantillons extraits de la même population s'appelle l'*erreur d'échantillonnage*.

Cette variation aléatoire dans la composition exacte des observations dans les échantillons occasionnée par l'erreur d'échantillonnage cause, à son tour, une différence dans la moyenne des échantillons : un échantillon qui contient plus d'observations dont les valeurs sont grandes aura une moyenne plus forte qu'un échantillon qui contient davantage d'observations dont les valeurs sont petites. Donc, l'erreur d'échantillonnage se répercute dans la moyenne des échantillons. L'*erreur type de la moyenne* est la statistique qui estime la taille de la fluctuation dans les moyennes des échantillons causée par l'erreur d'échantillonnage. Cette statistique est d'une importance primordiale pour distinguer l'hypothèse (H) de l'hypothèse nulle (H_0).

Un objectif des statistiques consiste à réaliser une inférence à la population à partir de l'échantillon. En particulier, la moyenne de l'échantillon (M) est utilisée pour inférer la moyenne de la population (μ). Même s'il est vrai que la moyenne de l'échantillon (M) est la meilleure estimation de μ, il est néanmoins possible que la moyenne de la population se situe loin de la moyenne de l'échantillon. Si nous connaissons la différence typique entre la moyenne d'un échantillon et celle de la population — l'erreur type de la moyenne —, nous pourrons alors déterminer si la moyenne obtenue dans notre échantillon est typique ou atypique, si elle est très ou peu différente de la moyenne de la population. Par exemple, si nous savons que typiquement les moyennes de 68 % des échantillons extraits d'une population ayant une moyenne de 100 se situent entre 85 et 115 (son écart-type étant 15) et que nous trouvons que notre échantillon a une moyenne de 130, nous concluons alors que cette moyenne est fort différente de la moyenne de la population (elle se situe à deux écarts types de la moyenne de la popula-

tion : [Z = (130-100)/15 = +2]. Cet échantillon n'est pas typique pour cette population et nous concluons qu'il appartient à une population différente. Bien sûr, tout cela présume que nous connaissons l'écart-type des moyennes – l'erreur type de la moyenne – des échantillons extraits de la population, une information rarement disponible directement.

Reste que l'erreur type de la moyenne est importante lorsqu'il s'agit d'évaluer l'hypothèse (H) et l'hypothèse nulle (H_0). Nous avons vu (chapitre 8) que l'hypothèse nulle est rejetée lorsque les moyennes des échantillons ne sont pas les mêmes. Lorsque deux échantillons n'ont pas la même moyenne, nous pouvons potentiellement conclure que ces deux échantillons proviennent de populations différentes. Il est également possible que les deux échantillons proviennent de la même population, mais que la différence entre leurs moyennes soit simplement attribuable à l'erreur d'échantillonnage. Dans ce dernier cas, le rejet de H_0 serait une erreur. Il faut donc trouver un mécanisme pour distinguer une différence attribuable à l'erreur d'échantillonnage d'une autre qui, elle, est attribuable à une différence de populations. Le mécanisme statistique qui permet de faire cette distinction exige la quantification de la taille de l'erreur type de la moyenne.

Une fois cette quantité déterminée, il est possible d'estimer la proximité des moyennes de deux échantillons ou la proximité de la moyenne de l'échantillon et de celle de la population. Par exemple, supposons que la différence typique (l'erreur type de la moyenne) entre la moyenne de deux échantillons est de 10 et que la moyenne de la population est de 100. On tire un échantillon ayant 90 comme moyenne. Cet échantillon est-il près ou loin de la moyenne de la population ? La différence entre les deux moyennes est de 10 (90 – 100 = 10), mais comment interpréter cette différence ? Une solution est de la standardiser en valeur étalon. Puisque nous connaissons la différence typique entre les moyennes des échantillons, c'est-à-dire leur écart-type, le calcul donne : $Z_M = (90 - 100)/10 = -1$ (l'erreur type, puisqu'elle n'est que l'écart-type des moyennes des échantillons, est donc égale à 10). Dans ce cas, nous observons que la moyenne de notre échantillon se trouve à une erreur type en dessous de celle de la population. Est-ce loin ou près de la moyenne ? Nous verrons. Mais on voit que la différence typique, l'écart-type entre les moyennes des échantillons dû à l'erreur d'échantillonnage, représente une statistique fort importante qui nous per-

met de faire l'interprétation d'une différence. Le défi consiste à déterminer sa valeur numérique.

Quiz rapide 9.1
Supposons qu'au Canada le salaire moyen des employés est de 50 000$. Nous tirons un échantillon de travailleurs canadiens qui détiennent tous un Ph.D. En moyenne, ces Ph.D. gagnent 90 000$. Pouvons-nous alors conclure que les Canadiens ayant un Ph.D. appartiennent à une population de salaire différente ?

QUANTIFIER L'ERREUR D'ÉCHANTILLONNAGE

L'erreur type de la moyenne est l'écart-type des moyennes des échantillons aléatoirement extraits de la même population. Cette statistique n'est pas la même pour tous les échantillons et toutes les populations. Elle peut être plus ou moins grande et sa taille dépend de deux facteurs : le nombre d'observations dans l'échantillon (N) et la variance de la population (σ^2).

1. N, le nombre d'observations dans les échantillons : *plus la taille de l'échantillon est grande, plus l'erreur d'échantillonnage est petite.* Ce principe est appelé *la loi des grands nombres*.

Imaginons un échantillon qui inclut tous les membres d'une population d'un million sauf un. Puisque presque tous sont présents dans l'échantillon, la moyenne de l'échantillon sera à un millionième près la moyenne de la population. Un deuxième échantillon de même taille tiré de cette population sera obligatoirement composé d'observations quasi identiques. Dans ce cas, il n'y aura virtuellement aucune erreur d'échantillonnage et la différence entre les moyennes de ces deux échantillons sera donc très proche de zéro. En revanche, si nous tirons un échantillon composé d'une seule observation, cette observation pourrait provenir de n'importe quelle partie de la population. La même chose serait vraie pour un deuxième échantillon extrait de cette population, composé lui aussi d'une seule observation. La différence entre les moyennes de ces deux échantillons sera grande, ce qui se traduira par une grande erreur d'échantillonnage et une plus grande différence typique entre les moyennes. Par exemple, la note obtenue à un examen par un seul étudiant est une piètre estimation des notes de toute la classe alors que la note moyenne obtenue par 99 % des étudiants sera très proche de la note moyenne obtenue par tous les étudiants. Ainsi,

nous comprenons que plus un échantillon contient d'observations, plus sa moyenne sera semblable à la moyenne de la population. L'erreur type de la moyenne sera petite.

2. La taille de l'erreur type de la moyenne est aussi fortement influencée par la variance (ou l'écart-type) de la population : *plus grande est la variance de la population, plus grandes sont l'erreur d'échantillonnage et l'erreur type de la moyenne.*

L'erreur d'échantillonnage est plus grande lorsque les observations dans la population diffèrent davantage les unes des autres. Lorsque les observations sont proches les unes des autres dans la population (la variance de la population, σ^2, est faible), les échantillons seront nécessairement composés d'observations qui sont plus similaires, plus proches les unes des autres et la variance des observations (s^2) sera plus faible. Lorsque les observations contenues dans les différents échantillons sont similaires, les échantillons auront des moyennes similaires et, dans ce cas, l'erreur type de la moyenne sera plus petite.

Pour illustrer le principe, prenons un cas de résultats à un examen où tous les étudiants obtiennent des notes entre 70 et 75. La variance de la population est donc faible et, par conséquent, tous les échantillons d'étudiants auront des moyennes plutôt similaires (elles seront toutes obligatoirement entre 70 et 75). Par conséquent, l'erreur-type de la moyenne sera faible. Mais supposons, à l'inverse, que les notes varient entre 0 et 100. Chaque échantillon risque fort d'être composé d'observations plus différentes, ce qui fera en sorte que la moyenne d'un échantillon sera différente de celle d'un autre échantillon. L'erreur type de la moyenne sera donc plus grande.

Ainsi, l'erreur d'échantillonnage augmente en fonction de la variance de la population (σ^2), mais elle se réduit en fonction de la taille de l'échantillon (N). Lorsqu'une population est très homogène, tous les échantillons extraits de cette population auront des moyennes proches les unes des autres. Similairement, les grands échantillons extraits d'une population détiendront tous des moyennes similaires. Dans les deux cas, l'erreur type de la moyenne sera petite.

Quiz rapide 9.2
Nous étudions l'attitude envers les hôpitaux de deux populations : les médecins et les citoyens. Quelle population aura probablement une variance plus grande dans les attitudes ?

L'expérience d'échantillonnage et l'erreur type de la moyenne

L'objectif d'un test statistique consiste à permettre une inférence à la moyenne de la population μ à partir de la moyenne de l'échantillon M. Mais nous savons que chaque échantillon (sauf si la taille des échantillons est infiniment grande ou la variance dans la population est nulle) produit une moyenne différente des autres. Par conséquent, presque toutes les moyennes des échantillons extraits de la même population seront au moins un peu différentes les unes des autres et différentes de la véritable moyenne de la population. C'est-à-dire que chaque échantillon produit une moyenne qui estime la moyenne de la population en faisant, au mieux, une légère erreur. Cette erreur, *l'erreur type de la moyenne*, doit être calculée si nous voulons interpréter une différence entre deux moyennes.

Nous pouvons comprendre et calculer cette erreur type en faisant une expérience particulière qui se nomme l'*expérience d'échantillonnage*. Supposons qu'à partir d'une population dont nous connaissons la vraie moyenne (μ), nous tirons tous les échantillons différents possibles (disons qu'il en existe K), chacun composé du même nombre d'observations *N*. Nous calculons, pour chaque échantillon, sa moyenne (M_j) et la différence entre la moyenne de chaque échantillon et la moyenne de la population μ (M-μ). Puisque tous les échantillons sont extraits de la même population, chaque différence entre la moyenne de l'échantillon et la moyenne de la population est en réalité une indication de l'erreur que la moyenne de chaque échantillon fait dans son estimation de la moyenne de la population. En ayant ces informations, il nous est alors possible de calculer l'erreur moyenne que nous pouvons aussi placer sous la rubrique de « l'erreur typique ». La Formule 9.1 formalise cette quantité. On remarquera, dans cette formule, que nous avons mis au carré les quantités M-μ afin d'empêcher que cette sommation donne zéro. Comme on l'aura peut-être deviné, cette formule n'est rien d'autre que celle utilisée pour calculer la variance

d'une population (dans ce cas, la variance de la moyenne des échantillons extraite de la même population d'observations).

$$\sigma_M^2 = \sum_{j=1}^{K} (M_i - \mu)^2 / K \qquad \text{Formule 9.1}$$

où σ_M^2 est la variance des moyennes des échantillons, μ est la moyenne de la population, M_i est la moyenne de chaque échantillon, et K est le nombre total d'échantillons.

Quiz rapide 9.3
Pourquoi la Formule 9.1 se sert-elle de K plutôt que de K – 1 comme dénominateur?

Si nous prenons la racine carrée de la quantité σ_M^2 (la variance des erreurs), nous obtenons son écart-type (l'écart-type des erreurs: σ_M). L'écart-type de ces erreurs est l'erreur typique que nous faisons en estimant μ à partir de M, la moyenne des échantillons, c'est-à-dire l'erreur type de la moyenne. Donc, à partir de l'expérience d'échantillonnage, nous obtenons l'information requise pour interpréter une différence entre la moyenne d'un échantillon et la moyenne de la population.

L'estimation de l'erreur type de la moyenne des échantillons

En pratique, nous ne pouvons jamais sélectionner tous les échantillons possibles d'une population et, en général, nous n'avons à notre disposition qu'un seul échantillon. Néanmoins, il faut connaître l'erreur type de la moyenne si nous voulons interpréter la moyenne d'un échantillon.

Heureusement, il est possible de faire *une estimation* de l'erreur type des moyennes attribuable à l'erreur d'échantillonnage à partir d'un seul échantillon. La Formule 9.2a fait cette estimation.

$$\sigma_M^2 = \frac{\sigma^2}{N} \qquad \text{Formule 9.2a}$$

La Formule 9.2a découle de la Formule 9.1. On en trouve la preuve mathématique (Comment trouver l'erreur type de la moyenne) à la fin de ce chapitre. Pour l'instant, examinons pourquoi la Formule 9.2a est appropriée.

On se souvient que la variance des moyennes des échantillons (σ^2_M) est plus grande lorsque la variance de la population (σ^2) est grande, mais que cette erreur est plus petite lorsque l'échantillon contient plus d'informations (N). La Formule 9.2a met en rapport ces deux influences sur la taille de l'erreur type de la moyenne pour produire la variance des erreurs. Plus la variance de la population (σ^2) est grande, plus grand est le numérateur de la Formule 9.2a, et plus grande est la quantité σ^2_M. Plus grande est la taille de l'échantillon (N), plus grand est le dénominateur et, par conséquent, plus petite est la quantité σ^2_M.

Comme précédemment, la Formule 9.2a produit l'*erreur type au carré,* ce qui n'est pas très commode. En calculant la racine carrée de l'erreur type au carré, nous obtenons la formulation de l'erreur type de la moyenne σ_M, calculée à partir d'un unique échantillon (Formule 9.2b).

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \qquad \text{Formule 9.2b}$$

Cet écart-type représente le degré avec lequel les moyennes (M) des échantillons fluctuent autour de la vraie moyenne (μ). C'est pour distinguer l'écart-type des moyennes de l'écart-type des observations à l'intérieur d'un échantillon que nous lui donnons son nom particulier : l'erreur type de la moyenne. Ainsi, l'erreur type de la moyenne est l'erreur typique qui existe entre la moyenne d'un échantillon et la moyenne de la population.

Si nous connaissons l'erreur type de la moyenne, il est facile de déterminer si la moyenne d'un échantillon particulier est près ou loin de la moyenne de la population. Un échantillon décrit (représente) fort bien la population lorsque sa moyenne est proche (située à moins d'une erreur type) de la moyenne de la population. Inversement, plus la moyenne de l'échantillon s'éloigne de la moyenne de la population, moins cet échantillon est capable de bien représenter la population.

L'estimation de l'erreur type de la moyenne en pratique

Jusqu'à présent, nous avons défini l'erreur type de la moyenne comme étant la variabilité des moyennes des échantillons, ce que nous pouvons calculer à condition de connaître l'écart-type de la population. Mais, en pra-

tique, nous ne connaissons (presque) jamais l'écart-type de la population (σ). Par conséquent, la formule de l'erreur type de la moyenne ($\sigma/\sqrt{N}$) ne peut (presque) jamais être calculée.

Mais nous connaissons l'écart-type de l'échantillon et nous savons (chapitre 8) que la meilleure estimation de l'écart-type de la population est l'écart-type de l'échantillon. Nous pouvons alors, en pratique, substituer l'écart-type de l'échantillon (s) à l'écart-type de la population (σ). La Formule 9.3 estime l'erreur type de la moyenne en pratique

$$s_M = s/\sqrt{N} \qquad \text{Formule 9.3}$$

où s est l'écart-type de l'échantillon, N est le nombre d'observations dans l'échantillon, et s_M est l'estimation de σ_M lorsque l'écart-type de la population n'est pas connu. Par exemple, si l'écart-type d'un échantillon de N = 100 observations est s = 10, l'erreur type estimée devient $s_M = 10/\sqrt{100} = 10/10 = 1$. Si la moyenne de l'échantillon est de 5, nous concluons que, typiquement, les échantillons extraits aléatoirement de cette population auront une moyenne se situant entre 4 et 6 (5±1). De la même manière, nous pouvons dire que la moyenne de la population se situe entre 4 et 6.

L'utilisation de l'erreur type de la moyenne : une illustration

Nous étudions le QI depuis presque un siècle et des millions de personnes ont passé ce test. Par conséquent, nous connaissons fort bien sa variance et sa moyenne dans la population. Le QI moyen est de 100 et son écart-type est de 16. Supposons que nous prenons un échantillon d'étudiants et que nous observons que le Q1 moyen dans cet échantillon est de 120. Est-ce que les étudiants de cet échantillon sont très différents de la population ? Supposons que l'erreur type de la moyenne est égale à 10. Nous pouvons alors calculer la position de la moyenne de notre échantillon par rapport à la moyenne de la population en transformant cette moyenne en valeur étalon Z.

La formule générale pour la valeur étalon d'un score *X* est $Z_x = (X - M)/s$, où s est l'écart-type des moyennes des échantillons, c'est-à-dire l'erreur type de la moyenne. Puisque nous voulons calculer la valeur Z pour la moyenne (M) d'un échantillon par rapport à la moyenne de la population (μ), nous utilisons la Formule 9.4

$$Z_M = (M - \mu)/\sigma_m \qquad \text{Formule 9.4}$$

où Z_M est la position de la moyenne de l'échantillon par rapport à la moyenne de la population, M est la moyenne obtenue dans l'échantillon, μ est la moyenne de la population et σ_m est l'erreur type de la moyenne.

Calculons ces valeurs pour notre échantillon d'étudiants: μ = 100, M = 120, et σ_m = 10. En appliquant la Formule 9.3, nous obtenons:

$$\begin{aligned} Z_M &= (M - \mu)/\sigma_m \\ &= 120 - 100\ /10 \\ &= 20/10 \\ &= +2 \end{aligned}$$

Nous savons maintenant que la moyenne de cet échantillon est à deux erreurs types au-dessus de la moyenne des QI dans la population. Comme nous le verrons plus tard, cet échantillon produit une moyenne que nous allons éventuellement qualifier de « statistiquement différente » de celle de la moyenne de la population.

Quiz rapide 9.4
Reprenez l'exemple précédent portant sur le QI. Supposons que le QI moyen d'un échantillon est de 140. Présumez que μ = 100 et que σ_m = 10. Quelle est la distance, en valeur étalon Z, entre la moyenne de cet échantillon et la moyenne de la population ? En vous référant au tableau de la courbe normale, quelle est la proportion des échantillons extraits de cette population qui auront une moyenne plus grande que 140 ?

Quiz rapide 9.5
Supposons maintenant que votre échantillon est composé de 100 personnes et que la variance de cet échantillon est de 100. Le QI moyen de cet échantillon est de 110. Pouvez-vous déduire la moyenne de la population de QI, à partir de ces informations ? Supposons maintenant la même moyenne (M = 110) et la même variance (100) mais un échantillon de 25 personnes seulement, quelle serait alors votre estimation de la moyenne de la population ? Ces deux estimations de la moyenne de la population sont-elles différentes ou non ? Pourquoi ?

Le théorème de la limite centrale

« L'expérience d'échantillonnage » décrite ci-dessus consiste à extraire tous les échantillons possibles d'une même taille d'une unique population d'ob-

servations. En calculant la moyenne de chaque échantillon, nous pouvons établir la distribution de ces moyennes et la distribution des différences entre chacune des moyennes et la moyenne de la population. Le *théorème de la limite centrale*[1] énonce une série de propositions qui sont vraies au sujet de la distribution de ces moyennes. Parmi ces propositions, trois sont particulièrement importantes et utiles.

- La moyenne de la distribution des moyennes des échantillons est égale à μ, la moyenne de la population.
- La variation entre les moyennes des échantillons sera plus petite que la variation entre les individus de la population. En fait, l'écart-type de cette distribution de moyennes est approximativement égal à l'erreur type de la moyenne ($\sigma/\sqrt{N}$).
- La forme de la distribution des moyennes s'approche de la distribution normale lorsque la taille des échantillons est grande (environ $N \geq 30$). *Cela demeure vrai même lorsque la distribution de la population n'est pas normale.* Si la distribution de la population est normale, la distribution de la moyenne des échantillons est normale même lorsque les échantillons extraits de cette population sont petits ($N<30$).

Les implications du théorème de la limite centrale pour l'inférence

Cette dernière proposition est particulièrement utile. Puisque la distribution des moyennes suit une distribution normale ayant un écart-type (c.a.d. une erreur type) connu ($s/\sqrt{N}$), nous sommes en mesure de faire un grand nombre d'inférences nous permettant, éventuellement, de choisir entre H et H_0.

Nous connaissons, à partir du tableau de la densité sous la courbe normale, la proportion des observations qui se situent à différentes distances de la moyenne. Nous pouvons appliquer cette connaissance à la distribution des moyennes des échantillons puisqu'elle est normale, ce qui est quasi toujours le cas.

1. Un théorème est une proposition qui est prouvée. Ce théorème a été conjecturé par Gauss lui-même en 1812, mais la preuve formelle n'a été découverte qu'en 1932 par Alan Turing, le fondateur de l'informatique.

La Figure 9.1 indique la proportion des échantillons dont la moyenne se trouve plus ou moins loin de la vraie moyenne de la population. On remarque dans cette figure que nous donnons les proportions pour Z = ±1,96 et Z = ±2,58 plutôt que pour Z = 2 ou Z = 3. Nous choisissons ces valeurs parce qu'elles seront utiles lorsqu'il sera question du concept de la signification statistique. Il vaut mieux s'y habituer, car elles vont revenir souvent à partir de maintenant !

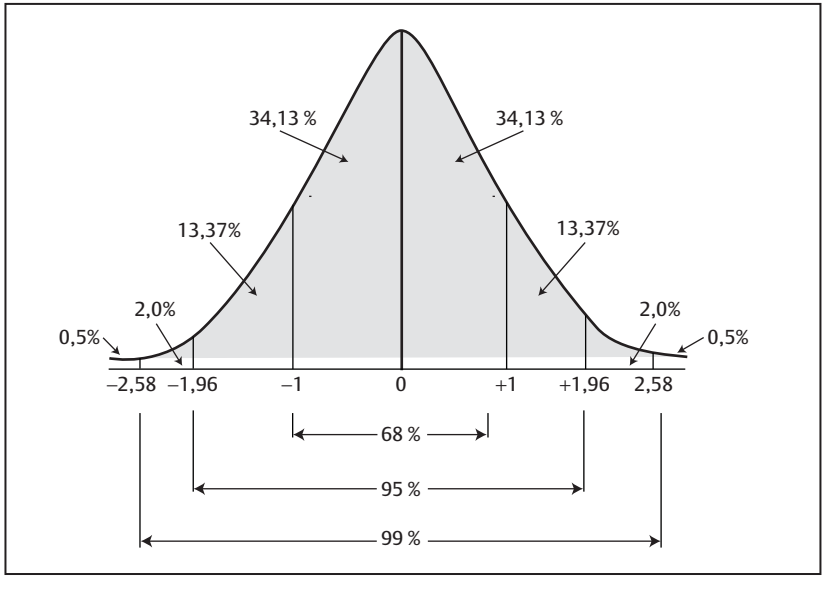
Puisque la distribution des moyennes est normale, en se servant du tableau de la distribution normale (chapitre 5), on connaît maintenant la proportion des échantillons qui ont des moyennes de différentes magnitudes.

À la Figure 9.1, nous voyons qu'environ 68 % des échantillons ont des moyennes situées à ±1 erreur type de la moyenne de la population. En utilisant des termes probabilistes, nous pouvons dire que la probabilité qu'un échantillon obtienne une moyenne le situant entre ±1 erreur type de la vraie moyenne μ est p = 0,68. La Figure 9.1 montre aussi qu'environ 95 % des échantillons ont une moyenne située entre ±1,96 erreur type de μ. Ainsi, la probabilité qu'un échantillon obtienne une moyenne se situant entre ±1,96 erreur type de la vraie moyenne μ est p = 0,95.

Remarquons que 95 % étant égal à 19 sur 20, on peut aussi dire que la moyenne des échantillons est la moyenne de la population avec une marge de 1,96 erreur type valable 19 fois sur 20. C'est exactement de cette façon que sont rapportés les sondages d'opinion dans les journaux. Enfin, puisque 99 % des échantillons ont des moyennes situées à ±2,58 erreurs types de la moyenne, la probabilité que la moyenne d'un échantillon se trouve entre ±2,58 erreurs types de la moyenne est p = 0,99.

Nous pouvons exprimer la même chose différemment. La probabilité que la moyenne de n'importe quel échantillon se situe *à ou au-delà* de ±1 erreur type de la moyenne (μ) est approximativement p = 0,32 (1,00 – 0,68 = 0,32) ; la probabilité que l'échantillon ait une moyenne à ou au-delà de ±1,96 erreur type est p = 0,05 (1 – 0,95 = 0,05) ; la probabilité qu'un échantillon ait une moyenne à ou au-delà de ±2,58 erreurs types est p = 0,01 (1 – 0,99 = 0,01). Obtenir un échantillon ayant une moyenne loin de la moyenne de la population est un événement beaucoup plus rare qu'obtenir un échantillon dont la moyenne est proche de μ.

FIGURE 9.1 Distribution de la moyenne des échantillons de même taille extraits d'une population quelconque



Reprenons l'exemple du groupe d'étudiants ayant un QI moyen égal à 120 extraits d'une population ayant un QI moyen de 100 et une erreur type de la moyenne de 10. Nous avons déterminé la position de la moyenne de notre échantillon par rapport à la moyenne de la population d'étudiants (QI = 100) comme étant $Z_M = (M - \mu)/\sigma_m = (120 - 100)/10 = 20/10 = +2$. La moyenne de cet échantillon se situe à +2 erreurs types de la moyenne de la population.

En se référant au tableau de la densité de la courbe normale (dans l'Annexe), on sait que 97,72 % (p = 0,9772 pour Z = +2) des échantillons devraient avoir une moyenne égale ou inférieure à un QI = 120. Autrement dit, seulement 2,28 % des échantillons d'étudiants devraient avoir un QI moyen supérieur à 120 (100 % – 97,72 % = 2,28 %) si notre échantillon provenait effectivement de la population générale du QI. Donc, la probabilité p que cet échantillon soit représentatif (provienne) de la population générale de QI est p = 0,0228. Or, il est exceptionnel (p = 0,0228) de tirer aléatoirement un échantillon d'étudiants dont le QI moyen est égal ou supérieur à 120 s'il provient d'une population dont le QI moyen est 100.

Puisque seulement 2,28 % des échantillons d'étudiants (tirés d'une population ayant un QI moyen de 100) peuvent avoir un QI supérieur à 120, nous concluons que cet échantillon ne provient pas (ou, plus exactement, a une faible probabilité de provenir) de la population générale. *Plus formellement, nous dirons que cet échantillon provient probablement d'une population d'intelligence différente de celle de la population générale.* En effet, cet échantillon est composé d'étudiants en moyenne très intelligents (différents du QI moyen de la population). Lorsque la moyenne d'un échantillon est suffisamment distante de la moyenne de la population, nous disons que cet échantillon provient d'une population différente. Il nous reste à définir ce que nous entendons par « suffisamment distant » – ce que les statisticiens appellent « la signification statistique ».

En formulant cette conclusion, nous acceptons un risque d'erreur, qui, dans ce cas, est de 0,0228, car 2,28 % des échantillons pourraient effectivement provenir de cette population. Cette logique est essentielle pour la compréhension du test de la *signification statistique* et du concept de *l'erreur d'inférence* qui sont expliqués plus loin.

LA SIGNIFICATION STATISTIQUE

La *signification statistique* est le critère utilisé pour conclure au sujet de H et de H_0 (rejet ou non de H_0 : voir le chapitre 8).

Le fait que la distribution de la moyenne des échantillons suit une distribution normale est extrêmement utile, car nous pouvons maintenant faire appel à ses caractéristiques (voir le chapitre 5 et la Figure 9.1) pour évaluer la proximité entre la moyenne de n'importe quel échantillon et la moyenne de la population. Ainsi, nous savons que 50 % des échantillons ont une moyenne égale ou inférieure à la moyenne de la population ; que 34,13 % des échantillons se trouvent entre la moyenne et 1 erreur type de la moyenne de la population ; qu'approximativement 68,26 % des échantillons ont une moyenne à ±1 erreur type de la moyenne de la population ; qu'environ 13,37 % des échantillons ont une moyenne entre +1 et +1,96 erreur type de la moyenne de la population et que 13,37 % des échantillons ont une moyenne entre –1 et –1,96 erreur type de la moyenne de la population. Enfin, seulement 2,5 % des échantillons ont une moyenne égale ou

plus grande et seulement 2,5 % des échantillons ont une moyenne égale ou plus petite que 1,96 erreur type de la moyenne de la population.

Supposons que nous savons que la moyenne de la population est égale à 4 et que l'écart-type des moyennes des échantillons (l'erreur type de la moyenne) est $\sigma_m = 1$. Nous savons alors, en nous fiant à la Figure 9.1, que 34,13 % des échantillons auront une moyenne se situant entre la moyenne de la population et 1 erreur type (entre 4 et 5) et que 34,13 % auront une moyenne se situant entre la moyenne de la population et -1 écart-type (entre 3 et 4). Au total, alors nous savons que 68,26 % (34,13 % +34,13 %= 68,26 %) des échantillons tirés de cette distribution ont une moyenne entre 3 et 5 (4 ± 1 erreur type de la moyenne). Nous savons aussi, toujours en nous référant au tableau de la densité des observations de la distribution normale, que 13,37 % ont une moyenne les situant entre +1 et +1,96 erreur type de la moyenne de la population (entre 5 et 5,96) et que 13,37 % ont une moyenne les situant entre –1 et –1,96 erreur type (entre 2,04 et 3). Ainsi, 95 % des échantillons ont une moyenne située entre ±1,96 erreur type (68,26 % + 13,37 % + 13,37 % = 95 %) de la moyenne de la population. Dans notre exemple, cela voudrait dire entre 2,04 et 5,96 lorsque la moyenne de la population est $\mu = 4$ et que l'erreur type est de 1,0.

Faisons le calcul inverse maintenant. Si 68 % des échantillons ont une moyenne égale ou inférieure à ±1 erreur type de μ, 32 % des échantillons ont une moyenne qui est plus distante. De la même façon, si 95 % des échantillons ont une moyenne qui est située à ±1,96 erreur type de μ, 5 % des échantillons ont une moyenne encore plus distante de la moyenne de la population. Enfin, lorsque 99 % des échantillons produisent une moyenne entre ±2,58 erreurs types, 1 % des échantillons ont une moyenne encore plus éloignée de μ.

Supposons maintenant qu'on ignore quelle est la moyenne de la population, mais qu'un quidam affirme que cette moyenne est de 4. Pour vérifier ses dires, on prend un échantillon et on trouve une moyenne de 5,96. Nous connaissons l'écart-type et le N associés à son échantillon et, à partir de ces données, nous calculons l'erreur type de la moyenne ($s_M = s/\sqrt{N}$) et nous trouvons qu'une erreur type est $s_M = 1$. Le quidam a-t-il raison ? S'il a raison, un échantillon comme celui qui a été obtenu se serait produit avec une faible probabilité p = 0,05 [(5,96-4)/1 = +1,96], indiquant que 95 %

des échantillons extraits de cette population auraient une moyenne égale ou inférieure à 5,96 et donc, que moins de 5 % des échantillons auraient une moyenne supérieure à celle obtenue dans notre échantillon. Nous concluons qu'il y a moins de 5 % des chances qu'un échantillon ayant une moyenne de 5,96 puisse provenir d'une population dont la moyenne est 4. Mais pourtant, c'est ce que nous avons effectivement obtenu. Le quidam se tromperait-il en affirmant que $\mu = 4$?

Quiz rapide 9.6
Imaginez que l'échantillon obtenu a une moyenne de 6,58. Quelle est la probabilité p d'obtenir un tel échantillon si le quidam a raison ?

Quiz rapide 9.7
Vous secouez la tête et découvrez que le quidam n'existe pas, sauf dans votre imagination. Pour sauver la face, pourriez-vous remplacer « quidam » par « hypothèse » ? Si oui, représente-t-il l'hypothèse ou l'hypothèse nulle ?

Par convention, lorsqu'un échantillon donne une moyenne qui se situe à plus de ±1,96 erreur type de la moyenne μ — c'est-à-dire que la probabilité qu'il puisse appartenir à cette population est plus petite que p = 0,05 —, nous disons que cet échantillon ne provient pas de la population : il est statistiquement différent de la moyenne de la population. Il est statistiquement significatif. Nous notons cela en écrivant $p < ,05$, ce qui indique qu'il y a moins de 5 % des chances que cet échantillon puisse effectivement provenir de cette population.

La signification statistique réfère donc à la probabilité que l'échantillon provienne de la population. Lorsque cette probabilité est plus petite que 0,05, nous disons (par convention) qu'il s'agit d'un échantillon qui ne peut être obtenu que rarement (5 fois sur 100). Puisqu'il s'agit d'une situation rare, nous concluons que cet échantillon n'appartient (probablement) pas à cette population.

Le concept de la signification statistique : une analogie

Supposons qu'une amie prétende avoir un pouvoir magique lui permettant de deviner si une pièce de monnaie lancée en l'air tombera du côté pile ou face. Pour la tester, vous lui demandez de lancer la pièce cinq fois et de deviner le résultat. Vous notez, à chaque fois, si elle a deviné le résultat correctement ou non. Au premier lancer, elle devine correctement. Cela ne prouve rien, car ses chances d'avoir raison sont de 1 sur 2 ($1/2 = p = 0{,}50$). Elle devine le deuxième lancer correctement aussi. La probabilité de deviner le deuxième lancer correctement est de 1/2. Mais la probabilité de deviner correctement deux lancers d'affilée est de $1/2 \times 1/2 = 1/2^2 = 1/4$ ($p = 0{,}25$). Elle devine le troisième lancer correctement. Si elle n'avait pas de pouvoir magique, la probabilité de deviner correctement trois lancers d'affilée serait de $1/2 \times 1/2 \times 1/2 = 1/2^3 = 1/8$ ($p = 0{,}125$). Elle devine les deux derniers lancers correctement aussi. La probabilité de deviner correctement tous les lancers est de $1/2^5 = 1/32$ ($p = 0{,}03$). Les chances qu'une personne sans pouvoir divinatoire puisse deviner correctement cinq lancers d'une pièce de monnaie sont $p = 0{,}03$, ce qui est une probabilité inférieure à $p = 0{,}05$, le seuil minimalement requis pour conclure à la signification statistique. Puisque cette amie est capable de le faire, nous concluons qu'elle n'appartient pas à la population habituelle. Serait-elle membre d'une population distincte, composée de personnes ayant des pouvoirs magiques ? Dans ce cas, la logique statistique permet une réponse affirmative.

Le choix de $p = 0{,}05$ pour définir la signification statistique est arbitraire. Il n'existe aucun motif rationnel pour choisir cette valeur. Il s'agit d'une convention sur laquelle nous nous accordons pour nous aider à prendre une décision : rejet ou non-rejet de H_0. Si nous prenons un échantillon d'une population et trouvons qu'il a une moyenne le situant à plus de 1,96 erreur type de la moyenne de la population attendue, nous concluons que cet échantillon est significativement différent de la population, qu'il ne provient pas de cette population : il provient d'une autre population. Mais sommes-nous certains de notre conclusion ? Cette dernière population existe-t-elle dans la réalité ?

Le risque d'erreur d'inférence et le seuil de signification α (alpha)

Lorsque nous obtenons pour notre échantillon une moyenne qui le situe à +1,96 (ou –1,96) erreur type de la moyenne de la population, nous faisons, par convention, l'inférence que cet échantillon provient d'une autre population. Mais nous en sommes venus à cette conclusion à partir de probabilités. Or, les probabilités ne sont jamais des certitudes. Après tout, toutes les populations contiennent des échantillons dont la moyenne se situe à

1,96 ou plus erreur type ou plus de la moyenne de la population, tout en y faisant partie. Bien sûr, de tels échantillons sont rares mais, néanmoins, ils sont possibles. Lorsque nous concluons à la signification statistique, nous courrons invariablement un certain risque d'avoir émis une fausse conclusion. Il nous faut donc trouver une façon de quantifier ce risque d'erreur.

Nous savons, en nous basant sur le théorème de la limite centrale, que la distribution des moyennes des échantillons est normale. Pour toutes les distributions normales, 5 % des échantillons auront une moyenne se situant à ±1,96 ou plus erreur type de la moyenne de la population.

Si nous obtenons un échantillon dont la moyenne est située à 1,96 erreur type ou plus, nous concluons, par convention, qu'il ne fait pas partie de la population (la moyenne de cet échantillon est statistiquement différente de celle de la population) et que cet échantillon appartient à une population différente. Malgré cette conclusion, il faut reconnaître que 5 % des échantillons de toutes les populations se situent véritablement à au moins 1,96 erreurs types de la moyenne de leur population. Par conséquent, nous avons une probabilité de $p = 0{,}05$ de faire une erreur en concluant que l'échantillon n'appartient pas à cette population (cet échantillon a 5 chances sur 100 d'appartenir réellement à cette population). Puisque nous avons conclu que l'échantillon ne provient pas de la population, nous courons alors un risque de 5 % de commettre une erreur. Nous donnons un nom particulier à ce risque d'erreur. Nous l'appelons le *seuil de signification alpha, l'erreur de type I,* ou encore *l'erreur alpha* (α).

Le seuil alpha indique le risque d'une erreur d'inférence associé à la conclusion que l'échantillon ne provient pas de la population. Supposons que l'on décide de juger statistiquement et significativement différent un échantillon dont la moyenne le situe à ±1 erreur type de μ. Puisque nous savons que 32 % des échantillons d'une population peuvent avoir une moyenne plus éloignée de la moyenne μ que ±1 erreur type, nous courons alors un risque d'erreur de $p = 0{,}32$ et le risque d'erreur de type 1 est $\alpha = 0{,}32$. Nous avons presque une chance sur trois de nous tromper avec ce seuil ! Cela étant un risque d'erreur plus grand que celui généralement accepté (0,05), nous concluons qu'il n'y a pas de preuve que notre échantillon appartienne à une autre population. Le rejet de H_0 n'est pas approprié et nous concluons à *la non-signification statistique* : rien indique que l'échantillon en question n'appartient pas à la population.

Quiz rapide 9.8
Supposons que nous voulons un risque d'erreur α très faible (disons $\alpha = ,01$). Supposons aussi que la moyenne hypothétique de la population est 100 avec une erreur type de la moyenne de 10. Si un échantillon obtient une moyenne de 140, allons-nous conclure qu'il ne provient pas de la population ? À partir de combien allons-nous commencer à conclure qu'il ne provient pas de cette population ?

Nous pouvons réduire le risque d'une erreur d'inférence alpha en choisissant un seuil de signification plus petit. Ainsi, nous pourrions décider que la différence entre la moyenne de l'échantillon et la moyenne de la population est statistiquement significative seulement si la moyenne de l'échantillon se situe à plus de 2,58 erreurs types de la moyenne μ. Nous savons que seulement 1 % des échantillons de cette population peuvent obtenir une moyenne à une telle distance de la moyenne de la population. Dans ce cas, si nous concluons que l'échantillon n'appartient pas à cette population, le risque d'une erreur alpha devient $\alpha = 0,01$.

Même si, par convention, nous concluons à la signification statistique lorsque le risque d'erreur alpha est plus petit que 0,05, dans certaines situations, il est permis de choisir des seuils alpha plus grands (par exemple $\alpha = 0,10$), lorsqu'il s'agit d'expériences pilotes, d'études exploratoires ou lorsque le risque de faire une erreur α est sans conséquence. Inversement, si une erreur d'inférence peut entraîner de graves conséquences, tels des dangers pour la santé, il est préférable de choisir $\alpha = 0,01$ ou même $\alpha = 0,001$ (1 chance sur 1000 de faire une erreur de type I) comme seuil de signification statistique.

Par exemple, supposons qu'un individu assure son ami que porter une pyramide sur la tête une heure par jour permet d'augmenter notablement le QI. Comme ce dernier ne le croit pas du tout et qu'il ne veut investir ni temps ni argent, il décide de faire une expérience avec peu de participants et un seuil de décision de 10 %. S'il ne rejette pas l'hypothèse nulle (pas d'effet de la pyramide), ça ne lui coûtera pas trop cher. Par contre, s'il rejette l'hypothèse nulle, il se promet d'aller au fond des choses à l'aide d'une seconde expérience plus élaborée et il testera l'hypothèse avec un seuil alpha plus sévère, tel que 5 ou 1 %.

Maintenant que nous avons en main une manière pratique de calculer l'erreur type de la moyenne, nous pouvons l'utiliser afin d'accroître notre degré de confiance dans les conclusions. Examinons le concept de l'intervalle de confiance ainsi que son calcul.

L'INTERVALLE DE CONFIANCE

La meilleure estimation que nous avons de la moyenne de la population est la moyenne de l'échantillon M. Nous ne devrions pas accepter aveuglément que M = μ. Après tout, un autre échantillon tiré aléatoirement de cette même population produira presque toujours une moyenne au moins un peu différente de celle trouvée dans le premier échantillon. L'*intervalle de confiance* est une statistique qui utilise l'erreur type de la moyenne afin de calculer une fourchette de valeurs dans laquelle la moyenne de la population a le plus de chances de se trouver.

Dans l'exemple du QI, notre échantillon produit une moyenne M = 120, et notre meilleure estimation est donc que μ = 120. Mais puisque l'erreur d'échantillonnage est un fait inévitable, il serait plus prudent de dire : « La meilleure estimation que nous avons est μ = 120, mais sa vraie valeur pourrait être, disons, aussi faible que 80 et aussi forte que 140. » Nous avons donc établi deux valeurs à l'intérieur desquelles la véritable moyenne de la population a beaucoup de chances de se trouver. Avant de discuter des calculs requis, examinons le principe.

Si nous avions simplement affirmé que le QI moyen de la population est μ = 120, nous nous serions trompés (il est en réalité μ = 100). Si nous avions calculé une fourchette de valeurs allant de 80 à 140, nous ne serions plus dans l'erreur puisque la vraie moyenne de la population (100) est incluse entre ces deux valeurs. Nous pouvons toujours nous tromper (la vraie moyenne de la population pourrait être 150), mais le risque d'erreur est minimisé.

Puisqu'il s'agit d'établir une fourchette de valeurs, il faut calculer une valeur inférieure à M et une autre supérieure à M, entre lesquelles la vraie valeur de μ se trouvera. La formule pour le calcul de l'intervalle de confiance (IC) est

$$IC = M \pm Z \times s_M \qquad \text{Formule 9.5}$$

où M est la moyenne de l'échantillon, Z est la taille de la fourchette de valeurs et s_M est l'erreur type de la moyenne. Le symbole ± indique que la quantité $Z \times s_M$ est ajoutée et soustraite de la moyenne obtenue dans l'échantillon afin de produire la limite supérieure et la limite inférieure que pourrait prendre la moyenne de la population.

Avant d'expliquer l'intervalle de confiance plus en détail, étudions le Tableau 9.1. La première ligne du tableau indique la note moyenne obtenue à un examen par une population de 100 étudiants que nous notons $\mu = 69{,}9$ (et non pas M = 69,9, car il s'agit de la moyenne de la population, et non pas celle de l'échantillon). De cette population, trois échantillons, chacun composé de 9 étudiants, sont aléatoirement tirés. Nous calculons la moyenne et l'écart-type des notes à l'examen pour chaque échantillon de N = 9 et, pour chaque, nous calculons l'erreur type de la moyenne à l'aide de la Formule 9.3. Chaque échantillon produit une estimation de μ qui comporte une erreur plus ou moins grande par rapport à la vraie moyenne dans la population (69,9). Au Tableau 9.1, on peut remarquer que, compte tenu de la moyenne de la population, il y a une erreur importante dans la moyenne produite par l'échantillon M_1. Sa moyenne (77,1) est très différente de la véritable moyenne de la population. Mais les deux autres échantillons, M_2 (71,0) et M_3 (69,0), donnent une moyenne très proche de μ.

Quiz rapide 9.9
Remarquez qu'au Tableau 9.1, nous ne calculons pas d'erreur type de la moyenne pour la première ligne du tableau. Pourquoi ?

Tableau 9.1
Intervalles de confiance à trois niveaux de confiance pour trois petits échantillons extraits aléatoirement d'une même population

	N	*M*		*Erreur type*	$IC_{68\%}$	$IC_{95\%}$	$IC_{99\%}$
Population	100	69,9	13,3	—	—	—	—
Échantillon 1	9	77,1	11,5	3,83	73,3 à 80,9	69,6 à 84,6	67,2 à 87,0
Échantillon 2	9	71,0	11,1	3,70	67,3 à 74,7	63,7 à 78,3	61,5 à 80,5
Échantillon 3	9	69,0	12,8	4,27	64,7 à 73,3	60,6 à 77,4	58,0 à 80,0

Examinons l'avant-dernière colonne du Tableau 9.1. Cette colonne donne l'intervalle de confiance à 95 % : la fourchette de valeurs à l'intérieur de laquelle la vraie valeur de la moyenne de la population se trouve probablement. Prenez l'échantillon M_3 (M = 69 et s_M = 4,27) : on voit que l'intervalle de confiance indique que la moyenne de la population se trouve entre 60,6 et 77,4. La vraie moyenne (μ = 69,9) se trouve effectivement entre ces deux bornes. Pour l'échantillon M_2 (71,0), l'intervalle de confiance situe la moyenne de la population entre 63,7 et 78,3, ce qui comprend aussi la véritable moyenne de la population. Regardez maintenant l'échantillon M_1 (M = 77,1), celui qui produit une estimation très erronée de la moyenne de la population. Même dans ce cas, l'intervalle de confiance produit une fourchette de valeurs qui inclut la véritable moyenne de la population (les bornes de son intervalle de confiance sont 67,2 à 87,0, ce qui comprend la véritable moyenne de la population, 69,9). Dans tous les cas, la fourchette de valeurs calculée par l'intervalle de confiance inclut μ. Tournons-nous maintenant vers le calcul de l'intervalle de confiance.

Le calcul de l'intervalle de confiance

Le calcul de l'intervalle de confiance, Formule 9.5, implique trois termes : M, Z et s_M. M, la moyenne de l'échantillon, est facile à calculer et l'erreur type de la moyenne se calcule aussi facilement ($s_M = s/\sqrt{N}$). Reste à expliquer Z.

Nous voulons calculer une fourchette de valeurs autour de la moyenne de l'échantillon M qui inclut μ. Supposons que nous voulons établir une fourchette de valeurs telle que ses bornes incluent 95 % des échantillons qui peuvent provenir de cette population. Nous savons, d'après le tableau de la densité sous la courbe normale, que 95 % des échantillons se trouvent à ±1,96 erreurs types de la moyenne. Ainsi, si nous multiplions l'erreur type de la moyenne par ±1,96, les bornes de l'intervalle prendront effectivement des valeurs qui incluront 95 % des échantillons extraits de cette population.

À titre illustratif, calculons l'intervalle de confiance pour l'échantillon M_1 du Tableau 9.1.

1. Calculons l'erreur type de la moyenne :
   $s_{M1} = s/\sqrt{N} = 11,5/\sqrt{9} = 3,83$.
2. Choisissons la valeur de Z, par exemple Z = 2,58 (intervalle de confiance à 99 %, parce que 99 % des échantillons se situent à ± Z = 2,58 erreurs types de la moyenne, tel qu'il est déterminé dans le tableau de la densité de la courbe normale).
3. Calculons la borne supérieure de l'intervalle :
   $IC_{sup} = M + Z \times s_M = 77,1 + (2,58 \times 3,83) = 86,98 = 87,0$.
4. Calculons la borne inférieure de l'intervalle :
   $IC_{inf} = 77,1 - (2,58 \times 3,83) = 67,2$.

Nous pouvons alors affirmer, à partir de notre unique échantillon, que la moyenne de la population se trouve entre 67,2 et 87,0 et que nous avons 99 % de chances de ne pas nous tromper dans cette conclusion. Dans ce cas, notre conclusion est juste mais nous aurions pu nous tromper. Moins on veut risquer de se tromper, plus grande doit être la fourchette de valeurs : l'intervalle de confiance.

La valeur Z et la taille de l'intervalle de confiance

On se souvient que Z est une valeur étalon, un indicateur de la distance entre la moyenne et une observation, ou, dans notre cas, la distance entre la moyenne d'un échantillon et la moyenne de la population. Nous avons vu qu'à partir du tableau de la densité de la courbe normale, nous pouvons trouver la proportion des échantillons qui se trouvent entre la moyenne et n'importe quelle valeur.

Par exemple, nous savons qu'environ 68 % des échantillons ont une moyenne se situant à ±1 erreur type de la moyenne de la population. Nous pouvons alors conclure qu'il y a 68 % de chances que la moyenne de la population se trouve à la moyenne trouvée dans notre échantillon ±1 erreur type. Mais s'il y a 68 % de chances que la moyenne μ se trouve à l'intérieur de cet intervalle, il y a 32 % de chances qu'elle ne s'y trouve pas. Ainsi, nous avons 32 % de chances que notre conclusion au sujet de la moyenne μ soit fausse (μ ne se trouve pas entre ces deux valeurs).

Nous pouvons réduire ce risque d'erreur en choisissant une fourchette de valeurs plus large. Nous savons que 95 % des observations d'une distri-

bution se trouvent à ±1,96 erreur type de la moyenne (Z = 1,96). Si nous calculons l'intervalle de confiance en utilisant cette nouvelle valeur, nous pouvons alors conclure qu'il y a 95 % de chances que μ se trouve entre ces nouvelles valeurs et 5 % de chances qu'elle ne s'y trouve pas. Nous savons aussi que 99 % des échantillons se trouvent à ±2,58 erreur type (Z = 2,58) de la moyenne et nous pouvons alors créer un intervalle de confiance où il y aurait 99 % de chances que la vraie moyenne de la population s'y trouve (et 1 % de chances qu'elle ne s'y trouve pas). En choisissant une valeur Z plus grande (±2,58 plutôt que ±1,96), nos chances de faire une erreur en concluant que l'échantillon n'appartient pas à la population chutent de 5 à 1 %. Nous avons plus confiance dans notre conclusion.

Le Tableau 9.1 montre les intervalles de confiance pour trois niveaux de confiance : 68 %, 95 % et 99 %. Les conclusions suivantes, tirées à partir de l'échantillon $M_{3,}$ sont valides :

1. Nous estimons que la moyenne obtenue à l'examen par la population d'étudiants est de 69 (en réalité elle est 69,9).
2. Il y a une probabilité p de 0,68 que la moyenne de la population se trouve entre 64,7 et 73,3, et il y a une probabilité p de 0,32 qu'elle ne se trouve pas entre ces deux valeurs. La fourchette est étroite mais le risque de se tromper est grand. Dans ce cas, nous ne nous sommes pas trompés, mais nous avons été chanceux.
3. Il y a une probabilité p de 0,95 que la moyenne de la population se trouve entre 60,6 et 77,4, et une probabilité p de 0,05 qu'elle ne soit pas entre ces deux valeurs. Le risque de se tromper est moins grand, mais la fourchette est beaucoup plus grande, passant de ±4,27 à ± 8,4.
4. Il y a une probabilité p de 0,99 que la moyenne de la population se trouve entre 58,0 et 80,0, et la probabilité que nous soyons dans l'erreur est de 0,01. La fourchette est très large et, par conséquent, le risque de se tromper est très faible.

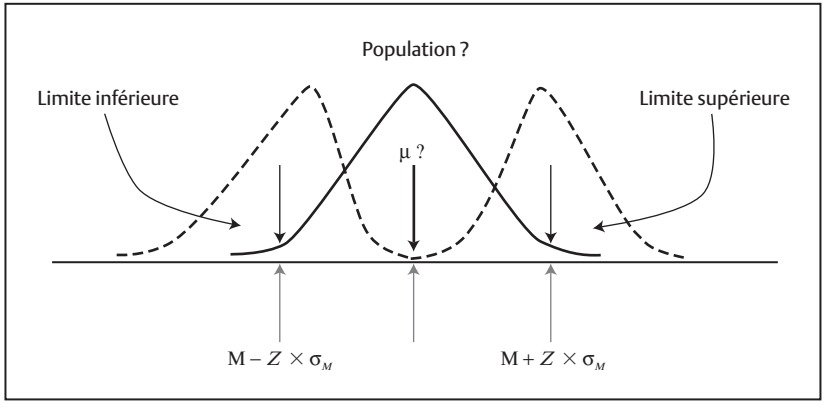
Regardez maintenant, au Tableau 9.1, les résultats obtenus pour le calcul de l'intervalle de confiance à 68 % pour l'échantillon 1 (M1=77,1). Nous trouvons une fourchette de valeurs qui indique que nous avons 68 % de chances d'avoir raison de conclure que la moyenne de la population se situe entre 73,3 et 80,9, mais nous avons du même coup 32 % de chances de nous tromper dans cette conclusion. Dans ce cas, nous nous sommes

effectivement trompés ! La vraie moyenne de la population ($\mu = 69{,}9$) n'est pas incluse dans l'intervalle de confiance. Ainsi, lorsque l'intervalle de confiance est plus étroit, les risques d'erreurs sont plus forts.

La Figure 9.2 illustre le concept de l'intervalle de confiance. La flèche noire épaisse représente la moyenne de l'échantillon et les flèches noires fines, les bornes de l'intervalle de confiance. La flèche épaisse en noir est la meilleure estimation que nous avons de la moyenne μ, c'est-à-dire la moyenne de l'échantillon. Les flèches étroites claires sont les valeurs possibles que la moyenne de la population peut prendre.

Ainsi, comme nous l'indique la Figure 9.2, l'intervalle de confiance donne les valeurs maximales et minimales que pourrait prendre la moyenne de la population. La valeur Z utilisée pour définir l'intervalle définit du même coup le risque que la moyenne de la population ne se trouve pas, en réalité, à l'intérieur de ces marges : le risque d'erreur d'inférence. Bien que, par tradition, nous établissions les intervalles de confiance avec des niveaux de certitude de 95 % ou de 99 %, nous pouvons établir les intervalles de n'importe quelle taille (68 %, 99,99 %, etc). Mais le principe demeure : plus grande est la valeur Z, plus larges sont les bornes de l'intervalle de confiance et plus grande est notre confiance que la moyenne de la population se situe en réalité entre elles. L'intervalle de confiance est la base sur laquelle repose la très célèbre « signification statistique ».

FIGURE 9.2 Intervalle de confiance et sa relation avec la moyenne de la population



Le principe du test de signification statistique sur un seul échantillon : H versus H_o

Nous avons maintenant en main tous les éléments requis pour comprendre la procédure à suivre pour opposer l'hypothèse à l'hypothèse nulle. Supposons, d'une part, que nous avons une population cible dont nous connaissons la moyenne et l'écart-type. Supposons, d'autre part, que nous avons un échantillon aléatoire que nous croyons appartenir à cette population. Nous voulons savoir si cet échantillon appartient (H_0) ou non (H) à cette population. Le jeu d'hypothèses se formule de la manière suivante.

$H_{0:}$ la moyenne de la population d'où est tiré l'échantillon est égale à la moyenne de la population cible (cet échantillon appartient à cette population cible ; $M = \mu$).

H : la moyenne de la population d'où est tiré l'échantillon n'est pas égale à la moyenne de la population cible (cet échantillon n'appartient pas à cette population cible ; $M \neq \mu$).

Pour décider, nous calculons la distance entre la moyenne de la population et la moyenne de l'échantillon. Si la probabilité p d'obtenir une telle différence dans un échantillon est inférieure à 0,05 (moins de 5 %), nous allons conclure que cet échantillon ne provient pas de la population cible (rejet de H_0). Si la probabilité p d'obtenir une telle moyenne dans l'échantillon est ≥ 0,05 (plus grande ou égale à 5 %), nous allons conclure que cet échantillon provient de la population cible (non-rejet de H_0).

En somme, il y a quatre étapes :

1. Poser les hypothèses :

   $H : M \neq \mu_0$

   $H_{0:}\ M = \mu_0$

   où μ_0 est la moyenne (connue) de la population cible.
2. Choisir le seuil de signification α désiré. Pour un niveau de confiance à 95 %, nous choisissons $\alpha = 0{,}05$. La valeur Z qui correspond à un niveau de confiance de 95 % est Z = 1,96.

   $\alpha = 0{,}05$
3. La décision est basée sur la règle suivante :

   Rejet de H_0 si M n'est pas inclus dans $\mu_0 \pm (Z \times \sigma_M)$.

4. Calculer la moyenne de l'échantillon, calculer l'erreur type de la moyenne et construire l'intervalle de confiance afin de conclure.

Par exemple, on prend un groupe de 16 autistes performants et on désire savoir si leur intelligence est comparable à l'intelligence de la population en général. Sur un test de QI, l'intelligence moyenne de la population μ est de 100 avec un écart-type de 16. Les hypothèses concernant la population des autistes performants sont donc les suivantes :

$$H : \mu \neq 100$$
$$H_{0:}\ \mu = 100$$

On utilise un seuil de signification α de 0,05, ce qui implique que $Z = 1{,}96$. La règle est donc de rejeter H_0 si la moyenne des autistes performants qu'on aura dans l'échantillon n'est pas incluse dans l'intervalle $100 \pm 1{,}96 \times \sigma/\sqrt{N}$. L'erreur type de la moyenne est $16/\sqrt{16} = 4$, ce qui donne l'intervalle de confiance $100 \pm 1{,}96 \times 4 = 100 \pm 7{,}84 =$ [92,16 à 107,84].

On calcule la moyenne de l'échantillon et on trouve 94. Cette moyenne est incluse à l'intérieur des bornes de l'intervalle de confiance : 95 % des échantillons extraits de cette population auront une moyenne qui se situe entre 92,16 et 107,84. Puisque l'échantillon a une moyenne (94) qui se situe entre ces deux valeurs, on conclut que cet échantillon fait partie de la population et qu'il ne représente pas une population de QI différente. S'il n'appartient pas à une population différente, on ne peut pas rejeter H_0 et on doit conclure qu'il n'y a pas de preuves voulant que les autistes appartiennent à une population de QI différente de celle de la population en général de QI. En jargon statistique, on dit qu'il n'y a pas de différence statistiquement significative entre le QI de la population en général et le QI des autistes performants.

Le test de signification statistique pour la différence entre deux échantillons

Il est possible d'étendre ce raisonnement pour opposer deux échantillons. Supposons que nous avons deux échantillons d'une population de patients qui souffrent de la maladie d'Alzheimer. Cette maladie du cerveau afflige certaines personnes âgées, créant des périodes de confusion mentale et de perte de mémoire de plus en plus sévères. On administre un médicament expérimental à un échantillon de patients, mais pas à l'autre. On mesure le

degré de confusion et de perte de mémoire des deux groupes. Si le nombre moyen de tels épisodes est plus petit pour le groupe qui reçoit le médicament que pour l'autre, nous concluons que le médicament est efficace.

Mais, à cause de l'erreur d'échantillonnage, une différence quelconque n'est pas une preuve suffisante pour conclure que le médicament est efficace. Après tout, les moyennes obtenues par deux échantillons qui n'ont pas reçu le médicament, ou deux qu'ils l'ont reçu, ne seront pas identiques.

On calcule alors un intervalle de confiance autour d'un échantillon. Si l'intervalle de confiance pour l'échantillon qui ne reçoit pas le médicament contient la moyenne du groupe qui reçoit le traitement (ou vice-versa), nous concluons que la moyenne obtenue dans cet échantillon aurait pu être obtenue par l'autre : la différence entre les deux échantillons n'est pas *statistiquement significative*, nous empêchant du coup de rejeter H_0 ; nous devons conclure que le médicament n'est pas efficace, car en jargon statistique, *les deux échantillons sont extraits de la même population* de malaises causés par la maladie.

Par contre, si la moyenne de l'échantillon qui reçoit le traitement n'est pas contenue dans l'intervalle de confiance construit autour de la moyenne de l'autre échantillon, nous concluons que cet échantillon appartient à une population différente, que cet échantillon est significativement différent de l'autre : *nous rejetons H_0 et nous concluons que le médicament est efficace.*

Reprenons cet exemple avec des chiffres :

1. Les hypothèses sont :

   H : $\mu_{\text{avec médicament}} \neq \mu_{\text{sans médicament}}$ (le nombre moyen de périodes de confusion ou de perte de mémoire pour ceux qui reçoivent le médicament n'est pas le même que pour ceux qui ne reçoivent pas le médicament).

   H_0 : $\mu_{\text{avec médicament}} = \mu_{\text{sans médicament}}$ (le nombre moyen d'épisodes de confusion ou de perte de mémoire pour ceux qui reçoivent le médicament est le même que pour ceux qui ne reçoivent pas le médicament).
2. Nous choisissons un seuil de signification de 5 %, d'où il s'ensuit que $Z = 1{,}96$.
3. Nous allons rejeter H_0 si la moyenne du groupe avec médicament est inférieure à la borne inférieure de l'intervalle de confiance de l'autre groupe. Supposons que l'écart-type des symptômes pour ce groupe s est

40 et qu'il y a 4 participants dans chaque échantillon, d'où $s_M = 40/\sqrt{4} = 40/2 = 20$.

4. Nous trouvons les résultats suivants :
   $M_{\text{avec médicament}} = 60$, d'où $IC_{95\,\%} = 60 \pm 1{,}96 \times 20 = 20{,}8$ à $99{,}2$
   $M_{\text{sans médicament}} = 100$, d'où $IC_{95\,\%} = 100 \pm 1{,}96 \times 20 = 60{,}8$ à $139{,}2$.

Dans le pire des cas, le degré de confusion/perte de mémoire moyen du groupe qui ne reçoit pas le médicament pourrait être aussi petit que 60,8. Le groupe qui reçoit le médicament obtient une moyenne de 60. Puisque cette moyenne n'est pas incluse dans l'intervalle de confiance du groupe qui ne reçoit pas le médicament (60,8 à 139,2), nous concluons que ceux qui reçoivent le médicament ne proviennent pas de la même population que ceux qui ne le reçoivent pas. La différence entre ces deux échantillons est statistiquement significative à $p < 0{,}05$ (le risque d'erreur dans cette conclusion est plus petit que 0,05) et, enfin, nous concluons que le médicament est efficace.

Supposons que la moyenne du groupe avec médicament est de 60,8, exactement à la valeur de la limite inférieure de l'intervalle de confiance. Par convention, lorsque la moyenne d'un échantillon est exactement à la limite de l'intervalle de confiance, on dit que la différence *n'est pas* statistiquement significative. La borne inférieure de l'intervalle de confiance d'un groupe doit être numériquement supérieure à la moyenne de l'autre échantillon. La moyenne obtenue dans l'échantillon avec traitement est significativement différente de la moyenne de l'autre échantillon lorsque les chances de se tromper dans la conclusion sont *moins* de 5 sur 100 ($p < 0{,}05$).

Ce que la signification statistique dit et ce qu'elle ne dit pas

La signification statistique est souvent mal interprétée non seulement par les étudiants mais aussi par les scientifiques. Voici quelques interprétations appropriées et inappropriées d'une différence statistiquement significative.

La signification statistique dit que :

- la population de laquelle un des échantillons est extrait est différente de la population de laquelle l'autre échantillon est tiré ;

- la probabilité que nous ayons incorrectement conclu qu'il existe deux populations au lieu d'une seule est égale au seuil alpha ;
- le traitement est efficace.

La signification statistique ne dit pas que :

- la différence entre deux populations est importante (une petite différence pourrait être statistiquement significative) ;
- la différence entre deux populations est réelle (nous courons invariablement un risque d'erreur qui est défini par le seuil alpha).

L'ERREUR DE TYPE I ET L'ERREUR DE TYPE II

Au chapitre 8, nous avons présenté le concept d'erreurs de type I et de type II. L'erreur de type I (l'erreur alpha « α ») se produit lorsqu'on conclut, à partir des échantillons, qu'il existe deux populations alors qu'en réalité il n'en existe qu'une. L'erreur de type II (l'erreur bêta « β ») survient lorsqu'on conclut, à partir de la moyenne des échantillons, qu'il n'y a pas de différence entre les populations, alors qu'en réalité il y en a une. Les erreurs α et β sont donc des images miroirs.

Supposons que le traitement pour la maladie d'Alzheimer ne soit pas efficace. C'est un coup de chance si la moyenne de l'échantillon qui a reçu le traitement est de 60. Dans ce cas, notre conclusion selon laquelle le traitement serait efficace (la différence entre la moyenne des deux échantillons est statistiquement significative) est erronée. Nous aurions donc commis une erreur d'inférence de type I.

Supposons que le traitement pour la maladie d'Alzheimer est véritablement efficace. Mais, malheureusement, la moyenne obtenue par l'échantillon de patients est de 60,8, à l'intérieur de l'intervalle de confiance de l'échantillon n'ayant pas reçu de médicament. La différence n'étant pas statistiquement significative, nous concluons, à tord dans ce cas, que le traitement est inefficace. Le traitement étant en réalité efficace, nous venons de faire une erreur d'inférence : une erreur de type II.

Les éléments qui affectent le risque d'une erreur de type I et de type II

L'inférence statistique (H et H_0) dépend de la taille de l'intervalle de confiance. Lorsque l'intervalle de confiance est très étroit, les bornes de cet intervalle sont plus proches l'une de l'autre. Il est donc moins probable que la moyenne d'un échantillon tombe à l'intérieur de l'intervalle de confiance de l'autre, ce qui se soldera, en bout de ligne, par une conclusion en faveur du rejet de H_0. Ainsi, lorsque l'intervalle de confiance est construit avec des bornes étroites (95 ou 90 % par exemple), le risque de commettre une erreur de type I augmente.

À l'inverse, lorsque l'intervalle de confiance a des bornes très larges, seuls les échantillons qui produisent une moyenne très différente se retrouveront à l'extérieur des bornes de l'intervalle de confiance. Par conséquent, seules les différences très grandes entre les moyennes des deux échantillons seront statistiquement significatives. Les chances de commettre une erreur de type I sont réduites mais les chances de commettre une erreur de type II sont plus grandes. Voyons pourquoi.

La taille de l'intervalle de confiance est déterminée par deux éléments : la valeur Z et la taille de l'erreur type de la moyenne.

La valeur Z est déterminée par le seuil de confiance α. Lorsque nous voulons minimiser le risque de commettre une erreur d'inférence de type I — et conclure correctement que deux échantillons diffèrent —, nous devons choisir un seuil α petit (ce qui équivaut à un Z plus grand). L'effet de ce choix sera d'élargir les bornes de l'intervalle de confiance. En élargissant les bornes, seules les grandes différences entre les moyennes des groupes pourront mener au rejet de H_0.

De son côté, l'erreur type de la moyenne est déterminée par deux éléments : le nombre d'observations (N) et l'écart-type des observations (*s*). Lorsque l'écart-type de l'échantillon est grand, l'erreur type de la moyenne est grande, et lorsque le nombre d'observations *N* est petit, l'erreur type de la moyenne est grande aussi. En général, nous ne pouvons pas vraiment agir pour réduire ou accroître la taille de l'écart-type de l'échantillon (ce sont les observations qui le déterminent). Mais nous pouvons avoir un impact sur le nombre d'observations. Nous pouvons choisir de mesurer 10 personnes ou 1 000. Par conséquent, en travaillant avec plus d'ob-

servations, nous réduisons la taille de l'erreur type de la moyenne, ce qui produira des intervalles de confiance plus étroits. Ces derniers étant plus étroits, les chances que les bornes de l'intervalle de confiance d'un groupe ne recoupent pas la moyenne de l'autre groupe augmentent, ce qui rend plus probable la conclusion en faveur du rejet de H_0.

Le Tableau 9.2 reprend l'exemple hypothétique du médicament pour traiter la maladie d'Alzheimer. Nous y présentons trois échantillons ayant un nombre d'observations différent (4, 16, 64). Tout en gardant constante la moyenne (100) et l'écart-type ($\sigma = 40$), nous voyons que l'erreur type de la moyenne se réduit (de 20 à 5) lorsque la taille des échantillons augmente (de 4 à 64 respectivement). Le syllogisme se comprend : plus grand l'échantillon, plus petite l'erreur type de la moyenne. Plus petite l'erreur type de la moyenne, plus étroit l'intervalle de confiance. Par conséquent, plus grand l'échantillon, plus étroit l'intervalle de confiance. Et, comme nous l'avons vu, il est plus probable que la différence entre les échantillons soit déclarée significative.

Ce constat est parfaitement raisonnable puisque l'échantillon plus grand contiendra une plus grande proportion des observations qui existent dans la population, ce qui devrait augmenter la confiance que nous avons dans la moyenne qu'il produit et dans la différence qui existe entre cette échantillon et la population ou un autre échantillon.

Lorsque nous travaillons avec de petits échantillons, l'erreur type de la moyenne est plus grande, ce qui cause des intervalles de confiance plus larges. Lorsque les bornes sont éloignées, il n'est possible de conclure à la signification statistique que lorsque les moyennes des deux échantillons sont très différentes. Lorsque la différence entre deux populations est réelle mais petite, et que nous la testons avec de petits échantillons, il est facile de commettre une erreur de type II, c'est-à-dire conclure que la différence n'est pas statistiquement significative. Qu'une seule population existe plutôt que deux.

Inversement, plus les échantillons sont de grande tailles, plus la probabilité de conclure qu'ils proviennent de la même population est petite. Donc, la probabilité de commettre une erreur de type II diminue.

Par ailleurs, plus petite est la valeur α, plus grande est la valeur Z : pour $\alpha = 5\,\%$, $Z = 1{,}96$ et pour $\alpha = 0{,}01$, $Z = 2{,}58$. L'accroissement de la valeur

Z entraîne l'accroissement des bornes de l'intervalle de confiance. Lorsque ces bornes s'élargissent, seules les grandes différences entre les moyennes permettent de conclure à la signification statistique. Par conséquent, lorsque le seuil α est plus petit, il devient plus difficile de rejeter H_0 et plus probable de conclure que la différence entre les moyennes n'est pas statistiquement significative. Le risque d'une erreur de type I est plus petit, mais le risque d'une erreur de type II est plus grand.

À l'inverse, l'intervalle de confiance est plus étroit lorsque le seuil de signification augmente (par exemple de $\alpha = 0{,}01$ à $\alpha = 0{,}05$). Augmenter ce seuil revient à dire que nous tolérons un risque d'erreur plus grand. L'intervalle se réduit et augmente nos chances de conclure à la signification statistique (rejet de H_0), occasionnant, en contrepartie, plus de risque de commettre une erreur de type I.

Tableau 9.2
Impact du nombre d'observations (*N*) et du seuil de signification (α) sur la taille de l'intervalle de confiance

				Intervalle de confiance		
M	***N***	***s***	σ_M	$IC_{68\%}$ ***(Z = 1, α = 0,32)***	$IC_{95\%}$ ***(Z = 1,96, α = 0,05)***	$IC_{99\%}$ ***(Z = 2,58, α = 0,01)***
100	4	40	20	80 à 120	60,8 à 139,2	48,4 à 151,6
100	16	40	10	90 à 110	80,4 à 119,6	74,2 à 125,8
100	64	40	5	95 à 105	90,2 à 109,8	87,1 à 112,9

Quiz rapide 9.10
Vous êtes le patron d'une compagnie pharmaceutique. Vous aurez le droit de mettre sur le marché votre nouvelle pilule seulement si vous êtes en mesure de démontrer statistiquement qu'elle est efficace. Vous voulez augmenter vos chances d'obtenir ce résultat. Devriez-vous tester l'efficacité de votre pilule sur un petit ou sur un grand échantillon de patients ? Votre α devrait-il être petit ou grand ?

Choisir entre les risques d'une erreur de type I ou de type II

De cet ensemble de considérations, il faut retenir que l'inférence statistique est un exercice qui consiste à établir les risques d'erreur d'inférence. En choisissant le seuil α ainsi que le nombre d'observations, nous choisissons automatiquement le risque d'erreurs de type I et II.

Le choix entre la réduction de l'erreur de type I ou l'erreur de type II dépend totalement du risque d'erreur que l'on désire minimiser. Lorsque le danger de rejeter incorrectement H_0 est plus élevé que le danger de ne pas le rejeter incorrectement, nous allons minimiser le risque d'une erreur de type II en utilisant des échantillons de grande taille et en choisissant un seuil de signification plus grand ($\alpha = 0{,}05$).

Par exemple, si on teste les effets secondaires d'un médicament, il est plus dangereux *de conclure à tort qu'il n'y a pas* d'effets secondaires nocifs (H_0) que de conclure, à tort, qu'il y en a (H). La compagnie pharmaceutique qui met sur le marché un médicament provoquant des effets secondaires importants sans avertir les patients s'expose à des poursuites judiciaires qui peuvent la mener à la faillite. Dans ce cas, le risque d'une erreur de type II (conclure qu'il n'existe pas d'effets secondaires alors qu'il en existe) est plus grave que de rapporter des effets secondaires qui n'existent pas. L'incidence d'effets secondaires causés par le médicament doit être testée avec de grands échantillons et/ou avec un seuil alpha plus grand ($\alpha = 0{,}05$ plutôt que $\alpha = 0{,}01$). Dans ce cas, même une légère différence sur le plan des effets secondaires sera statistiquement significative, ce qui encouragera la compagnie à signaler à sa clientèle un risque d'effets secondaires.

À l'inverse, avant d'investir d'énormes sommes d'argent dans le développement d'un nouveau médicament ainsi que dans des études cliniques de grande envergure légalement requises pour la mise en marché, la compagnie pharmaceutique se doit de vérifier s'il a de bonnes chances d'être efficace. Elle désire, dans ce cas, minimiser le risque d'une erreur de type I (conclure à tort que le médicament est efficace). Par conséquent, elle utilisera un échantillon de petite taille et un seuil alpha plus petit ($\alpha = 0{,}01$ plutôt que $\alpha = 0{,}05$), car seule une grande différence sera statistiquement significative. Si, avec ce petit N et ce seuil, elle conclut que le médicament est efficace (statistiquement significatif), elle a de très bonnes chances d'en

arriver à la même conclusion lorsqu'elle testera son efficacité avec des échantillons plus grands et plus coûteux.

SOMMAIRE DU CHAPITRE

Le processus d'échantillonnage aléatoire cause une variation inévitable entre la moyenne des échantillons extraits d'une population. La taille de cette erreur d'échantillonnage (l'erreur type de la moyenne) peut être estimée à partir de la variance et du nombre d'observations d'un seul échantillon. Il devient alors possible de créer un intervalle de confiance qui reflète le degré de variabilité aléatoire des échantillons. Dès lors, on peut comparer la moyenne de l'échantillon à la moyenne à laquelle on pourrait s'attendre (moyenne de la population cible). Cette comparaison est au cœur des tests de signification. On dit que deux échantillons sont statistiquement différents lorsqu'il est peu probable qu'ils puissent tous les deux provenir de la même population. Le test de signification indique le risque d'erreur de type I que l'on accepte en concluant que l'échantillon ne provient pas de la population ou que deux échantillons ne proviennent pas de la même population. Mais on doit aussi faire attention au risque d'une erreur de type II, le risque de conclure à tort que les deux échantillons proviennent de la même population. Lorsqu'il s'agit de minimiser le risque d'une erreur de type I, on utilise un seuil α et un N de petite taille. Lorsqu'il s'agit de minimiser l'erreur de type II, on fait l'inverse : on utilise un seuil α et un N plus grand.

COMMENT TROUVER L'ERREUR TYPE DE LA MOYENNE

Pour obtenir l'erreur type de la moyenne, il faut calculer la variance de M, notée Var(M). Une façon d'y arriver serait de prendre plusieurs échantillons puis de calculer la variance entre les moyennes de ces échantillons. En fait, c'est comme si l'on bâtissait un « méta-échantillon » Z contenant comme données brutes les moyennes $\{M_1, M_2, M_3, \ldots, M_e\}$. Évidemment, dans la pratique, nous n'avons pas le loisir de constituer plusieurs échantillons uniquement pour connaître l'erreur type de la moyenne. Heureusement, les statistiques ont résolu ce problème. Premièrement, il faut savoir

que, de façon générale, $\mathrm{Var}(X) = E(X^2) - E^2(X)$. Cette formule s'applique aussi pour M :

$$\mathrm{Var}(M) = E(M^2) - E^2(M) = E(M^2) - \mu^2.$$

Si on détaille le premier terme de la soustraction, on obtient :

$$\begin{aligned} M^2 &= \left(\frac{1}{n}\sum_i X_i\right)^2 = \frac{1}{n^2}(X_1 + X_2 + X_3 + \ldots + X_n)^2 \\ &= \frac{1}{n^2}\frac{1}{n^2}\left(X_1^2 + X_2^2 + X_3^2 + \ldots + X_n^2 + 2\sum_{i<j} X_i X_j\right) \\ &= \frac{1}{n^2}\left(\sum_i X_i^2 + 2\sum_{i<j} X_i X_j\right) \end{aligned}$$

Or, puisque $Var(X) = E(\mathbf{X}^2) - E^2(\mathbf{X}) = E(\mathbf{X}^2) - \mu^2$ et que $Var(\mathbf{X})$ est la meilleure estimation de σ^2, cela implique que, par simple réarrangement, $E(\mathbf{X}^2) = \sigma^2 + \mu^2$. De plus, $E(\mathbf{XX}) = E(\mathbf{X})E(\mathbf{X}) = \mu^2$. Finalement, si une variable *i* peut prendre toutes les valeurs de 1 à *n*, et que pour un *i* donné, la variable *j* peut prendre toutes les valeurs de 1 à *i* exclusivement, nous nous retrouvons avec $\frac{n(n-1)}{2}$ combinaisons de *i* et de *j*. Si on intègre tous ces éléments, nous pouvons noter que :

$$\begin{aligned} E(M^2) &= \frac{1}{n^2}\left(\sum_i (\sigma^2 + \mu^2) + 2\sum_{i<j} \mu^2\right) \\ &= \frac{1}{n^2}\left(n(\sigma^2 + \mu^2) + 2\frac{n(n-1)}{2}\mu^2\right) \\ &= \frac{1}{n^2}(n\sigma^2 + n\mu^2 + n^2\mu^2 - n\mu^2) \\ &= \frac{1}{n^2}(n\sigma^2 + n^2\mu^2) = \frac{\sigma^2}{n^2} + \mu^2 \end{aligned}$$

En intégrant la première équation et la dernière, nous obtenons :

$$Var(M) = \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}$$

Il ne reste plus qu'à prendre la racine carrée pour trouver l'erreur type de la moyenne.

EXERCICES DE COMPRÉHENSION

1. La meilleure estimation que nous avons de μ et de σ est _____ et _____ respectivement.
   a) la taille de l'échantillon ; sa variance
   b) M ; s
   c) N ; N – 1
   d) Toutes ces réponses sont justes.

2. L'échantillon A est composé de 100 observations, alors que l'échantillon B est composé de 1 000 observations. Toutes choses étant égales par ailleurs, l'erreur type de la moyenne pour l'échantillon B sera _____ l'échantillon A ?
   a) plus grande que celle de
   b) moins grande que celle de
   c) identique à celle de
   d) parfois plus grande et parfois plus petite que celle de

3. Si nous voulons minimiser nos chances de commettre une erreur d'inférence de type II, _____ .
   a) il faut choisir de petits échantillons et utiliser un seuil alpha plus petit (0,01 plutôt que 0,05)
   b) il faut choisir de grands échantillons et utiliser un seuil alpha plus petit (0,01 plutôt que 0,05)
   c) il faut choisir de petits échantillons et utiliser un seuil alpha plus grand (0,05 plutôt que 0,01)
   d) il faut choisir de grands échantillons et utiliser un seuil alpha plus grand (0,05 plutôt que 0,01)

4. Si nous voulons minimiser nos chances de commettre une erreur d'inférence de type I, _____ .
   a) il faut choisir de petits échantillons et utiliser un seuil alpha plus petit (0,01 plutôt que 0,05)
   b) il faut choisir de grands échantillons et utiliser un seuil alpha plus petit (0,01 plutôt que 0,05)
   c) il faut choisir de petits échantillons et utiliser un seuil alpha plus grand (0,05 plutôt que 0,01)

d) il faut choisir de grands échantillons et utiliser un seuil alpha plus grand (0,05 plutôt que 0,01)

5. Les chances de rejeter H0 sont plus grandes lorsque nous analysons _____.

a) des échantillons de petite taille
b) des échantillons ayant des moyennes proches de μ
c) des échantillons provenant d'une population ayant une grande variance
d) Aucune de ces réponses

6. Les bornes de l'intervalle de confiance sont plus larges lorsque nous choisissons un seuil alpha _____ et un N _____.

a) plus petit ; plus petit
b) plus petit ; plus grand
c) plus grand ; plus petit
d) plus grand ; plus grand

7. Nous avons une population qui est asymétrique positive. Nous tirons de cette population 1 000 échantillons, chacun composé de 100 observations, et nous calculons la moyenne pour chaque échantillon. Enfin, nous établissons la distribution des effectifs pour ces moyennes. Cette distribution sera approximativement _____.

a) asymétrique positive
b) asymétrique négative
c) normale
d) asymétrique positive, négative ou normale, selon le test statistique utilisé.

8. La moyenne de cet échantillon est égale à 11 et l'erreur type de la moyenne pour cet échantillon est de 1. La moyenne de la population de laquelle est extrait cet échantillon est égale à 10. Si $\alpha = 0{,}01$

a) Il est fort probable que cet échantillon provienne de cette population.
b) Il est fort probable que cet échantillon ne provienne pas de cette population.
c) Selon le résultat au test de signification, cet échantillon pourrait ou non provenir de cette population.

d) Ces trois réponses sont toutes également justes.

9. Nous avons trouvé une différence qui est statistiquement significative à $\alpha = 0{,}05$ entre deux échantillons. Par conséquent, _____ que les deux échantillons _____ de la même population.

a) il est certain ; proviennent
b) il est probable ; proviennent
c) il est certain ; ne proviennent pas
d) il est probable ; ne proviennent pas

Réponses

1. b
2. b
3. d
4. a
5. d
6. a
7. c (voir le théorème de la limite centrale)
8. a
9. d

CHAPITRE 10
UNE OU DEUX POPULATIONS ? LE TEST t

CHAPITRE 10

UNE OU DEUX POPULATIONS ? LE TEST t

Le test t — comme le test z décrit au chapitre 9 ou celui qui sera décrit dans le prochain chapitre (l'ANOVA) — est un test statistique qui permet de déduire, avec un risque d'erreur connu, si deux échantillons sont statistiquement différents, c'est-à-dire s'ils proviennent d'une seule population ou de deux. *La grande différence entre le test t et les autres est que celui-ci est optimisé pour fournir des inférences valides pour des échantillons de petite taille*. Bien que le test t soit utilisé principalement pour comparer deux petits groupes, il peut aussi être utilisé pour déterminer si un échantillon unique n'appartient pas à une population connue ou si le même groupe d'informateurs produit des résultats différents sur deux mesures différentes et/ou si le même groupe d'informateurs fournit une réponse moyenne différente sur la même variable lorsque celle-ci est administrée à deux moments différents.

Dans son utilisation principale, le principe du test t se comprend assez facilement. On calcule la différence entre la moyenne des deux échantillons que l'on va comparer à la différence typique à laquelle on peut s'attendre de deux échantillons tirés aléatoirement d'une population. Si la différence entre les deux moyennes est plus grande que la différence typique, on conclut, avec une probabilité d'erreur connue, que les deux échantillons sont extraits de populations différentes : la différence est statistiquement significative. Si la différence entre les deux moyennes n'est pas plus grande que la différence typique entre deux échantillons tirés de la même population,

nous concluons qu'il n'y a pas de preuves voulant que les deux échantillons n'appartiennent pas à la même population. La différence n'est pas statistiquement significative.

William S. Gosset, la statistique et la bière

Nous devons la statistique t et le test t, indirectement, à la bière ! Au début du xx^{e} siècle, William S. Gosset, chimiste et mathématicien employé par la brasserie britannique Guinness, prit congé de son employeur pour entrer, à titre d'étudiant, au laboratoire de Karl Pearson — le même Pearson qui nous a donné le coefficient de corrélation. Gosset décida de se pencher sur un problème pratique et théorique qui préoccupait les statisticiens, ainsi que les brasseurs, de l'époque.

Les caractéristiques de la distribution normale étaient fort bien connues et les spécialistes savaient s'en servir pour tirer des inférences. Mais était-il possible de se servir de cette distribution normale pour tirer des inférences alors que les échantillons étaient de petite taille ? Après tout, le théorème de la limite centrale (chapitre 9) indique que la distribution des échantillons s'approche de la normalité, mais seulement lorsque le nombre d'observations dans les échantillons est assez grand ($N \geq 30$).

Les systèmes de production de denrées alimentaires (y compris la bière) sont soumis à des contrôles de qualité. Des échantillons du produit sont aléatoirement choisis et analysés afin de tirer une inférence au sujet de la chaîne de production. Donne-t-elle un produit qui est conforme aux exigences de qualité et de pureté ? Ces analyses étant complexes et coûteuses, elles n'étaient appliquées que sur de petits échantillons (seulement quelques bouteilles de Guinness). Il fallait donc tirer une inférence au sujet de la population (la chaîne de production de la bière) à partir d'un très petit échantillon (quelques bouteilles de Guinness). C'est Gosset, le chimiste-brasseur-mathématicien, qui réussit le premier à résoudre le problème de l'inférence à partir d'un petit échantillon. Sa contribution : la statistique t, la distribution t et le test t.

POURQUOI UN « PETIT » ÉCHANTILLON ?

Les statisticiens préfèrent utiliser de grands échantillons plutôt que des petits. Les grands échantillons sont en effet plus aptes à nous renseigner sur la moyenne de la population, car l'erreur type de la moyenne est plus petite lorsque le nombre d'observations est plus grand. Des simulations montrent que, lorsque les échantillons contiennent au moins une trentaine d'observations, la distribution de la moyenne de ces échantillons commence à ressembler à la distribution normale. Ainsi, les statisticiens considèrent un échantillon « petit » lorsqu'il est composé de moins de 30 observations et « grand » lorsqu'il en contient plus.

Naturellement, ce critère est approximatif. Pour certaines populations, celles qui sont très symétriques, l'approximation à la distribution normale se fera avec des échantillons comprenant moins de 30 observations. À l'inverse, lorsque la population est très asymétrique (tels les salaires des joueurs de la NHL), seuls les échantillons contenant plus (et parfois beaucoup plus) de 30 observations conduiront à une approximation raisonnable de la distribution normale.

L'erreur type de la moyenne et les petits échantillons

On se souvient (voir le chapitre 9) que le calcul de l'erreur type de la moyenne (σ_M) permet de positionner la moyenne de l'échantillon par rapport à la moyenne de la population. Grâce à cette statistique, il est possible de calculer un intervalle de confiance qui, à son tour, est utilisé pour réaliser une inférence au sujet de la signification statistique.

L'erreur type de la moyenne se définit par le rapport entre l'écart-type de la population (σ) et le nombre d'observations N dans l'échantillon ($\sigma_M = \sigma/\sqrt{N}$). Puisque nous connaissons rarement l'écart-type de la population, cette formule est inutile en pratique. Mais comme nous l'avons vu au chapitre 9, nous pouvons estimer l'écart-type de la population à partir de l'écart-type de l'échantillon ($s_M = s/\sqrt{N}$). Le théorème de la limite centrale (chapitre 9), quant à lui, indique que l'approximation de l'écart-type de la population sera bonne à condition que le nombre d'observations N soit grand ($N \geq 30$).

Qu'arrive-t-il lorsque les échantillons sont petits ? La distribution de ces moyennes est-elle la même que la distribution Z ? W. S. Gosset eut l'idée lumineuse (et la persistance) d'établir empiriquement la forme de la distribution des moyennes pour les petits échantillons.

Gosset construit une population normale d'observations et il calcule la moyenne de cette population (μ). Utilisant la procédure d'échantillonnage avec remise (voir l'encadré), il tire de cette population plusieurs centaines d'échantillons ayant la même petite taille (par exemple N = 2). Pour chacun de ces petits échantillons, il calcule sa moyenne (M_i) qu'il compare à la moyenne (connue) de la population ($M_i - \mu$). Puisque les échantillons sont tous extraits de la même population, nous nous attendons à ce que la

différence entre leurs moyennes et la moyenne de la population soit égale à zéro. Mais à cause de l'erreur d'échantillonnage, nous savons que cela ne sera pas le cas. Il calcule, alors, pour chaque échantillon, l'erreur type de la moyenne ($s_M = s/\sqrt{N}$).

La sélection aléatoire avec ou sans remise

Supposons que nous avons une population comprenant cinq familles (A à E) de laquelle nous tirons des échantillons aléatoires de deux familles. Quelle chance la famille A a-t-elle d'être choisie dans le premier échantillon ? Puisque nous avons cinq familles, la probabilité pour n'importe quelle famille d'être choisie est de 1/5, p = 0,20. Mais selon quelle probabilité la famille B sera-t-elle choisie dans ce même échantillon ? Puisqu'il ne reste que quatre familles dans la population, cette probabilité est de 1/4, p = 0,25. La probabilité d'être choisi n'est pas la même pour les deux membres de cet échantillon. Cette inégalité dans les chances d'être choisi viole un principe fondamental de la sélection aléatoire (voir le chapitre 8). On nomme échantillonnage *sans remise* cette procédure d'échantillonnage.

Pour pallier cette difficulté, on a créé la procédure de *sélection aléatoire avec remise*. Son but est d'égaliser les chances d'inclusion dans un échantillon de tous les membres d'une population. À la suite de chaque tirage au sort, l'observation choisie est replacée dans la population la rendant admissible pour le prochain tirage. La sélection aléatoire avec remise donne à chaque membre de la population une chance d'être choisi qui est exactement égale.

En pratique, on utilise rarement la sélection aléatoire avec remise parce qu'elle n'est pas nécessaire. Nous travaillons généralement avec des populations de très grande taille (des millions d'observations potentielles). Le biais de sélection que la procédure de sélection aléatoire sans remise occasionne est, par conséquent, négligeable. Mais lorsque l'on travaille avec de petites populations (par exemple les patients atteints d'une maladie très rare), l'échantillonnage avec remise est obligatoire.

Ces deux informations — a) la différence entre la moyenne de chaque échantillon et la moyenne de la population ($M_i - \mu$) et b) l'erreur type de la moyenne ($s/\sqrt{N}$) — sont divisées pour produire la statistique t décrite par la Formule 10.1

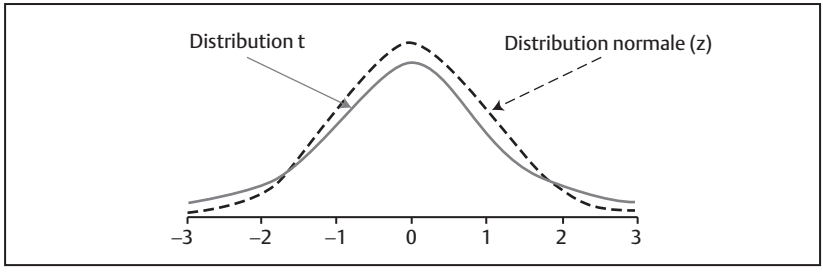
$$t = \frac{M - \mu}{s/\sqrt{N}} \qquad \text{Formule 10.1}$$

Nous voyons alors que la statistique t décrit la distance qui existe entre la moyenne d'un échantillon et la moyenne de la population ($M - \mu$) par rapport à la différence typique (l'erreur type de la moyenne).

Gosset construit une distribution des statistiques t obtenues sur les innombrables échantillons de même taille (N = 2, N = 3, etc.). Empirique-

ment, il découvre que cette distribution prend la forme générale d'une distribution unimodale (où les trois mesures de tendance centrale coïncident ; voir le chapitre 5). Mais les extrémités de la distribution des valeurs t ne sont pas les mêmes que celles de la distribution normale : la proportion des valeurs t plus extrêmes est plus grande que celle à laquelle nous pourrions nous attendre si la distribution des valeurs t suivait la distribution normale standardisée (Z). La Figure 10.1 représente le graphe des polygones décrivant la distribution normale et la distribution de la statistique t.

FIGURE 10.1 Distribution normale et distribution de la statistique t



Gosset répète ce processus d'échantillonnage en augmentant le nombre d'observations systématiquement (N = 3, N = 4, N = 30, etc.). Pour chaque série d'échantillons de même taille, il calcule la statistique t et établit pour chacune la distribution de ces statistiques t. En comparant les divers polygones des fréquences de la statistique t, il constate que la forme exacte de la distribution de la statistique t varie en fonction de la taille de l'échantillon. Lorsque le nombre d'observations est petit, la distribution t s'éloigne de la distribution normale Z. À l'inverse, lorsque les échantillons sont de grande taille, la distribution t est plus similaire à la distribution Z, et avec un nombre infini d'observations, les distributions Z et t sont parfaitement identiques.

L'intervalle de confiance pour les petits échantillons

On se souvient (voir chapitre 9) que l'intervalle de confiance est déterminé par la densité des observations sous la courbe normale (Z) et par l'erreur type de la moyenne (σ_M) que nous estimons avec s_M. La formule finale pour

le calcul de l'intervalle de confiance est $\mu \pm Z \times \sigma_M$. Nous utilisons la statistique Z parce qu'elle nous indique la proportion des échantillons qui se trouvent entre n'importe quelle valeur et la moyenne de la population. Par exemple, nous savons que 95 % des échantillons tirés d'une distribution normale se situent à une distance Z = ±1,96 de la moyenne de la population. Ceci découle directement du théorème de la limite centrale.

Le travail de Gosset indique que la distribution des moyennes des petits échantillons par rapport à la moyenne de la population est décrite avec plus de précision par la distribution t. Le calcul d'un intervalle de confiance pour les petits échantillons ne peut pas, par conséquent, suivre la forme habituelle : $\mu \pm Z \times \sigma_M$ puisque la distribution Z n'est pas la distribution qui décrit le mieux la forme de la distribution pour les petits échantillons. Il faudrait plutôt faire appel à la distribution de la statistique t. La Formule 10.2 décrit la forme qui est appropriée lorsque l'intervalle de confiance est construit sur de petits échantillons.

Pour calculer cet intervalle de confiance, il faut trouver une valeur t qui inclut 95 % des échantillons extraits de la même population. Similairement, pour avoir plus de certitude dans notre conclusion, nous pouvons trouver une valeur t qui recoupe 99 % des échantillons, c'est-à-dire choisir un seuil $\alpha = 0{,}01$ (voir le chapitre 9). Nous allons appeler cette valeur le $t_{critique}$ (certains statisticiens préfèrent le terme t [dl] et nous utilisons ici ces deux termes de façon interchangeable).

Mais il y a un problème : les valeurs $t_{critiques}$ dépendent de la taille de l'échantillon (en revanche, la valeur critique Z ne dépend pas de N). Comme nous le verrons plus loin, les valeurs $t_{critiques}$ sont déjà établies pour toutes les tailles des échantillons entre N = 3 et environ N = 120[1]. Elles sont reproduites dans le tableau des valeurs critiques de t dans l'Annexe. Nous verrons plus loin comment lire et interpréter ce tableau. Présumons pour l'instant que la valeur $t_{critique}$ est trouvée.

Une fois cette valeur $t_{critique}$ trouvée, nous pouvons alors utiliser la Formule 10.2 pour calculer l'intervalle de confiance autour de la moyenne de n'importe quel échantillon

1. Le tableau des valeurs critiques de t va jusqu'à *N* = 120 parce que, pour les échantillons de plus grande taille, la distribution t devient quasi identique à la distribution Z. Il n'est donc plus nécessaire de s'en servir.

$$\mu \pm t_{critique} \times s_M \qquad \text{Formule 10.2}$$

où $t_{critique}$ est une valeur qui définit la proportion des valeurs t qui inclut 95 % (ou 99 %) des valeurs t de la distribution.

Cette formule d'intervalle de confiance pour les petits groupes est identique à celle utilisée pour les grands groupes, sauf que la valeur critique se trouve à partir de la distribution t plutôt qu'à partir de la distribution Z. Le calcul de cet intervalle de confiance nécessite le calcul de l'erreur type de la moyenne ($s_M = s/\sqrt{N}$), où s est l'écart-type de l'échantillon. Il faut donc calculer l'écart-type de l'échantillon (s) que nous divisons par la racine carrée du nombre d'observations. Il faut aussi connaître la valeur critique de t que nous trouvons dans un tableau (voir l'Annexe). Il faut maintenant apprendre à lire le tableau des valeurs critiques de t.

Le tableau des valeurs critiques de t

Le tableau des valeurs critiques de t se trouve dans l'Annexe A.2 et le Tableau 10.2 (p. 314) en présente un extrait. Ce tableau est composé de rangées et de colonnes. Chaque rangée définit le nombre de degrés de liberté dans l'échantillon. Le nombre de degrés de liberté pour chaque échantillon est donné par N – 1 : le calcul de l'erreur type dépend de l'écart-type de l'échantillon. L'écart-type, à son tour, est calculé en fonction du nombre de degrés de liberté, N – 1. Si on a six observations dans un échantillon, il contient donc cinq degrés de liberté.

Pour trouver le $t_{critique}$ requis pour l'établissement de l'intervalle de confiance, il faut préalablement calculer le nombre de degrés de liberté, N – 1. Nous trouvons alors la rangée du tableau des valeurs critiques de t qui correspond au nombre de degrés de liberté dans l'échantillon du Tableau 10.2.

Il faut ensuite déterminer un seuil α approprié. Désirons-nous produire un intervalle de confiance doté de bornes étroites ou larges ? Tout comme nous l'avons étudié au chapitre 9, si nous désirons réduire le risque d'une erreur alpha, nous choisissons un seuil α très petit ($p < 0{,}01$) plutôt qu'un seuil plus grand ($p < 0{,}05$). Les colonnes du tableau des valeurs critiques identifient le seuil alpha désiré.

La valeur critique de t est la valeur qui est inscrite dans le tableau à l'intersection de la rangée qui correspond au degré de liberté et de la colonne

qui correspond au seuil alpha désiré. Par exemple, si nous avons un échantillon composé de 7 personnes, les degrés de liberté sont N – 1 = 6, et pour un seuil de $\alpha = 0{,}05$, la valeur $t_{critique} = 2{,}447$.

Quiz rapide 10.1
Trouvez dans le tableau des valeurs critiques de t (dans l'Annexe) la valeur critique $\alpha = 0{,}05$ et 0,01 pour un échantillon contenant un total de 12 observations.

L'idée à retenir est que la distribution des moyennes des petits échantillons s'apparentant à une distribution t n'est pas la même que celle produite par des grands échantillons, qui, elle, est la distribution Z. À partir de ces considérations, il est possible d'expliquer l'utilisation de la statistique t dans trois applications distinctes.

Le test t pour un seul échantillon

Cette version du test t est utilisée pour déterminer si un petit échantillon est différent de la moyenne hypothétique de la population lorsque la variance de la population est inconnue (l'échantillon appartient-il à cette population X ?). Par exemple, une nouvelle marque de voiture a-t-elle le degré de consommation de carburant que prétend le manufacturier ?

Le test t pour deux échantillons indépendants

Cette version sert à déterminer si deux petits échantillons ont des moyennes différentes, c'est-à-dire s'ils appartiennent à deux populations différentes. Par exemple, une technique chirurgicale est-elle plus efficace qu'une autre ?

Le test t pour deux échantillons non indépendants, ou le test t pour les données jumelées

Pour déterminer si le même petit échantillon diffère sur deux variables. Cette dernière application est très utile lorsqu'il s'agit d'évaluer le changement. Par exemple, la compréhension de la statistique dans un cours s'est-elle améliorée à la suite d'un premier examen ?

LE TEST T POUR UN ÉCHANTILLON

Le test t sur un seul échantillon est utilisé afin de déterminer si un échantillon provient ou non d'une population dont on croit connaître la moyenne, mais pas la variance. Il consiste à établir un intervalle de confiance (par exemple à 95 %) autour de la moyenne de la population. Si la moyenne de l'échantillon tombe à l'intérieur des bornes de cet intervalle de confiance, on n'aura pas de raison de conclure que cet échantillon n'appartient pas à la population. Mais si la moyenne de l'échantillon tombe à l'extérieur de l'intervalle, on aura alors de bonnes raisons de croire que l'échantillon n'appartient pas à cette population.

Supposons que, dans une grande manufacture de circuits électroniques, l'employé moyen monte 100 circuits par jour. Nous pouvons dire que la moyenne de productivité de cette population est $\mu = 100$. Un cadre met sur pied un programme de formation qui vise à accroître la productivité. Vingt-cinq employés tirés au hasard participent à ce programme. On mesure ensuite la productivité de ce groupe d'employés et on trouve qu'en moyenne ces N = 25 employés produisent M = 107 circuits par jour et que l'écart-type de son échantillon s = 15. Ce programme de formation améliore-t-il la productivité ? Formalisons le jeu d'hypothèses. Nous postulons (H) que la productivité des employés qui ont reçu la formation n'est pas la même que celle des employés en général. L'hypothèse nulle (H_0) est que leur productivité est en réalité la même que celle de la population.

$$H_0 : \mu = 100$$
$$H : \mu \neq 100$$

Pour que ce programme soit jugé efficace, il faut démontrer qu'il est peu probable d'avoir une productivité de 107 circuits dans un échantillon, alors que la population en produit en moyenne 100. Il faut donc établir un intervalle de confiance en se servant de la Formule 10.2.

Nous choisissons un seuil de signification α de 0,05. Le test est de la forme :

$$\text{Rejet de } H_0 \text{ si M n'est pas inclus dans } \mu \pm t_{\text{critique}} \times s_M.$$

Il faut préciser les degrés de liberté. Ici, nous avons dû calculer l'écart-type de l'échantillon où toutes les données sauf une sont libres. Nous avons

donc $N - 1$ degrés de liberté, où N est le nombre d'observations. Pour N = 25, les degrés de liberté sont 24.

Notre échantillon contient N = 25 observations, sa productivité moyenne est M = 107 et l'écart-type de cette productivité est s = 15. Nous pouvons maintenant calculer l'intervalle de confiance et tester notre hypothèse.

a) Calculer $s_M = s/\sqrt{N} = 15 / 5 = 3$.
b) Chercher dans la table t la valeur du $t_{critique}$. Les degrés de liberté étant $N - 1 = 24$, nous trouvons dans le tableau des valeurs critiques de t, la valeur $t_{critique} = 2{,}06$ à l'intersection de 24 degrés de liberté et de la colonne $\alpha = 0{,}05$. Le test est donc :

$$\text{Rejet de } H_0 \text{ si 107 n'est pas inclus dans } 100 \pm 2{,}06 \times 3{,}00$$
$$= 100 \pm 6{,}18 = 93{,}82 \text{ à } 106{,}18.$$

Quatre-vingt-quinze pour cent des échantillons de 25 travailleurs aléatoirement extraits de cette population auraient une productivité moyenne variant entre 93,8 et 106,2 circuits électroniques. Notre échantillon de personnes formées produit, en moyenne, 107 circuits, un degré de productivité qui n'est pas inclus dans l'intervalle de confiance. Par conséquent, nous rejetons H_0 et concluons que la productivité de cet échantillon n'appartient pas à la distribution de productivité de la population générale de travailleurs de cette entreprise. Dans ce cas, nous concluons que le programme de formation est efficace (il résulte en un degré de productivité plus grand que celui de la population de travailleurs qui n'ont pas reçu de formation).

Lorsque les observations qui appartiennent à un échantillon ne peuvent pas appartenir à un autre, on dit que les échantillons sont indépendants. Le test t utilisé dans ces conditions est appelé le *test t pour deux échantillons indépendants*. Par contre, dans certaines études, les mesures sont prises sur les mêmes individus. Ces études sont particulièrement utiles lorsqu'il s'agit d'évaluer le changement. Par exemple, les symptômes de maladie sont-ils aussi fréquents avant qu'après un traitement médical ? Le test t que l'on utilise prend alors le nom de *test t pour échantillons pairés ou jumellés*, aussi appelé *test t pour échantillons dépendants*. Nous expliquerons cette forme du test t plus loin de ce chapitre.

Le lien entre un intervalle de confiance et le test t pour un échantillon

Utilisons le symbole t(dl) pour indiquer le $t_{critique}$. Une autre façon de voir le test t consiste à noter que :

$$M \text{ n'est pas inclus dans } \mu \pm t(dl) \times s_M.$$

Cela revient au même que de dire :

$$M < \mu - t(dl) \times s_M \text{ ou } M > \mu + t(dl) \times s_M$$

e. g. M est en bas de la limite inférieure ou au-dessus de la limite supérieure. Si l'on réaménage quelque peu ceci, on obtient :

$$M - \mu < -t(dl) \times s_M \text{ ou } M - \mu > +t(dl) \times s_M$$

ou de façon équivalente :

$$\frac{M - \mu}{s_M} < -t(dl) \text{ ou } \frac{M - \mu}{s_M} > +t(dl)$$

Cela signifie que $\frac{M - \mu}{s_M}$, ignorant le signe, doit excéder t(dl). On dit que la valeur absolue de $\frac{M - \mu}{s_M}$, notée $\frac{|M - \mu|}{s_M}$, doit être plus grande que la valeur absolue de t(dl).

Cela permet un raccourci :

$$\frac{|M - \mu|}{s_M} > t(dl)$$

où t(dl) est sans signe. Aussi, une façon concise d'écrire le test t sur un échantillon est :

$$\text{Rejet de } H_0 \text{ si } \frac{|M - \mu|}{s_M} > t(dl)$$

Le test t sur un échantillon est généralement connu sous cette dernière forme, mais en fait, c'est exactement le même test que lorsqu'on a utilisé des intervalles de confiance !

La partie gauche de l'équation, sans valeur absolue, est parfois appelée la statistique $t_{observé}$, à ne pas confondre avec t(dl), le $t_{critique}$:

$$t_{observé} = \frac{M - \mu}{s_M}$$

Quiz rapide 10.2

Nous testons une thérapie avec un schème avant-après. Est-ce que les données forment un échantillon ? deux échantillons indépendants ? deux échantillons pairés ? Nous testons une nouvelle méthode d'enseignement du français en 6e année. Nous essayons la nouvelle méthode pendant une année et nous comparons les résultats à ceux de l'année précédente. S'agit-il d'échantillons indépendants ou d'échantillons pairés ?

LE TEST T POUR DEUX ÉCHANTILLONS INDÉPENDANTS

Le test t pour deux échantillons indépendants est la forme qui est la plus utilisée. Imaginons la situation où nous voulons déterminer si un nouveau médicament améliore l'état de santé de patients souffrant de la maladie d'Alzheimer. Nous tirons aléatoirement deux petits échantillons de cette population de patients. À un groupe, nous administrons le médicament, alors que nous ne le faisons pas pour l'autre groupe. Quelques semaines ou quelques mois plus tard, nous mesurons l'état de santé des patients dans chaque groupe et nous calculons une moyenne pour chacun des groupes. La question est : l'état de santé moyen du groupe qui reçoit le traitement est-il différent de (ou supérieur à) celui du groupe qui ne reçoit pas le traitement ? Plus formellement, les deux groupes appartiennent-ils ou non à la même population ?

La logique de base pour le test t pour échantillons indépendants

Si les deux échantillons sont extraits de la même population (c'est-à-dire que le médicament ne change rien), nous pouvons nous attendre à n'obtenir aucune différence entre les moyennes des deux groupes. Or, à cause de l'erreur d'échantillonnage, il est quasi certain que la différence entre ces deux échantillons ne sera pas exactement de zéro. Il faut donc examiner la différence entre la moyenne des deux groupes et l'interpréter à la lumière de l'erreur d'échantillonnage.

Nous pouvons estimer la différence typique qui existe entre deux échantillons aléatoirement tirés de la même population. Il s'agit de calculer *l'erreur type de la différence* entre deux échantillons (nous allons voir comment procéder plus loin). À partir de cette erreur type de la différence, nous pouvons générer un intervalle de confiance en fonction du seuil de signification désiré. Puis, nous calculons la différence observée entre les deux échantillons. Si la différence entre eux tombe à l'extérieur de (est plus grande que) l'intervalle de confiance, nous concluons que la différence observée dans ces échantillons est statistiquement significative : les échantillons n'appartiennent pas à la même population. Puisque la seule différence entre les deux groupes est que l'un prend un médicament et l'autre

pas, force est de conclure que le médicament a un effet. Si la différence entre les moyennes est incluse dans l'intervalle de confiance, la différence n'est pas significative et il n'est pas possible de conclure que le médicament produit l'effet escompté.

Le calcul de la statistique $t_{observé}$ pour les échantillons indépendants

Le calcul de la statistique $t_{observé}$ est plus complexe pour un test sur deux échantillons indépendants. Nous présentons les diverses formules requises, mais, en pratique, les logiciels d'analyses statistiques (SPSS ou Excel) font ces calculs automatiquement.

La première étape est d'obtenir une estimation de l'erreur type. Il faut estimer σ, l'écart-type de la population, ce que l'on fait à partir de l'écart-type de l'échantillon : l'erreur type est obtenue en divisant l'écart-type de l'échantillon (s) par la racine carrée du nombre de sujets. Mais là, un choix est à faire : il y a deux échantillons. Va-t-on utiliser l'écart-type de l'échantillon 1 ou de l'échantillon 2 ? Lequel est le meilleur pour estimer l'écart-type de la population ?

En fait, aucun ne l'est. Selon l'hypothèse nulle, les deux échantillons proviennent de la même population (le test t pour échantillons indépendants dira si l'on a raison ou tort). Si toutes les données des deux échantillons viennent de la même population, pourquoi ne pas les regrouper ensemble pour estimer σ ? Appelons S_c^2 la variance combinée des deux groupes. La variance combinée se calcule par

$$S_c^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 - 1) + (N_2 - 1)} \qquad \text{Formule 10.3}$$

où s_1^2 et s_2^2 sont les variances de chaque échantillon et N_1 et N_2 représentent le nombre d'observations dans chaque échantillon.

En fait, cette formule indique que S_c^2 est la moyenne des deux variances. L'ajout des termes ($N_1 - 1$ et $N_2 - 1$) au numérateur de la formule est nécessaire lorsque les deux échantillons ne sont pas de même taille (N). En multipliant les variances de chaque échantillon par $N_1 - 1$, nous créons une variance moyenne pondérée, qui donne plus d'importance à l'échantillon qui contient plus d'observations. Cela est raisonnable puisque l'échantillon

qui contient plus d'observations produit une estimation de la population qui est plus précise que celle produite par un échantillon plus petit.

L'erreur type de la moyenne se calcule par $s/\sqrt{N}$. Puisque S_c^2 est la variance combinée, en la divisant par N_1, nous obtenons l'erreur type (au carré) dans l'échantillon 1, et en divisant S_c^2 par N_2, nous obtenons l'erreur type (au carré) dans l'échantillon 2. Quelle erreur type doit-on prendre pour estimer l'erreur type de la différence entre la moyenne 1 et la moyenne 2 ? La plus grande ? La moyenne des deux ? Il faut savoir que l'erreur de mesure, ou erreur d'échantillonnage, est toujours croissante : si on soustrait deux mesures, chacune entachée d'erreurs, l'erreur totale est la somme des erreurs individuelles. En ce qui concerne les erreurs d'échantillonnage, ce sont les erreurs carrées qu'on doit additionner, puis il faut prendre la racine carrée pour obtenir une erreur typique, ce qui donne :

$$s_{M_1 - M_2} = \sqrt{\frac{s_c^2}{N_1} + \frac{s_c^2}{N_2}}$$

ce qui se simplifie en :

$$s_{M_1 - M_2} = s_c\sqrt{1/N_1 + 1/N_2} \qquad \text{Formule 10.4}$$

où $S_{M_1 - M_2}$ est l'erreur type qui résulte du calcul de la différence entre deux moyennes et s_c est l'écart-type de la variance combinée S_c^2, obtenu en faisant la racine carrée.

La Formule 10.4 nous donne $S_{M_1 - M_2}$ qu'on appelle *l'erreur type de la différence.* Elle indique la différence typique entre les moyennes de deux groupes. Cette mesure peut finalement être utilisée pour calculer une statistique t qui, elle, sera en mesure de tester la différence entre les deux groupes indépendants. Cette valeur t, que l'on nomme le $t_{observé}$, est celle que nous allons comparer éventuellement au tableau des valeurs critiques de t. La statistique présente donc le rapport de la différence observée entre les deux moyennes et l'erreur type de la différence moyenne entre deux échantillons extraits de la même population. La Formule 10.5 décrit la forme finale que prend le test t pour deux échantillons indépendants.

$$t_{observé} = \frac{M_1 - M_2}{s_{M_1 - M_2}} \qquad \text{Formule 10.5}$$

Il s'agit maintenant de tirer une conclusion. Nous avons le $t_{observé}$ et à partir du tableau des valeurs critiques de t, nous trouvons la valeur $t_{critique}$ qui correspond au nombre de degrés de liberté et au seuil alpha désiré. Si la valeur $t_{observé}$ est égale ou plus grande que la valeur du $t_{critique}$, nous concluons que les deux groupes n'appartiennent pas à la même population, qu'ils sont statistiquement différents.

Les degrés de liberté du test t pour les échantillons indépendants

Pour trouver la valeur du t *critique,* nous nous servons du tableau des valeurs critiques de la statistique t (voir l'appendice). Il nous faut donc trouver la cellule qui correspond à nos degrés de liberté pour le seuil d'erreur choisi. Mais supposons que le nombre d'observations dans chaque échantillon n'est pas identique. Alors, quel sera le nombre de degrés de liberté ? Celui qui correspond au premier ou au deuxième échantillon ? Comme pour le calcul de l'erreur type de la différence, ni l'un ni l'autre, mais les deux ! Ainsi, nous additionnons le nombre d'observations dans chaque groupe. Puisque nous additionnons ensemble les N, nous devons aussi additionner les degrés de liberté. Nous perdons un degré de liberté pour chaque groupe et, au total, nous en perdons deux. Ainsi, le nombre de degrés de liberté devient $(N_1 - 1) + (N_2 - 1)$ ou, plus simplement, $N_1 + N_2 - 2$. Nous cherchons donc le t *critique* à l'intersection de la colonne désirée et du nombre de degrés de liberté $N_1 + N_2 - 2$. Si nous avons deux groupes, chacun ayant 10 observations, le nombre de degré de liberté est de 18 (10 + 10 – 2).

Un exemple de calcul pour le test t pour les échantillons indépendants

Le calcul de la variance combinée

Au Tableau 10.1, nous reprenons un exemple médical. Un groupe de patients (l'échantillon 1) reçoit un médicament et l'autre (l'échantillon 2) n'en reçoit pas. Après quelques mois, on mesure, pour chaque patient de chaque groupe, le niveau de symptômes, un nombre élevé voulant dire plus de symptômes. Chaque groupe est composé de $N_1 = N_2 = 50$ observations.

Nous calculons la moyenne de chaque échantillon ($M_1 = 10$ et $M_2 = 20$) et la variance dans ces deux échantillons ($S_1^2 = 12$ et $S_2^2 = 20$). L'application de la Formule 10.3 donne la variance combinée qui est indiquée au Tableau 10.1.

Tableau 10.1
Calcul de la variance combinée S_c^2i

	Échantillon 1	*Échantillon 2*
	(reçoit le médicament)	(ne reçoit pas le médicament)
M	10	20
s^2	12	20
N	50	50

$$
\begin{aligned}
S_c^2 &= \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 - 1) + (N_2 - 1)} \\
&= \frac{(50 - 1)12 + (50 - 1)20}{(50 - 1) + (50 - 1)} \\
&= \frac{49 \times 12 + 49 \times 20}{98} \\
&= (588 + 980) / 98 \\
&= 16
\end{aligned}
$$

Le calcul de la variance combinée, dans ce cas, donne 16. En fait, comme les groupes sont égaux, il s'agit de la moyenne entre 12 et 20. L'écart-type de cette variance combinée se calcule en extrayant sa racine carrée. Dans ce cas, $S_c = \sqrt{16} = 4{,}0$.

Quiz rapide 10.3
Recalculez la variance combinée du Tableau 10.1 avec la Formule 10.3, mais, cette fois, le nombre d'observations est de 50 pour l'échantillon 1 et de 500 pour l'échantillon 2. La variance combinée est-elle toujours 16 ? Pourquoi ?

Le calcul de l'erreur type de la différence entre deux moyennes

À partir de la variance combinée S_c^2, nous pouvons calculer l'erreur type de la différence en utilisant la Formule 10.4. L'erreur type de la différence

indique la différence moyenne à laquelle nous pourrions nous attendre si les deux échantillons provenaient de la même population (les deux ayant ou n'ayant pas reçu de médicaments).

$$\begin{aligned} s_{M_1 - M_2} &= S_c \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \\ &= \sqrt{16} \sqrt{\frac{1}{50} + \frac{1}{50}} \\ &= 0{,}80 \end{aligned}$$

Dans ce cas, la différence typique à laquelle nous pouvons nous attendre entre ces deux échantillons, s'ils sont tirés de la même population, est de 0,80.

Le calcul de la statistique $t_{observé}$

Ayant maintenant en main l'erreur type de la différence, nous pouvons enfin calculer la statistique $t_{observé}$ en utilisant la formule pour son calcul (Formule 10.5).

$$t_{observé} = \frac{M_1 - M_2}{s_{M_1 - M_2}}$$

Pour les données du Tableau 10.2, nous calculons la différence entre les deux moyennes (10 – 20 = –10) et nous divisons cette différence par l'erreur type de la différence $t_{observé}$.

$$\begin{aligned} t_{observé} &= \frac{M_1 - M_2}{s_{M_1 - M_2}} \\ &= \frac{10 - 20}{0{,}8} = -12{,}5 \end{aligned}$$

$$t_{observé} = -12{,}5$$

Il faut maintenant tirer une conclusion. Une différence entre deux échantillons qui correspond à $t_{observé} = -12{,}5$ est-elle un événement rare ou fréquent si les deux échantillons proviennent effectivement de la même population ? Cette différence est-elle statistiquement significative ? Pour répondre à cette dernière question, il faut se référer au tableau des valeurs critiques de t.

La valeur du $t_{critique}$ (t[dl]) et les degrés de liberté

Comme nous l'avons vu à propos du test t pour un seul échantillon, la valeur $t_{critique}$ est celle à laquelle on oppose la statistique $t_{observé}$. *Lorsque le $t_{observé}$ est égal ou supérieur à la valeur $t_{critique}$, nous concluons au rejet de H_0 (l'hypothèse nulle) :* les deux échantillons proviennent de populations différentes avec un risque d'erreur d'inférence égal au seuil α.

Pour trouver la valeur critique pertinente dans le tableau des valeurs critiques, nous devons calculer le nombre de degrés de liberté et décider du seuil alpha. Nous savons que les degrés de liberté se donnent par $N_1 + N_2 - 2$. Dans ce cas, nous avons $N_1 = N_2 = 50$. Le nombre de degrés de liberté est donc $(50 - 1) + (50 - 1) = (50 + 50 - 2) = 98$.

Nous pouvons maintenant trouver la valeur critique du t. Choisissons un seuil $\alpha = 0{,}001$. Au tableau des valeurs critiques dans l'Annexe, nous voyons que pour 98 degrés de liberté (dl = 98) et pour un seuil $\alpha = 0{,}001$, $t_{critique} = 3{,}73$. Nous comparons maintenant le $t_{observé}$ au $t_{critique.}$ Puisque le $t_{observé} = -12{,}5$ est plus extrême que 3,73 (nonobstant le signe), nous concluons au rejet de H_0 car le résultat indique qu'il existe moins d'une chance sur mille ($\alpha = 0{,}001$) qu'une telle différence puisse être observée entre deux échantillons provenant de la même population. Le résultat est significatif avec une probabilité p d'erreur de type I inférieure à 0,001. Nous interprétons ces statistiques en disant que le médicament change significativement le nombre de symptômes de la maladie ($t(98) = -12{,}5$, $p < 0{,}001$).

Quiz rapide 10.4
Nous obtenons un $t_{observé}$ de 10, 74 avec 40 degrés de liberté. Cette différence est-elle statistiquement significative à $\alpha = 0{,}05$, 0,01 et 0,001 respectivement ?

Le signe de la statistique $t_{observé}$

Lorsque nous calculons la statistique $t_{observé}$, elle peut prendre des valeurs positives ou négatives. Ce signe est déterminé par l'ordre dans lequel nous calculons la différence entre les deux moyennes M_1 et M_2. Lorsque la moyenne du groupe 1 est numériquement supérieure à celle du groupe 2, la statistique $t_{observé}$ prendra un signe positif. Si l'inverse est vrai, $M_2 > M_1$,

le signe sera négatif. Puisque nous sommes libres de spécifier l'ordre des calculs, le signe du test t n'a pas de signification particulière. Les valeurs $t_{critique}$ tabulées ne contenant pas de signes, lorsque nous comparons le $t_{critique}$ au $t_{observé}$, nous ignorons le signe de ce dernier.

Hypothèse unicaudale ou hypothèse bicaudale ?

Lorsque nous concevons notre hypothèse, nous devons prendre une décision à son sujet. Proposons-nous une hypothèse *directionnelle* ou une hypothèse *non directionnelle* ? Une hypothèse directionnelle prend le nom technique d'*hypothèse unicaudale* et une hypothèse non directionnelle prend celui d'*hypothèse bicaudale*.

Une *hypothèse non directionnelle* (bicaudale) signifie qu'on cherche à démontrer l'existence d'une différence, peu importe sa direction. Ainsi, dans l'exemple portant sur l'efficacité du programme de formation, n'importe quelle différence significative aurait appuyé notre hypothèse. Les employés pouvaient avoir une productivité moyenne moindre ou supérieure à 100. L'hypothèse non directionnelle dans ce cas est :

> H : $\mu_{avec\ formation} \neq 100$; la performance des personnes formées sera *différente* de celle de la population.

Par contre, l'*hypothèse directionnelle* (unicaudale) indique que l'on veut démontrer que la différence sera dans une seule direction. Dans l'exemple qui porte sur l'efficacité de la formation, nous choisirions fort probablement une hypothèse unicaudale, car il nous importe de savoir si la formation mène à un accroissement de la productivité. Dans ce cas, notre hypothèse ne serait soutenue que si la moyenne pour le groupe ayant reçu la formation était supérieure à celle des travailleurs qui ne l'ont pas reçue. L'hypothèse directionnelle prendrait la forme suivante :

> H : $\mu_{avec\ formation} > 100$; la performance des personnes formées sera *supérieure* à la moyenne de la population.

Dans l'exemple médical, nous comparons deux groupes, avec ou sans traitement de la maladie d'Alzheimer. Nous faisons l'expérience décrite afin de déterminer si le médicament est efficace. Le médicament ne sera efficace que dans un seul cas : lorsque les patients qui le reçoivent ont *moins*

de symptômes que les autres. Si la condition des patients qui reçoivent le médicament s'aggrave, ou elle demeure inchangée, nous ne pouvons pas conclure que le traitement est efficace. En l'occurrence, poser la question « le traitement est-il efficace ? » revient à vérifier si les symptômes de ceux qui sont traités avec le médicament sont amoindris par rapport à ceux qui n'en bénéficient pas. L'hypothèse s'écrit :

$$H : \mu_{\text{avec médicament}} < \mu_{\text{sans médicament.}}$$

Il s'agit d'une *hypothèse directionnelle* puisqu'elle ne sera confirmée que si la différence que nous observons est dans une seule direction : les patients recevant le traitement ont moins de symptômes que ceux qui n'en reçoivent pas. Deux résultats peuvent invalider l'hypothèse directionnelle. D'une part, si les deux moyennes sont statistiquement égales, nous ne pouvons pas rejeter H_0 et nous sommes contraints de conclure qu'il n'y a pas de preuve pour H. Mais nous ne pouvons pas plus rejeter H_0 si le résultat obtenu est l'inverse de notre hypothèse : les patients qui reçoivent le traitement démontrant plus de symptômes.

Lorsque nous présentons une hypothèse directionnelle, nous postulons à l'avance non seulement qu'il existera une différence, mais, plus spécifiquement, quel groupe aura une moyenne supérieure. Advenant une hypothèse directionnelle, nous allons tester la statistique $t_{\text{observé}}$ dans la partie du tableau donnant les valeurs t_{critique} pour un test unicaudal.

Quiz rapide 10.5
Reprenez le résultat obtenu t = –12,5 et comparez-le avec la valeur critique de t unicaudale. Le médicament est-il efficace ?

La valeur critique de t pour les hypothèses unicaudale et bicaudale

Arrêtons-nous au Tableau 10.2, qui est un extrait du tableau des valeurs critiques de t, et suivons une rangée de degrés de liberté à travers toute sa longueur. On remarquera que le t_{critique} augmente (devient plus grand) lorsque l'on passe d'un seuil α de 0,05 à un seuil plus petit (α = 0,01 ou 0,001). Pour que le $t_{\text{observé}}$ soit significatif, il lui faut être égal ou supérieur au t_{critique}. Toutes choses étant égales par ailleurs, la taille du t reflète la taille de la dif-

férence entre les moyennes, par conséquent la différence entre les moyennes doit être plus grande.

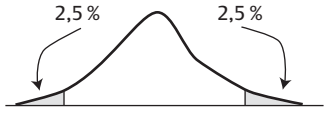
Lorsque nous rejetons H_0 avec un certain α (disons $\alpha = 0{,}05$), cela veut dire en réalité que moins de 5 % des différences entre les échantillons extraits aléatoirement d'une même population auront une différence de moyenne aussi forte que le $t_{observé}$. Examinons cela plus précisément encore en étudiant les graphiques du Tableau 10.2. Lorsque nous faisons un test unicaudal avec un α de 0,05 (disons $H_0: \mu_1 > \mu_2$), nous voulons que moins de 5 % des échantillons donnent un $t_{observé}$ plus grand que le t critique. Puisque le test unicaudal spécifie la direction de la différence, nous n'avons qu'à démontrer que la différence observée ($t_{observé}$) est au bon endroit et notre inférence sera juste. Graphiquement, dans ce tableau, nous n'avons qu'à démontrer que le $t_{observé}$ se situe à l'intérieur de la zone de la distribution des différences grises, en l'occurrence le 5 % supérieur de la distribution des différences.

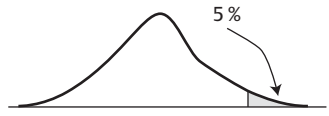
Lorsque nous travaillons avec une hypothèse non directionnelle, il s'ensuit que 5 % des échantillons auront une différence plus grande que celle obtenue dans nos échantillons. Mais nous avons deux façons différentes de conclure au rejet de H_0 : soit que M_1 est plus grand que M_2, soit que M_1 est plus petit que M_2. Le test non directionnel prend cela en considération en divisant le risque d'erreur α en deux, assignant 2,5 % dans la partie supérieure ($\alpha/2$) et 2,5 % dans la partie inférieure ($\alpha/2$) de la distribution des différences. Pour ces raisons, le test directionnel est aussi appelé test unicaudal et le test non directionnel, test bicaudal.

L'utilisation du tableau des valeurs critiques pour les tests unicaudaux et bicaudaux

Selon la forme de l'hypothèse unicaudale ou bicaudale, nous allons utiliser une partie différente du tableau des valeurs critiques de t. Lorsque nous posons une hypothèse bicaudale, nous utilisons la partie gauche du Tableau 10.2. Pour les hypothèses unicaudales, il faut faire appel à la partie droite. Supposons que nous comparons deux petits groupes ($N_1 = N_2 = 4$) ayant les moyennes suivantes : $M_1 = 10$ et $M_2 = 8$, et qui produisent le $t_{observé} = 2{,}0$. Ici, nous avons 6 degrés de liberté. Si notre hypothèse est

Tableau 10.2
Valeurs $t_{critique}$ pour le test t directionnel et non directionnel





Hypothèse bicaudale (non directionnelle)				*Hypothèse unicaudale (directionnelle)*			
	Seuil α				*Seuil α*		
dl	*0,05*	*0,01*	*0,001*	*dl*	*0,05*	*0,01*	*0,001*
1	12,706	63,657	636,62	1	6,314	31,821	318,31
6	2,447	3,707	5,959	6	1,943	3,143	5,208
11	2,201	3,106	4,437	11	1,796	2,718	4,025
16	2,12	2,947	4,073	16	1,746	2,583	3,686
40	2,021	2,704	3,551	40	1,684	2,423	3,307
120	1,980	2,617	3,373	120	1,658	2,358	3,160
∞	1,960	2,576	3,291	∞	1,645	2,326	3,090

bicaudale, nous utilisons la partie gauche du tableau. Nous trouvons, pour dl = 6 et $\alpha = 0,05$, le $t_{critique} = 2,447$. Puisque $t_{observé} = 2,0$ est inférieur au $t_{critique} = 2,447$, nous ne pouvons pas rejeter l'hypothèse nulle. Nous devons conclure que les deux groupes ne proviennent pas de populations différentes, qu'ils ne sont pas statistiquement différents. Mais supposons que l'hypothèse est directionnelle et qu'elle postule que le groupe 1 sera supérieur au groupe 2 ($M_1 = 10 > M_2 = 8$) et que nous trouvons le même $t_{observé} = 2,0$. Nous cherchons alors dans la partie droite du Tableau 10.2, la partie unicaudale, et nous trouvons $t_{critique} = 1,943$. En comparant cette valeur au $t_{observé} = 2,0$, nous voyons que cette dernière est supérieure au $t_{critique}$. Nous concluons maintenant au rejet de l'hypothèse nulle. Ainsi, il est d'une extrême importance de bien choisir la partie du tableau (unicaudale ou bicaudale) qui correspond correctement à la forme de l'hypothèse.

Le seuil α

Le seuil α d'un test t a exactement la même signification que celle que nous avons vue au chapitre 9. Il s'agit du risque de tirer une conclusion fausse (erreur de type I) en rejetant l'hypothèse nulle. En choisissant un α de 0,05, nous acceptons un risque de 5 % de faire une erreur de type I. Avec $\alpha = 0{,}01$, le risque d'erreur α tombe à 1 chance sur 100, et avec $\alpha = 0{,}001$, le risque d'une erreur de type I est de 1 sur 1000. On peut remarquer la différence entre les valeurs critiques de t pour les différents seuils pour les mêmes degrés de liberté. La magnitude du $t_{critique}$ augmente à fur et à mesure que le niveau α passe de 0,05 à 0,001. Cela est raisonnable. Lorsque la différence de moyenne est grande, on a plus confiance qu'il existe une différence sur le plan de la population. Un seuil de signification de 0,001 donne davantage de poids à notre conclusion qu'un seuil de signification de 0,05 (avec $\alpha = 0{,}001$, nous avons 1 chance sur 1 000 de nous tromper en concluant qu'il y a une différence, alors qu'avec $\alpha = 0{,}05$, notre risque d'erreur est de 5 chances sur 100). Cela donne donc plus de poids à notre rejet de H_0 s'il se base sur un seuil α de 0,001 plutôt que sur un seuil α de 0,05. Mais pour obtenir un $t_{observé}$ supérieur au $t_{critique}$ pour $\alpha = 0{,}001$, la différence entre les moyennes doit être plus grande que celle requise pour conclure à la signification statistique avec 5 chances sur 100 de se tromper.

Si on fait passer un test d'arithmétique à un groupe d'élèves du primaire et qu'on compare cette performance avec un groupe de professeurs de mathématiques à l'université, la différence entre les deux groupes sera très grande, ce qui se traduira par une valeur $t_{observé}$ très grande. Notre conclusion, selon laquelle les mathématiciens universitaires sont meilleurs en mathématiques que les enfants de l'élémentaire, aura plus de poids. Si on répète l'expérience en comparant des élèves de 5e année à des élèves de 6e année, la différence sera plus petite et notre conclusion aura moins de poids. La différence pourrait être significative à $\alpha = 0{,}05$, mais pas à $\alpha = 0{,}01$.

Un exemple de test t sur deux groupes indépendants

Supposons que l'on compare deux échantillons dans le but de vérifier s'ils ont des moyennes différentes, s'ils proviennent de populations différentes.

$$H: \mu_1 \neq \mu_2$$
$$H_0: \mu_1 = \mu_2$$

L'hypothèse est non directionnelle (bicaudale). La statistique $t_{observé}$ qui évalue la différence entre les deux moyennes est égale à 3,0. Nous avons 21 observations dans chaque groupe. Le nombre de degrés de liberté est 21 + 21 – 2 = 40. Nous désirons tester notre hypothèse avec un risque d'erreur α inférieur à 0,01 (nous rejetterons H_0 seulement si le risque d'erreur est plus petit que 1 %). Dans l'Annexe, pour dl = 40 et $\alpha = 0{,}01$, la valeur critique $t_{critique} = 2{,}704$ (bicaudale). Le $t_{observé} = 3{,}0$ étant plus grand que le $t(40) = 2{,}704$, nous concluons que la différence entre les deux groupes est statistiquement significative, et il y a moins de 1 chance sur 100 ($p < 0{,}01$) que cette conclusion soit fausse. Nous écrivons : « La différence entre les deux échantillons est significative ($t(40) = 3{,}0$, $p < 0{,}01$). »

Quiz rapide 10.6
Refaites le problème ci-dessus, mais cette fois, testez le $t_{observé}$ en utilisant une hypothèse directionnelle. La conclusion (H ou H_0) change-t-elle ?

LE TEST T POUR DES DONNÉES PAIRÉES

La dernière utilisation du test t concerne les données pairées ou jumelées. Cette application du test t est particulièrement utile lorsqu'il s'agit de mesurer le changement. On prend ainsi des mesures avant et après un événement, et on évalue si les deux moyennes, avant et après, sont les mêmes statistiquement. Par exemple, nous pourrions évaluer si l'introduction d'un programme d'accès à des aiguilles sanitaires réduit le taux d'hépatite chez les héroïnomanes. On pourrait alors mesurer l'incidence d'hépatites dans la population d'héroïnomanes avant et après l'introduction du programme.

Il est important que ce soit les mêmes sujets (par exemple les mêmes personnes) qui soient mesurés deux fois ; si une personne ne peut pas être mesurée au second passage, il faut retirer sa première mesure de

l'échantillon. Par conséquent le nombre d'observations pour la mesure pré-intervention est invariablement égal au nombre d'observations de la mesure post-intervention. Donc, il faut toujours s'assurer que $N_1 = N_2$.

Supposons que nous voulons étudier l'impact d'une intervention sur un groupe de personnes. Nous avons donc, pour chaque personne, deux informations : sa performance avant et après l'intervention. Si le traitement n'a aucun effet, chaque personne produira la même performance sur les deux mesures. De manière équivalente, nous dirons que la différence entre les performances de la même personne aux deux prises d'information est égale à zéro. Pour chaque personne, nous aurons donc une performance pré-intervention (symbolisée par X_i) et une performance évaluée après l'intervention (Y_i). Nous pouvons calculer la différence entre les deux informations pour chaque personne. On peut alors écrire : $D_i = (X_i - Y_i)$, qui représente la différence D, pour chaque personne. On calcule ensuite la moyenne de cette valeur D que nous appelons M_D. Si cette valeur est égale à zéro, on ne peut pas conclure que le traitement a eu un effet. Si la moyenne des différences D n'est pas égale à zéro, on peut alors (potentiellement) conclure que le traitement a eu un effet, qu'il existe une différence entre avant et après. Pour que cela soit potentiellement vrai, il faut que la différence moyenne (M_D) ne soit pas égale à zéro. Ainsi, le jeu d'hypothèses pourrait prendre la forme suivante :

$$H : \Delta \neq 0$$
$$H_0 : \Delta = 0$$

où Δ (delta) représente la vraie différence en ce qui a trait à la population entière. Il s'agit maintenant de faire un test t sur la différence moyenne de l'échantillon. Cette hypothèse, comme d'habitude, peut être formulée de manière directionnelle ou non directionnelle.

En fait, nous avons créé, à partir des deux valeurs (avant et après), une nouvelle variable, D, qui reflète la différence de performance pour chaque personne. La moyenne de cette variable D est M_D. Mais parce que D est une variable, comme toutes les variables, on peut calculer son écart-type s_D.

À partir de l'écart-type s_D, il devient possible de calculer l'erreur type s_{MD} à l'aide de la formule habituelle ($s/\sqrt{N}$) qui, dans ce cas, devient :

$$s_{M_D} = \frac{s_D}{\sqrt{N}}$$

Puisque l'hypothèse nulle postule que la différence moyenne entre les valeurs avant et après sera égale à zéro, nous connaissons maintenant la valeur présumée de la moyenne de la population ($\Delta = 0$).

On peut maintenant construire l'intervalle de confiance avec la formule de calcul pour un unique groupe avec variance inconnue : $\Delta \pm s_{M_D} \times t(dl)$. Puisqu'on postule que la moyenne dans la population est 0, la formule se simplifie pour devenir $\Delta \pm s_{M_D} \times t(dl) = 0 \pm s_{M_D} \times t(dl) = \pm s_{M_D} \times t(dl)$. Si la différence moyenne observée (M_D) se situe à l'extérieur de l'intervalle de confiance, nous concluons au rejet de H_0.

On peut aussi établir la statistique $t_{observé}$ directement en utilisant la forme suivante :

$$t_{observé} = M_D / s_{M_D}$$

où M_D est la différence moyenne entre les observations avant-après et s_{M_D} est l'erreur type de cette différence. Il ne reste alors qu'à comparer le $t_{observé}$ au $t_{critique}$ pour conclure. Encore une fois, et selon la teneur de l'hypothèse (uni ou bicaudale), il faudra faire appel à la bonne colonne du tableau des valeurs critiques pour repérer le $t_{critique}$.

Les degrés de liberté dans le test t pour échantillons pairés

Pour analyser la différence entre les deux mesures de chaque observation, on a créé une nouvelle variable, D. Nous calculons la moyenne et l'écart-type de cette variable D. Puisque nous n'avons qu'une seule moyenne et qu'un seul écart-type, nous ne perdons qu'un seul degré de liberté. Par conséquent, le nombre de degrés de liberté pour cette forme du test t est $N - 1$.

Une illustration du test t pour échantillons pairés

Prenons pour illustration un programme de relaxation par le yoga visant à réduire le stress ressenti au travail chez les cadres supérieurs. On mesure le stress au travail avec un questionnaire dans lequel un score élevé indique un degré de stress élevé. Nous postulons que le programme de relaxation réduit le stress ressenti par les cadres supérieurs, ce qui donne :

$$H_0 : \Delta = 0$$
$$H : \Delta > 0$$

où Δ est la moyenne des différences dans la population. Il s'agit d'une hypothèse directionnelle. Nous adoptons le seuil α de 0,05.

Pour tester notre hypothèse, nous mesurons un échantillon de 25 personnes avant le début du cours de yoga et trois mois après la fin du cours. Pour chacune, nous calculons le degré de stress ressenti avant le cours moins le score obtenu après. La différence moyenne M_D est de 12 (le degré de stress après l'intervention est moins élevé qu'avant). Est-ce une amélioration notable ? Nous calculons l'écart-type de la différence et nous trouvons que $s_D = 20$.

Nous calculons l'erreur type de D (s_{M_D} en utilisant la formule habituelle)

$$s_{M_D} = \frac{s_D}{\sqrt{N}} = 20/5 = 4.$$

La valeur critique t(dl) est trouvée dans la table t avec 24 degrés de liberté et pour $\alpha = 0{,}05$ (unicaudal). La valeur critique du t(dl) est 1,71.

Le test statistique de la différence (pas égale à zéro) est de la forme :

Rejet de H_0 si M_D est plus grand que la limite supérieure de l'intervalle de confiance $s_{M_D} \times$ t(dl).

Ou de façon équivalente :

$$\text{Rejet de } H_0 \text{ si } \frac{M_D}{s_{M_D}} > \text{t(dl)}.$$

Nous trouvons :

$$M_D / s_{M_D} = 12/4 = 3.$$

La valeur obtenue étant plus grande que la valeur critique t(dl), nous rejetons l'hypothèse nulle et nous écrivons : « Les cours de yoga ont diminué significativement le stress des cadres supérieurs [$t(24) = 3$, $p < 0{,}05$]. »

SOMMAIRE DES ÉTAPES POUR RÉALISER UN TEST T

1. Poser les hypothèses ; décider si elles sont directionnelles ou non directionnelles.

2. Choisir le seuil de confiance α.
3. Décider de la forme du test (un groupe, deux groupes, échantillons pairés).
   a) Calculer la statistique $t_{observé}$.
   b) Calculer les degrés de liberté et trouver t(dl) (le $t_{critique}$) dans le tableau des valeurs critiques de t (Annexe) en fonction des dl, de α et selon que le test est unicaudal ou bicaudal.
4. Conclure.
   a) Si le $t_{observé}$ est égal ou plus grand que la valeur critique t(dl), conclure que la différence est statistiquement significative au niveau α choisi et rejeter H_0.
   b) Si le $t_{observé}$ est plus petit que la valeur critique t(dl), conclure que la différence n'est pas statistiquement significative au niveau α choisi et ne pas rejeter H_0.

RÉDIGER UNE INTERPRÉTATION DES DONNÉES

La rédaction d'une interprétation des résultats n'est pas chose aisée. D'un côté, un travail important de statistique a été réalisé. Or, le lecteur de la recherche n'est pas nécessairement un statisticien. On doit donc lui expliquer les résultats en termes accessibles et significatifs pour lui. Il est probable que H_0, μ, M, etc., ne feront qu'égarer le lectorat. D'un autre côté, pour des raisons de crédibilité, on ne peut pas faire d'affirmations gratuites. Chaque fois qu'on vous rapporte une différence ou un effet, on doit mettre dans un rapport des signes linguistiques qui disent en substance : « Je n'affirme pas cela gratuitement, j'ai posé mes hypothèses et fait le test statistique approprié, et l'effet est significatif, ou ne l'est pas. »

Dans à peu près toutes les disciplines scientifiques, il y a : 1) l'utilisation du mot « significatif » ; 2) l'inclusion du résultat du test entre parenthèses, suivi du seuil α selon cette écriture très stricte : « (nom-de-la-stat [degrés de liberté, s'il y a lieu] = résultat, $p < \text{seuil } \alpha$) » si le test est significatif. Par exemple, un résultat statistiquement significatif, à la suite d'un test t, serait présenté de la manière suivante : $t(12) = 10{,}45$, $p < 0{,}01$. Le signe plus petit (<) signifie que la probabilité d'obtenir le résultat par pur hasard est plus petite que α, ce qui veut dire qu'on a rejeté H_0. Si le test n'est pas signifi-

catif, il faut aussi rapporter la statistique, mais cette fois, « p > seuil α ». Par exemple, nous pourrions écrire t(12) = 1,45, p > 0,05. Le signe plus grand (>) signifie que la probabilité d'obtenir ce résultat par pur hasard est plus grande que α, ce qui veut dire qu'on n'a pas rejeté H_0.

Voici un exemple tiré d'un rapport de recherche scientifique.

Interprétation des résultats

Pour les 135 personnes composant notre échantillon, nous trouvons une amélioration significative à la suite de la thérapie (t(134) = 6,4, p < ,05).

Comme on peut le voir, à part l'utilisation du mot « significative » et la présence de codes dans les parenthèses, il n'y a pas de jargon statistique (« hypothèse », « population », « μ », etc.).

Quiz rapide 10.7

Pouvez-vous dire quel test statistique a été fait dans l'exemple précédent ? Pouvez-vous dire quel risque le chercheur était prêt à prendre quand il a écrit sa conclusion ? Croyez-vous que s'il avait été prêt à prendre un risque plus faible (disons un sur mille), la conclusion aurait tenu la route ?

SOMMAIRE DU CHAPITRE

La statistique, la distribution et le test t sont tous attribuables à W. S. Gosset. Ce test a été développé spécifiquement pour être utilisé avec de petits échantillons, généralement définis comme étant inférieurs à N = 30. Ce test statistique sert à déterminer si un échantillon a de fortes ou de faibles chances d'appartenir à une population en particulier, ou à déterminer si deux échantillons appartiennent à la même population ou à des populations différentes. La statistique t, comparant deux groupes indépendants, est le rapport entre la différence qui existe entre ces deux moyennes et la différence qui existe entre deux moyennes aléatoirement extraites de la même population. L'interprétation de la statistique t se fait en la comparant avec une valeur standard, le $t_{critique}$. Cette valeur standard est tabulée. L'utilisation des valeurs tabulées est différente selon que l'hypothèse est directionnelle ou non directionnelle. Lorsque le t obtenu est numériquement

égal ou supérieur au $t_{critique}$, nous pouvons conclure que les deux groupes ne proviennent pas de la même population.

EXERCICES DE COMPRÉHENSION

1. Étant donné deux groupes indépendants avec respectivement 12 et 10 données brutes, quel est le degré de liberté pour réaliser un test t comparant la moyenne des deux échantillons ?
   a) 22
   b) 21
   c) 20
   d) un autre nombre
2. Nous testons la différence entre deux échantillons, A et B, et la différence entre deux autres échantillons, C et D. Les échantillons A, B, C et D sont tous de la même taille N et ils ont tous la même variance. La différence entre A et B est statistiquement significative seulement à $\alpha = 0{,}05$, alors que la différence entre C et D est statistiquement différente à $\alpha = 0{,}01$. La différence entre les moyennes des groupes A et B est _____ que la différence entre les moyennes des groupes C et D.
   a) plus grande
   b) plus petite
   c) de la même taille
   d) Toutes ces réponses sont possibles.
3. Nous postulons que les hommes sont moins consciencieux au travail que ne le sont les femmes. Pour tester notre hypothèse, nous choisissons aléatoirement un groupe d'hommes et un groupe de femmes, et nous mesurons leur degré de concentration au travail. Dans ce cas, il _____.
   a) nous faudra faire appel à un test statistique uni ou bicaudal selon l'erreur type
   b) nous faudra faire appel à un test statistique unicaudal
   c) nous faudra faire appel à un test statistique bicaudal
   d) n'est pas possible de faire un test statistique

4. Nous voulons examiner si l'étude de ce volume a un impact sur la compréhension que les étudiants ont de la statistique. Nous administrons un test de statistique aux 12 personnes qui suivent le cours le premier jour de classe et nous l'administrons à nouveau le dernier jour de classe. Il nous faudra alors tester _____ en faisant appel au test _____ et les degrés de liberté seront de _____.

a) la différence entre les moyennes ; t ; 11
b) la différence entre les moyennes ; t ; 10
c) la différence entre les variances ; Z ; 11
d) la différence entre les variances ; Z ; 12

5. Dans la population, nous savons que le salaire moyen est de 30 000 $. Nous examinons un échantillon de forgerons pour trouver qu'en moyenne ils gagnent 35 000 $ et que l'intervalle de confiance autour de cette moyenne est de 2 000 $. Laquelle des conclusions suivantes est juste ?

a) Les forgerons sont, en général, mieux payés que la moyenne des gens.
b) Les forgerons sont, en général, payés autant que la moyenne des gens.
c) Tous les forgerons gagnent plus que 30 000 $.
d) Compte tenu des informations disponibles, toutes ces réponses sont possiblement justes.

6. Une compagnie pharmaceutique en est au début du processus d'évaluation d'un médicament. Elle compare deux groupes, l'un recevant un médicament, l'autre non. La compagnie désire minimiser l'erreur de type II. Par conséquent, elle choisit de comparer de _____ groupes, elle compare les moyennes avec la statistique t et elle fait appel à un seuil de signification plus _____.

a) grands ; petit
b) petits ; petit
c) petits ; grand
d) grands ; grand

7. Voici les résultats des tests t exécutés pour chacune des trois études suivantes. Dans chaque cas, il s'agit d'études qui comparent des groupes indépendants et, dans tous les cas, l'hypothèse faite est non directionnelle avec $\alpha = 0{,}05$. Il faut indiquer, pour chaque résultat, si les deux échantillons proviennent d'une ou de deux populations.
   Étude A : t = 2,58 ; N = 7 ; Étude B : t = 2,1 ; N = 22 ;
   Étude C : t = 1,99 ; N = 62.
8. En y pensant bien, le test t pour les échantillons indépendants compare la différence entre les moyennes de deux groupes à la différence à laquelle nous pourrions nous attendre entre deux groupes extraits de la même population. Cette phrase est-elle vraie ou fausse ?
9. Pour le contraste entre les moyennes de deux groupes, on pourrait faire appel au test t ou au test Z. Lorsque le N est _____, nous devons faire appel au test t, alors que lorsque nous avons au moins _____ observations, la distribution des valeurs t et des valeurs Z est _____.
   a) grand ; 30 ; identique
   b) petit ; 1 000 ; identique
   c) grand ; 30 ; très différente
   d) petit ; 120 ; identique

Réponses

1. c
2. b
3. b
4. a
5. a
6. d
7. Étude A = 2 ; Étude B = 2 ; Étude C = 1
8. Vraie
9. d

CHAPITRE 11
L'ANALYSE DE VARIANCE À UN FACTEUR

CHAPITRE 11

L'ANALYSE DE VARIANCE À UN FACTEUR

Le test t est utile lorsque l'objectif est de comparer deux conditions expérimentales (avant et après) ou deux groupes (les hommes et les femmes). Il est spécialement conçu pour conclure si deux échantillons de petite taille proviennent ou non de la même population. Avec le test t, nous calculons la statistique t qui standardise la différence entre les moyennes des deux groupes et nous vérifions, à partir d'une distribution des valeurs t, la probabilité d'obtenir une telle valeur t lorsque deux échantillons proviennent de la même population. Lorsque cette probabilité est faible, souvent $p < 0,05$ ou moins, nous inférons que les deux échantillons ne proviennent pas de la même population, en sachant que nous courons le risque (de 5 %) que cette conclusion soit erronée. Lorsque nous voulons minimiser la probabilité de commettre une telle erreur d'inférence, nous choisissons un seuil de signification statistique alpha qui est plus petit, plus conservateur ($p < 0,01$ ou même $p < 0,001$).

L'analyse de variance simple, l'ANOVA (de l'anglais *ANalysis Of VAriance*), est un test statistique qui généralise le test t. Elle permet l'analyse des différences entre deux groupes ou plus de toute taille. L'ANOVA fait partie des tests statistiques les plus utilisés aujourd'hui et sa compréhension est nécessaire pour l'interprétation de la plupart des textes scientifiques ou pour lire les résultats des études évaluatives. L'ANOVA fonctionne de la même façon que le test t. On calcule une nouvelle statistique, la statistique F, qui standardise la différence entre les moyennes de plusieurs groupes. On recherche la valeur F observée dans un tableau de distribution de la

statistique F afin de déterminer la probabilité d'obtenir le F observé lorsque les groupes proviennent de la même population. Lorsque cette probabilité est faible ($p < 0,05$), on conclut que les groupes ne proviennent pas d'une seule population, c'est-à-dire qu'au moins un des groupes provient d'une population différente. Comme avec le test t, les conclusions tirées avec la statistique F comprennent un risque d'erreur, défini par le seuil alpha.

L'UTILISATION DE L'ANOVA

Nous avons souvent utilisé des exemples où nous comparons un échantillon à un autre. En réalité, la majorité des études comparent plusieurs groupes, pas seulement deux. En recherche pharmaceutique par exemple, il est habituel de comparer trois groupes de patients. Un groupe reçoit le médicament, un deuxième reçoit un médicament concurrent ou un placebo et un dernier groupe ne reçoit ni le médicament ni le placebo. Enfin, on mesure le niveau de guérison pour les patients dans chacun des trois groupes et le test statistique compare l'efficacité moyenne enregistrée dans chaque groupe. L'ANOVA est la technique statistique spécifiquement conçue pour faire ces comparaisons. Il n'y a pas de limites techniques au nombre de groupes pouvant être simultanément comparés par l'ANOVA et, par conséquent, son utilisation est fort répandue.

En psychologie clinique, un psychiatre pourrait choisir de traiter la dépression avec l'une ou l'autre de ces quatre approches thérapeutiques : comportementale, psychanalytique, cognitive et chimique. Sont-elles toutes également efficaces ? Pour répondre à cette question, on répartit aléatoirement des patients en quatre groupes et on administre un traitement différent à chacun des groupes. Après la thérapie, on mesure le degré de dépression de chaque patient dans chaque groupe. Si, à la suite du traitement, le niveau de dépression moyen pour chaque groupe de patients est le même, on ne peut pas conclure que les diverses thérapies sont *inégalement* efficaces (ou inefficaces). Si le niveau de dépression moyen entre les groupes est différent, on conclut que les thérapies ne sont pas également efficaces : les différents traitements produisent différentes « populations de réduction de la dépression ».

Ce que l'ANOVA dit

Comme pour le test t, le test d'ANOVA est utilisé afin de conclure si oui ou non les groupes appartiennent à la même population. L'hypothèse nulle veut que tous les groupes soient identiques, qu'ils proviennent tous de la même population. Lorsqu'on rejette l'hypothèse nulle, on conclut que les groupes ne proviennent pas de la même population. Comme avec le test t, le concept de la signification statistique est central pour l'ANOVA, et fait appel à un tableau de valeurs critiques. Dans le cas de l'ANOVA, il s'agit du tableau des valeurs critiques de la statistique F. Lorsque la différence entre les groupes est statistiquement significative, on sait qu'*au moins un des groupes* provient d'une population différente des autres; que tous les groupes ne proviennent pas de la même population.

Ce que l'ANOVA ne dit pas

Les résultats qu'une ANOVA produit permettent de déterminer si les groupes proviennent ou non de la même population, mais ils ne peuvent pas indiquer où ces différences se situent. Ainsi, si on compare 15 groupes et que l'ANOVA confirme qu'ils ne proviennent pas de la même population, on ne saura toujours pas si cette différence provient d'un seul ou de plusieurs groupes. De plus, l'ANOVA n'identifie pas le ou les groupes qui sont différents. Pour cela, nous verrons les tests de comparaison multiple plus loin. En outre, les résultats produits par l'ANOVA ne nous disent pas, même lorsque la différence est statistiquement significative, si la différence est grande ou petite. Pour cela, il faudra faire appel à une technique statistique particulière — la taille de l'effet — qui est discutée plus loin dans ce chapitre.

POURQUOI L'ANOVA ET PAS LE TEST T ?

Supposons que nous avons trois groupes, groupe 1, groupe 2 et groupe 3, et que nous voulons décider si les groupes diffèrent (proviennent de populations différentes). À première vue, nous pourrions les comparer deux par deux en faisant appel au test t pour deux échantillons indépendants. Un

premier test t comparerait le groupe 1 au groupe 2, un autre, les groupes 2 et 3 et un dernier, la différence entre les groupes 1 et 3. Si nous avons cinq groupes, nous pourrions comparer chaque paire de groupes avec le test t (1 vs 2, 1 vs 3, 1 vs 4, 1 vs 5, 2 vs 3, 2 vs 4, 2 vs 5, 3 vs 4, 3 vs 5 et 4 vs 5). Malheureusement, cette approche est sous-optimale et il ne faut pas y faire appel. D'une part, l'utilisation d'une multitude de tests t n'est pas une technique pratique mais, plus important encore, une telle tactique cause un problème important : *le cumul des risques de l'erreur de type I* qui, lui, produira presque certainement une erreur d'inférence.

Les tests t multiples : une stratégie peu pratique

L'utilisation des tests t répétés amplifie le nombre de calculs requis. Lorsque nous avons deux groupes, une seule comparaison (et donc un seul test t) est requise (1 vs 2). Lorsque nous avons trois groupes, trois comparaisons sont requises (1 vs 2, 2 vs 3 et 1 vs 3). Le problème pratique ne devrait pas sembler si énorme. Cependant, avec 5 groupes, il faut faire 10 comparaisons et donc exécuter 10 tests t. Avec 10 groupes, on a besoin d'en faire 45, ce qui commence à friser l'absurde. Avec cette approche, le nombre de tests à effectuer devient rapidement excessif.

À partir du nombre de groupes K, il est possible de calculer le nombre de paires de comparaisons c, et par conséquent le nombre de tests t qu'il faudrait faire. La Formule 11.1 nous indique comment le faire.

$$c = K\,(K - 1)/2 \qquad \text{Formule 11.1}$$

En appliquant la Formule 11.1 à 10 groupes, le nombre de test t requis est $c = (10 \times 9)/2 = 45$. Pour 20 groupes, il faut exécuter 190 tests t !

Quiz rapide 11.1
Vous avez 25 groupes. Combien de tests t faut-il faire si vous voulez vérifier la différence entre chaque paire de groupes ?
À une soirée entre amis, vous faites un toast en l'honneur de l'hôtesse. S'il y a 10 personnes autour de la table et que chacun porte un toast avec tous les autres convives, combien de tintements de verre entendrez-vous ?

Les tests t multiples : une stratégie qui cumule les risques d'une erreur de type I (α)

La deuxième raison de faire appel à l'ANOVA plutôt qu'au test t lorsqu'on veut comparer plus de deux groupes est plus subtile mais encore plus importante. Une telle utilisation du test t induit une distorsion dans l'inférence, ce qui mènera très probablement à une fausse conclusion.

Pour comprendre le problème, il faut revenir sur le concept de l'erreur de type I. On se souvient qu'à chaque fois que nous concluons, à la suite d'un test t, que les groupes sont statistiquement différents, nous courons le risque de commettre une erreur de type I. La taille de ce risque est déterminée par le seuil alpha choisi (5 % de risque d'erreur lorsque $\alpha = 0{,}05$, 1 % lorsque $\alpha = 0{,}01$, etc.). C'est le cas lorsqu'on fait une comparaison entre deux groupes. Si on a trois groupes, il faut faire trois comparaisons. La probabilité *qu'au moins une* de ces conclusions soit fausse n'est plus de 5 % ; le risque de commettre au moins une erreur de type I est plus élevée.

Pour mieux saisir ce concept, voyons une analogie. Supposons qu'un oracle a une chance sur deux de commettre une erreur en devinant la prochaine carte qu'il va tirer. Avec deux cartes, il a aussi une chance sur deux de se tromper sur chaque carte. Cependant, le risque qu'il commette *au moins* une erreur sur les deux cartes est plus grand (75 %). Avec 10 cartes, il est quasi certain de s'être trompé au moins une fois (le risque est de 99,9 %).

De la même manière, si avec une comparaison (un test t), on a 1 chance sur 20 ($p < 0{,}05$) de commettre une erreur en concluant au rejet de H_0, le risque de commettre au moins une erreur de type I est de $p = 0{,}14$ avec 3 comparaisons, alors qu'avec 45 comparaisons (10 groupes), le risque de commettre au moins une erreur de type I s'élève à 0,90 ! Nous allons, fort probablement, commettre au moins une erreur de type I : conclure à une différence qui n'existe pas.

Il est possible de calculer la probabilité d'une erreur d'inférence avec la Formule 11.2 :

$$p = 1 - (1 - \alpha)^c \qquad \text{Formule 11.2}$$

où p est la probabilité de commettre au moins une erreur de type I, α est le niveau de signification choisi pour les tests t individuels et c est le nombre de comparaisons à effectuer.

Supposons que nous voulons comparer trois groupes ($K = 3$). Pour établir le cumul des risques d'erreur de type I, nous calculons d'abord le nombre de comparaisons requises avec la Formule 11.1.

$$c = K\,(K - 1)/2 = 3(2)/2 = 3$$

Ensuite nous calculons le cumul des risques d'erreur de type I avec la Formule 11.2 pour le seuil $\alpha = 0{,}05$.

$$p = 1 - (1 - \alpha)^c = 1 - (1 - 0{,}05)^3 = 1 - (0{,}95)^3 = 1 - 0{,}857 = 0{,}14$$

Alors que nous pensions conclure qu'il y aurait un risque d'erreur de 5 %, la Formule 11.2 indique que le véritable risque *d'au moins une erreur de type I* est de 14 %, soit presque le triple ! Lorsqu'on compare 15 groupes, deux à la fois, avec α=0,05, nous devons faire 105 comparaisons [(15 X 14)/2 = 105] et le cumul des risques d'erreur devient astronomique : $(1-(1-0{,}05)^{15}$= p = 0,99 ! Il y a 99 chances sur 100 qu'au moins une des différences jugées statistiquement significatives soit erronée. Et, bien entendu, nous ne savons pas laquelle !

Le Tableau 11.1 présente le cumul de l'erreur de type I pour différents nombres de groupes lorsque α est égal à 0,05 et 0,01.

Tableau 11.1
Cumul de l'erreur de type I (p) pour différents nombres de groupes avec $\alpha = 0{,}05$ et $\alpha = 0{,}01$

Nombre de groupes	*c*	*p (α = 0,05)*	*p (α = 0,01)*
2	1	0,05	0,01
3	3	0,14	0,03
4	6	0,26	0,06
5	10	0,40	0,10
7	21	0,66	0,19
10	45	0,90	0,36
15	105	0,99	0,65
20	190	0,99	0,85

Le risque de conclure à tort qu'au moins une des différences est statistiquement significative augmente lorsque le nombre de comparaisons augmente.

Lorsqu'on utilise un seuil α plus conservateur (0,01 plutôt que 0,05), le risque d'en arriver à au moins une fausse conclusion se réduit bien qu'il demeure souvent très élevé (par exemple $p = 0{,}65$ pour 15 groupes) et, de toute façon, il dépasse de beaucoup le seuil conventionnel de la signification statistique ($p < 0{,}05$), qui est généralement défini comme étant minimal pour justifier le rejet de l'hypothèse nulle.

Néanmoins, cela ouvre la porte à une solution potentielle au problème du cumul de l'erreur d'inférence. En réduisant le seuil alpha pour chacune, lorsque le nombre de comparaisons augmente, nous pouvons réduire le risque du cumul de l'erreur.

Supposons que l'on désire comparer 15 groupes, ce qui nécessite 105 tests t. En appliquant la Formule 11.2, nous trouvons que si nous choisissons $\alpha = 0{,}0001$, le risque cumulé de conclure à tort qu'une paire de groupes diffère est $p < 0{,}04$. Cela indique que nous courons un risque de moins de 5 % de nous tromper, au moins une fois, lorsque nous concluons qu'un test t est significatif.

Bien que cette stratégie soit en mesure de contrôler l'erreur cumulée — et donc l'erreur de type I —, elle cause un effet secondaire fort nocif ! Elle augmente le risque d'une erreur de type II. Peu de tests t seront significatifs lorsque le niveau alpha est minuscule : seuls les comparaisons entre les groupes ayant d'immenses différences de moyennes seront significatifs. D'autres différences, plus modestes mais néanmoins importantes, ne seront pas détectées. Conclure qu'il n'y a pas de différence, alors qu'en réalité il en existe une, représente une erreur d'inférence de type II. Lorsque plusieurs comparaisons sont requises, il faudrait utiliser une autre procédure statistique que le test t.

L'alternative — l'analyse de variance — est la technique requise. Elle est spécifiquement conçue pour éviter le problème du cumul des risques d'erreur de type I, lorsque l'on compare plus de deux groupes, sans accroître le risque d'une erreur de type II.

Quiz rapide 11.2

A. Si vous ne faites qu'un test t, est-ce qu'il y a cumul des risques d'erreur de type I ? Pouvez-vous le confirmer avec la Formule 11.2 ?
B. En vous référant au Tableau 11.1, si vous choisissez de contrôler le cumul de l'erreur à $p < ,05$, quel est le nombre maximum de comparaisons qui peut validement être fait lorsque le seuil de signification pour chacune est fixé à $p < ,01$?

LA VARIABLE INDÉPENDANTE ET LA VARIABLE DÉPENDANTE POUR L'ANOVA

L'ANOVA est une série de procédures statistiques qui comparent la moyenne de la variable dépendante pour chaque « niveau » de la variable indépendante. La *variable indépendante* représente la caractéristique qui distingue les groupes, alors que les *niveaux* définissent chacun des groupes qui vont être comparés. En voici quelques illustrations dans différents contextes.

- Un chercheur s'intéresse au comportement des rats dans un labyrinthe lorsque différentes intensités de chocs électriques leur sont administrées à la suite des erreurs qu'ils font. Un groupe de rats reçoit des chocs de 5 mA (milliampères), un autre groupe est exposé à des chocs de 10 mA et un dernier reçoit une intensité de 15 mA. La variable indépendante est donc « l'intensité du choc électrique », laquelle est de trois niveaux (5, 10 et 15 mA), chacun administré à un groupe distinct.
- La ministre de l'Éducation désire savoir si le niveau d'éducation a un impact sur le salaire des citoyens. Elle compare alors quatre groupes de citoyens qui définissent la variable indépendante, l'éducation, en quatre niveaux. Un groupe est composé de personnes ayant obtenu un diplôme d'études secondaires, un deuxième ayant un baccalauréat, un troisième détenant une maîtrise et un dernier composé de personnes ayant un doctorat. La variable indépendante est « le niveau de scolarité » composée de quatre groupes ou « niveaux ».
- La vice-présidente des ressources humaines dans une entreprise s'intéresse à l'impact du type de rémunération sur le rendement des employés. Elle sélectionne alors trois groupes d'employés : un groupe payé à la commission, un autre touchant un salaire annuel et le dernier

payé à l'heure. La variable indépendante est « la structure de la rémunération » qui, dans ce cas, comprend trois niveaux.

La variable indépendante (intensité des chocs électriques, scolarité, structure salariale) est divisée en niveaux. Il n'y a pas de limite au nombre de niveaux d'une variable indépendante qui peuvent être comparés par ANOVA.

Quiz rapide 11.3
Dans l'exemple des thérapies pour la dépression, décrit en début de chapitre, quelle est la variable indépendante ? Quels sont les niveaux pour cette variable ?

La *variable dépendante* est la variable qui est mesurée. On a besoin de la valeur obtenue par la variable dépendante pour chaque observation de chaque groupe. Le chercheur qui s'intéresse au comportement des rats par rapport aux chocs électriques compterait le nombre d'erreurs faites par chacun des rats dans chacun des groupes. La variable dépendante est « le nombre d'erreurs ». La ministre de l'Éducation pourrait choisir le salaire de chaque personne dans chaque groupe comme variable dépendante. La vice-présidente des ressources humaines devra mesurer le niveau de rendement de chaque employé dans chacun des groupes de rémunération.

Quiz rapide 11.4
Dans l'exemple des thérapies contre la dépression, quelle pourrait être la variable dépendante ?

LE PRINCIPE FONDATEUR DE L'ANALYSE DE VARIANCE : LES DIFFÉRENCES INTERGROUPES ET INTRAGROUPES

La logique de l'ANOVA suit étroitement celle du test t, et les tests d'hypothèses sont eux aussi analogues à tous ceux que nous avons déjà étudiés. Les termes qui définissent l'ANOVA semblent à première vue assez différents de ceux qu'on utilise pour le test t, et les formules spécifiques le sont aussi, mais les principes sont dans l'ensemble très similaires.

Commençons par un exemple hypothétique. Nous avons trois groupes d'élèves de troisième secondaire qui font leurs études dans trois pays hypo-

thétiques, Pays 1, Pays 2 et Pays 3. Chaque groupe est composé de cinq élèves aléatoirement choisis parmi la population des élèves ayant des résultats scolaires équivalents dans leur pays. On fait passer le même test d'algèbre aux trois groupes d'élèves. La variable indépendante est le pays d'origine ; cette variable a trois niveaux (Pays 1, Pays 2, Pays 3). La variable dépendante est la performance à l'examen d'algèbre notée sur 100 pour chaque élève.

Nous voulons déterminer si les élèves des trois pays ont des performances différentes ou non. Les hypothèses sont :

$H_{0:}$ La performance en algèbre n'est pas différente pour les élèves des trois pays (les élèves des trois pays proviennent tous de la même population de performance en algèbre).

H : La performance en algèbre est différente pour les élèves des trois pays (les élèves des trois pays ne proviennent pas tous de la même population de performance en algèbre).

Ce qui devient formellement :

$H_{0:}$ $\mu_1 = \mu_2 = \mu_3$ (il n'y a aucune différence entre les groupes).

H : $\mu_i \neq \mu_j$ pour au moins une paire de moyennes.

Le Tableau 11.2 montre les résultats obtenus par cinq élèves dans chacun des pays (dans une vraie étude, on mesurerait plus de cinq élèves).

La meilleure estimation que nous avons de la performance des élèves dans chaque pays est la moyenne obtenue dans chaque groupe : $M_1 = 40$; $M_2 = 50$ et $M_3 = 60$. Les trois moyennes ne sont pas numériquement identiques. Il existe une variabilité entre les groupes, variabilité qu'il nous faudra quantifier. Nous donnons le nom de *différence intergroupe* à cette quantité.

Tableau 11.2
Performance en algèbre pour les élèves dans 3 pays

Élève	*Pays 1*	*Pays 2*	*Pays 3*
1	30	40	50
2	35	45	55
3	40	50	60
4	45	55	65
5	50	60	70
Moyenne	**40**	**50**	**60**

Mais est-ce que la présence de différences entre les groupes indique nécessairement qu'il existe plus d'une population de performance en algèbre (rejet de H_0)? En examinant cette différence intergroupe isolément, nous ne pouvons pas répondre à la question. Après tout, plusieurs élèves provenant de pays différents obtiennent la même note et, entre autres, l'élève 5 du pays 1 obtient une note supérieure ou égale aux élèves 1, 2 et 3 du pays 2. Il se pourrait que la différence entre les élèves d'un même pays soit aussi grande que celle existant entre les élèves des différents pays.

Nous devons comparer la différence entre les groupes à un étalon (un standard). Un étalon possible est la différence moyenne qui existe entre les observations du *même* groupe. Dans l'exemple du Tableau 11.2, ce serait la différence entre les élèves provenant du même pays. En calculant, pour chaque pays, la différence qui existe entre chaque élève et la moyenne des élèves de ce pays, nous calculons la différence interne au pays. Cette quantité s'appelle la *différence intragroupe*. En additionnant les différences intragroupes, nous obtenons la totalité de la différence à l'intérieur des groupes. La statistique F que nous allons calculer (et la conclusion que nous allons tirer) sera le rapport entre la différence intergroupe et la différence intragroupe.

Lorsque la différence intergroupe est beaucoup plus grande que la différence intragroupe, nous rejetons H_0 et concluons que la différence est statistiquement significative. Ainsi, dans l'exemple du Tableau 11.2, avant de conclure que les élèves de certains pays sont supérieurs en algèbre aux élèves d'autres pays, *il faut démontrer que la différence moyenne entre les pays est plus grande que la différence moyenne entre les élèves, tous pays confondus.* Il nous faut donc établir le rapport entre la différence intergroupe et la différence intragroupe, ce qui produit une nouvelle statistique, la statistique F. La statistique F est nommée en hommage à Ronald Fisher (voir le texte ci-dessous), le célèbre statisticien qui a découvert la distribution F. La Formule 11.3 donne le calcul de la statistique F.

$$F = \frac{\text{différence intergroupe}}{\text{différence intragroupe}} \qquad \text{Formule 11.3}$$

Lorsque ce rapport F est proche de 1,0, cela implique qu'il existe autant de différence entre les groupes qu'il y en a à l'intérieur des groupes. Lorsque ce rapport est nettement plus grand que 1, cela implique que la différence

entre les groupes est plus grande que la différence intragroupe. Puisque nous nous intéressons tout particulièrement à la différence entre les groupes, lorsque le rapport F est « suffisamment » grand, nous concluons qu'il y a peu de chances que tous les groupes proviennent de la même population. En jargon statistique, nous concluons que la différence entre les groupes est significative.

Quiz rapide 11.5
Faites le rapprochement entre la statistique F et la statistique $t_{observée}$ vue au chapitre précédent. Que représente le numérateur pour ces deux statistiques ? Que représente le dénominateur pour ces deux statistiques ?

Sir Ronald A. Fisher : un géant parmi les grands statisticiens

Sir Ronald A. Fisher (1890-1962) fait partie des fondateurs de la statistique telle que nous l'utilisons aujourd'hui, avec William S. Gosset (1876-1937) et Karl Pearson (1857-1936). Fisher est à l'origine de nombreux concepts et procédures statistiques, tels que la variance, la répartition aléatoire, le concept de l'inférence statistique, l'hypothèse nulle et le test de l'hypothèse et, bien sûr, la distribution et la statistique F ainsi que l'ANOVA. Travaillant initialement sur les problèmes liés à l'agriculture (quel type de semences ou d'engrais est plus productif pour quel type de sol), E. F. Lindquist intégra les idées de Fisher au monde de la psychologie et de l'éducation. Si on trouve les statistiques difficiles, il faut blâmer Fisher et Lindquist !

Largement respecté pour son génie mais totalement détesté pour son style interpersonnel, Fisher, aux dires de ses contemporains, avait une personnalité exécrable et un style pédagogique opaque. Ses idées politiques n'étaient guère plus attrayantes : partisan du mouvement eugéniste et craignant une dilution de la qualité génétique des classes sociales supérieures, il favorisait la procréation pour les riches et la décourageait chez les autres. Pour Fisher, même l'infanticide (dans la classe ouvrière) n'était pas nécessairement une mauvaise chose. Heureusement, l'eugénisme étant largement discrédité de nos jours, seules ses idées concernant les statistiques lui ont survécu.

L'ANOVA consiste à faire une comparaison entre deux types de différences (les différences intergroupes et intragroupes). Puisqu'en statistiques, les différences prennent le nom de variance, nous parlons dans ce contexte de variance intergroupe et de variance intragroupe. Ainsi, l'ANOVA comparera la variance intergroupe à la variance intragroupe et c'est pour cette raison que nous appelons cette procédure l'analyse de variance ou, en bref, l'ANOVA (de l'anglais *Analyse of variance*).

Les composantes de la statistique F

Le calcul de la statistique F implique deux termes qui sont expliqués à tour de rôle.

- *La variance intergroupe:* La différence moyenne entre les moyennes de chaque groupe et la moyenne des moyennes (ce qui exige le calcul de la moyenne de tous les groupes, aussi appelée la grande moyenne ou *la moyenne globale*).
- *La variance intragroupe:* La différence moyenne entre chaque observation et la moyenne de son propre groupe.

Ces deux termes sont mis en rapport pour produire la statistique F qui est décrite par la Formule 11.4.

$$\mathrm{F} = \frac{\sum n_j (M_j - M.)^2}{K - 1} \Big/ \frac{\sum_{j=1}^{K} \sum_{i=1}^{n_i} (X_{ij} - M_j)^2}{N - K} \qquad \text{Formule 11.4}$$

Cette formule peut paraître intimidante à première vue mais elle ne l'est pas. Décomposons-la.

La moyenne globale (M.)

Dans les calculs qui suivent, nous aurons besoin de la moyenne globale. L'hypothèse nulle postule que les groupes proviennent tous de la même population (ayant une moyenne unique, disons μ). Nous construisons la meilleure estimation possible de μ en calculant la moyenne de tous les groupes, la *moyenne globale*, aussi appelée la *grande moyenne M.* (M suivi d'un point). La Formule 11.5 donne le calcul de la grande moyenne à partir des moyennes de chaque groupe.

$$\mathrm{M.} = \sum_{j=1}^{K} M_j / K \qquad \text{Formule 11.5}$$

où $M.$ est la grande moyenne, M_j est la moyenne obtenue dans chaque groupe j et K est le nombre de groupes.

La grande moyenne est donc la moyenne des moyennes. Pour les données du Tableau 11.1, la grande moyenne est :

$$\text{M.} = \sum_{j=1}^{K} M_j / K = (40 + 50 + 60)/3 = 150/3 = 50.$$

Cette grande moyenne est la meilleure estimation que nous ayons de μ (la compétence en algèbre) sous l'hypothèse nulle qui spécifie que les élèves des trois pays proviennent de la même population de connaissance de l'algèbre. La grande moyenne sera utile pour le calcul de la différence moyenne entre les groupes, c'est-à-dire la différence intergroupe.

La différence entre les groupes : la somme des carrés intergroupe (SC_{inter})

Nous désirons calculer la différence entre les groupes. Puisque nous connaissons la grande moyenne (M.), nous pouvons calculer la différence entre la moyenne de chaque groupe et la grande moyenne (M_j – M.), puis faire la somme de toutes ces différences [$\sum M_j$ – M.)]. Cependant, il faut *pondérer* chaque différence pour donner plus d'importance aux groupes qui contiennent plus d'observations [$\sum n_j(M_j$ – M.)]. Cette pondération est nécessaire puisque nous savons que les échantillons qui contiennent plus d'observations estiment la population avec plus de précision.

Ainsi, chaque différence obtenue entre la moyenne de chaque groupe et la grande moyenne est multipliée par le nombre d'observations dans le groupe (n_j). Lorsque nous faisons la somme de toutes ces différences, nous obtenons une quantité qui s'appelle la *somme des écarts intergroupe* (entre les moyennes des groupes et la grande moyenne). La Formule 11.6 décrit le calcul.

$$\sum_{j=1}^{K} n_j(M_j - M.) \qquad \text{Formule 11.6}$$

où M_j et n_j sont respectivement la moyenne et le nombre d'observations dans chaque groupe, et M. est la grande moyenne.

Pour les données du Tableau 11.1, il y a 5 observations dans chaque groupe (d'où $n_1 = n_2 = n_3 = 5$) et la grande moyenne est de 50. La somme des écarts intergroupe, est :

$$[5 \times (40 - 50)] + [5 \times (50 - 50)] + [5 \times (60 - 50)] = -50 + 0 + 50 = 0.$$

Cette manière de calculer la somme des écarts produit invariablement la même réponse : zéro ! Il n'y a rien de nouveau ici, puisque la moyenne est le point d'équilibre des données, la somme des écarts entre la moyenne des groupes et la grande moyenne sera toujours égale à zéro.

Comme pour le calcul de la variance, nous contournons ce problème en mettant chaque différence au carré. Cette quantité se nomme la *somme des carrés intergroupe* (SC_{inter}) :

$$SC_{inter} = \sum_{j=1}^{K} n_j (M_j - M.)^2 \qquad \text{Formule 11.7}$$

où tous les termes sont identiques à ceux de la Formule 11.6.

Pour poursuivre l'exemple du Tableau 11.1, la somme des carrés intergroupes est :

$$\begin{aligned} SC_{inter} &= [5 \times (40 - 50)^2] + [5 \times (50 - 50)^2] + [5 \times (60 - 50)^2] \\ &= [5 \times 10^2] + [5 \times 0^2] + [5 \times 10^2] \\ &= 1\,000 \end{aligned}$$

La différence au carré entre les moyennes est 1 000. Cette quantité, la somme des carrés intergroupe, est la statistique qui nous intéresse. Lorsque cette statistique est proche de zéro, cela indique que la différence entre les groupes est proche de zéro. Si, par contraste, la différence intergroupe est grande, il est plus probable que les groupes proviennent de populations différentes.

Il n'est malheureusement pas possible de faire une utilisation directe de cette quantité, SC_{inter}. La raison provient du fait que cette quantité mélange deux grandeurs : la différence entre chaque moyenne et la grande moyenne, d'une part, et le nombre de groupes, d'autre part. Plus le nombre de groupes est grand, plus la somme des carrés intergroupe est grande. Il faut donc séparer ces deux influences.

La solution au problème est simple. Il s'agit de calculer la différence moyenne. Nous calculons la différence moyenne entre les groupes en divisant la SC_{inter} par le nombre de degrés de liberté entre les groupes (que nous expliquons plus loin)

$$dl_{inter} = K - 1 \qquad \text{Formule 11.8}$$

où K est le nombre de groupes.

Cette statistique, la *moyenne des carrés intergroupe*, ou plus simplement le *carré moyen* (*CM*), se calcule avec la formule suivante :

$$CM_{inter} = \frac{SC_{inter}}{dl_{inter}} = \frac{\sum n_j (M_j - M.)^2}{K - 1} \qquad \text{Formule 11.9}$$

Le carré moyen intergroupe pour les données du Tableau 11.1 est CM_{inter} = 1 000/(3 – 1) = 1 000/2 = 500.

La Formule 11.9 est très similaire à la formule pour le calcul de la variance (chapitre 3). On divise la somme des carrés intergroupe par les degrés de liberté K – 1, c'est-à-dire par le nombre de groupes (K) moins un. Tout comme pour le calcul de la variance autour de la moyenne d'un échantillon, une des différences entre les moyennes et la grande moyenne n'est pas libre, ce qui va à l'encontre du postulat de l'indépendance des observations (voir le chapitre 8). Nous corrigeons ce biais en divisant la somme des carrés intergroupe par le nombre de moyennes qui peuvent varier librement : c'est-à-dire toutes sauf une, et nous obtenons K – 1 pour les degrés de liberté. Cette correction produit une estimation non biaisée de la différence entre les groupes dans la population.

Le carré moyen intergroupe (CM_{inter}) est toujours positif, car il est impossible d'avoir moins que zéro différence. Nous trouvons dans l'exemple portant sur la connaissance en algèbre des élèves que la différence moyenne au carré entre nos moyennes est égale à 500. Malheureusement, ce résultat ne nous donne pas encore assez d'informations, puisque nous ne savons pas si une différence de cette taille est grande ou petite, habituelle ou rare. Cette différence entre les moyennes n'est peut-être pas plus grande que la différence typique à laquelle nous pourrions nous attendre si les trois échantillons provenaient de la même population. Par conséquent, il faut comparer cette différence entre les échantillons avec la différence qui existe à l'intérieur des groupes.

Quiz rapide 11.6
Si vous avez trois groupes pour lesquels la moyenne est parfaitement égale, quelle sera, dans ce cas, la quantité CM_{inter}?

La différence intragroupe : la somme des carrés moyens intragroupe

Chaque échantillon est composé d'un certain nombre d'observations. Or, il existe (presque certainement) de la variabilité à l'intérieur de chaque groupe. Dans l'exemple du Tableau 11.2, les élèves d'un même pays n'obtiennent pas tous le même résultat. Il est possible de quantifier cette variation avec la Formule 11.10, la somme des carrés intragroupe (SC_{intra}). On peut remarquer que la Formule 11.10 établit la différence entre chaque observation d'un groupe et la moyenne du groupe et qu'elle met au carré cette différence afin d'éviter que la somme donne zéro :

$$SC_{intra} = \sum_{j=1}^{K} \sum_{i=1}^{n_i} (X_{ij} - M_j)^2 \qquad \text{Formule 11.10}$$

où X_{ij} est le score du sujet i dans le groupe j et M_j est la moyenne pour ce groupe.

La double sommation ($\sum\sum$) indique que nous faisons d'abord la somme des différences au carré entre chaque observation (X_{ij}) et la moyenne de son propre groupe (M_j), puis nous faisons la somme de toutes ces quantités. La formulation suivante est l'expansion de la Formule 11.10. Elle explicite la nature des calculs qui doivent être faits.

$$= \sum_{i=1}^{n_1} (M_{i1} - M_1)^2 + \sum_{i=1}^{n_2} (X_{i2} - M_2)^2 + \ldots + \sum_{i=1}^{n_K} (X_{iK} - M_K)^2$$

Pour nos données du Tableau 11.2, la somme des carrés intragroupe pour le Pays 1 est : $(30 - 40)^2 + (35 - 40)^2 + \ldots + (50 - 40)^2$. Pour le Pays 2, nous calculons $(40 - 50)^2 + (45 - 50)^2 + \ldots + (60 - 50)^2$. Et pour le Pays 3, nous calculons $(50 - 60)^2 + (55 - 60)^2 + \ldots + (70 - 60)^2$. En additionnant chaque somme, nous obtenons la somme des carrés intragroupe : $SC_{intra} = 750$.

Comme pour la somme des carrés intergroupe, la SC_{intra} sera plus grande s'il y a plus d'observations et plus de groupes. Il faudra donc séparer ces deux influences en divisant la quantité SC_{intra} par le nombre de degrés de liberté, qui, lui, devra prendre en considération le nombre total d'observations (N) aussi bien que le nombre de groupes (K).

Inévitablement une observation dans chaque groupe n'est pas libre. Nous perdons donc un degré de liberté par groupe. Au total, les degrés de liberté deviennent N – K, où N est le nombre total d'observations et K est le

nombre de groupes. Ce qui nous amène à la Formule finale 11.11, formule requise pour le calcul du carré moyen intragroupe, CM_{intra}.

$$CM_{intra} = \frac{\sum_{j=1}^{K} \sum_{i=1}^{n_i} (X_{ij} - M_j)^2}{N - K} \qquad \text{Formule 11.11}$$

Lorsque nous avons 5 groupes, chacun composé de 10 observations, le nombre de degrés de liberté intragroupe est $N - K = (5 \times 10) - 5 = 50 - 5 = 45$.

Quiz rapide 11.7
Supposons 10 groupes, chacun ayant 10 observations. Quel est le nombre total de degrés de liberté intergroupe et intragroupe ?

Pour les données du Tableau 11.2, nous avons $SC_{intra} = 750$ pour 15 observations (N = 15) réparties dans 3 groupes (K = 3). Puisque nous avons N = 15 et K = 3, les degrés de liberté intragroupe sont N – K = 15 – 3 = 12. Nous pouvons maintenant calculer le carré moyen intragroupe :

$$CM_{intra} = 750/(15 - 3) = 750/12 = 62{,}5.$$

Une fois les calculs des quantités CM_{inter} (le carré moyen intergroupe) et CM_{intra} (le carré moyen intragroupe) terminés, nous pouvons enfin calculer la statistique F, leur rapport.

Quiz rapide 11.8
Si, dans votre étude, les observations dans chaque groupe sont égales à la moyenne de leur propre groupe, quelle sera la quantité CM_{intra} ? Quelle sera la quantité CM_{inter} ?

Le calcul de la statistique F

La statistique F est le rapport entre la différence moyenne intergroupe et la différence moyenne intragroupe. La Formule 11.12 reprend la Formule 11.3, mais avec ses composantes maintenant formalisées :

$$F = CM_{inter}/CM_{intra} \qquad \text{Formule 11.12}$$

où CM_{inter} est le carré moyen intergroupe et CM_{intra} est le carré moyen intragroupe.

Pour les données du Tableau 11.2, nous savons déjà que $CM_{inter} = 500$ et $CM_{intra} = 62,5$. Nous pouvons donc calculer la statistique F :

$$\begin{aligned} F &= CM_{inter}/CM_{intra} \\ &= 500 / 62,5 \\ &= 8,00 \end{aligned}$$

Si F avait valu 1, nous aurions conclu qu'il existe autant de différence entre les groupes qu'il en existe à l'intérieur des groupes. Mais nous avons obtenu F = 8,00. Cela signifie que la différence moyenne entre les groupes est huit fois plus grande que la différence moyenne à l'intérieur des groupes. Obtenir une différence moyenne entre les groupes huit fois plus grande que la différence moyenne à l'intérieur des groupes (F = 8,00) est possible même lorsque tous les groupes proviennent de la même population. Mais est-ce probable ? Si une telle différence est probable, nous allons conclure que la différence n'est pas statistiquement significative. Mais si elle n'est pas probable, nous allons tirer la conclusion inverse. Il nous faut alors établir la probabilité d'obtenir un rapport F de cette taille si tous les groupes proviennent de la même population. Pour établir cette probabilité, il faut d'abord examiner la distribution de la statistique F (Tableau A.3).

La distribution théorique de la statistique F

En principe, la distribution de la statistique F est construite de manière similaire à celle qui a servi à la construction de la distribution de la statistique t ou celle de la distribution normale. On commence par créer une unique population normale d'observations. Par conséquent, cette distribution n'a qu'une seule moyenne. On choisit au hasard deux groupes, chacun ayant la même taille (N), un nombre infini de fois. À chaque fois, on calcule la statistique $F = CM_{inter} / CM_{intra}$ et on établit une distribution des effectifs de la statistique F pour deux échantillons de taille N.

Puisque les deux échantillons sont extraits de la même population, il ne devrait pas y avoir de différence entre eux, c'est-à-dire que la différence intergroupe (CM_{inter}) devrait être égale à la différence intragroupe (CM_{intra}),

ce qui produira un F proche de 1,0. Mais par pur hasard (l'erreur d'échantillonnage), une certaine proportion des différences sera différente de 1,0. Puisqu'on obtient un nombre infiniment grand de statistiques F calculées à partir d'échantillons extraits de la même population, on peut calculer la proportion des F de différentes tailles.

On répète cette simulation pour différentes tailles d'échantillons *N* et pour différents nombres de groupes K. À la fin, on obtient la distribution des valeurs F pour n'importe quelle combinaison de groupes K – 1 (les degrés de liberté intergroupes) et d'observations N – K (les degrés de liberté intragroupes).

La valeur critique F et le tableau des valeurs critiques de la statistique F

Pour chacune de ces distributions de F, nous identifions la valeur F qui est plus grande que 95 % des valeurs F contenues dans la distribution. Cette valeur F prend le nom de *valeur critique F* ou $F_{critique,}$ et c'est cette valeur qui est inscrite au tableau des valeurs critiques de F que nous retrouvons à l'appendice A3. Par exemple, lorsque nous avons 3 groupes (K = 3, dl_{inter} = K–1=2) et 33 observations (N = 33, dl_{inter} = N–K=30), 95 % des F sont inférieurs à F = 3,32, et moins de 5 % ($p < 0,05$) des F de cette distribution sont égaux ou plus grands que F = 3,32. De la même manière, nous trouvons le F qui correspond à 1 % ($p < 0,01$) ou même 0,1 % ($p < 0,001$). Les tableaux de l'appendice A.3 montrent les valeurs critiques de la statistique F lorsque le nombre de groupes va de 2 à 13 et lorsque le nombre d'observations est moins de 1013. Les valeurs critiques F sont données pour trois seuils α fréquemment utilisés (0,05, 0,01 et 0,001).

Dans les Tableaux A.3 de la distribution des valeurs critiques F dans l'Annexe, chaque colonne représente le nombre de degrés de liberté intergroupe, soit le nombre de degrés de liberté associé au carré moyen intergroupe (CM_{inter}). Nous indiquons ce nombre par la lettre grecque ν_1 (« nu » 1). Le nombre de degrés de liberté intergroupes est $\nu_1 = K - 1$.

Chaque rangée du tableau des valeurs critiques de F représente le nombre de degrés de liberté intragroupes, *N* – K. Nous indiquons ce nombre par ν_2 (« nu » 2).

En utilisant ce tableau, nous trouvons la valeur critique F. Elle se trouve à l'intersection de la colonne $\nu_1 = K - 1$ et de la rangée $\nu_2 = N - K$, qui correspondent à notre analyse. Par exemple, si nous comparons quatre groupes ayant un nombre total de 24 observations, nous calculons d'abord le nombre de degrés de liberté ν_1 et ν_2. Les degrés de liberté intergroupes sont $\nu_1 = K - 1 = 4 - 1 = 3$. Les degrés de liberté intragroupes sont $\nu_2 = N - K = 24 - 4 = 20$. Nous trouvons alors la cellule qui se trouve à l'intersection de la colonne $\nu_1 = 3$ et $\nu_2 = 20$. Dans ce cas, nous trouvons que la valeur critique $F(3, 20) = 3{,}098$ lorsque nous choisissons un risque d'erreur $\alpha = 0{,}05$. Si nous avions choisi un seuil alpha de 0,01, la valeur critique pour ce problème aurait été 4,938. Pour alpha = 0,001, le seuil critique pour $\nu_1 = 3$ et $\nu_2 = 20$ est 8,098.

Quiz rapide 11.9
Dans votre étude, vous comparez 10 groupes ayant chacun 10 observations. Quelle est la valeur critique pour un α de 5 % et pour un α de 1 % ?

L'utilisation du tableau des valeurs critiques de F pour faire une inférence

L'utilisation du tableau des valeurs critiques de F est quasi identique à celle que nous avons vue pour le test t. À partir des valeurs CM_{inter} et CM_{intra}, nous calculons la statistique F. Nous allons maintenant au tableau des valeurs critiques de F (appendice A3) et nous repérons, pour les degrés de liberté et le niveau alpha choisi, la cellule qui y correspond. Enfin, nous comparons le F qui a été calculé à partir des données à celui qui se trouve dans le tableau. Si notre F est inférieur à celui que l'on trouve dans le tableau, nous ne pouvons pas affirmer qu'il existe une différence entre les groupes. Les statisticiens disent que la différence entre les groupes n'est pas statistiquement significative. Mais si le F calculé à partir des données est égal ou plus grand que le $F_{critique}$, nous concluons que la différence entre les groupes est statistiquement significative, au niveau alpha choisi. Au moins un groupe diffère des autres.

SOMMAIRE DU TEST DE L'HYPOTHÈSE POUR K GROUPES

Nous pouvons maintenant formaliser le tout pour montrer comment fonctionne le test d'ANOVA. Comme pour le test t, il y a quatre étapes.

Poser les hypothèses

L'hypothèse nulle présume toujours l'égalité entre tous les K groupes, c'est-à-dire qu'ils proviennent tous d'une population unique. L'hypothèse est qu'il y a au moins un groupe qui diffère des autres. Si on veut être formel :

H_0 : $\mu_1 = \mu_2 = ... = \mu_K$ (les groupes proviennent de la même population).
H : $\mu_i \neq \mu_j$ pour au moins une paire de moyennes (au moins un groupe ne provient pas de la même population).

Choisir le seuil de signification α

Il faut choisir le seuil α avant de regarder les données, pour que ce choix soit objectif. Souvent, on prend $\alpha = 5\,\%$, mais selon le problème et, surtout, l'importance de minimiser les erreurs d'inférence (de types I et II), on pourrait choisir un seuil plus grand ($\alpha = 0{,}10$) ou plus petit ($\alpha = 0{,}01$ ou même $\alpha = 0{,}001$).

Spécifier la règle décisionnelle pour choisir entre H et H_0

Le test d'ANOVA est de la forme :

Rejet de H_0 si $F_{observé} \geq F_{critique}$: nous rejetons l'hypothèse nulle si le $F_{observé}$ est égal ou supérieur au $F_{critique}$.

La valeur $F_{critique}$ s'obtient dans le tableau des valeurs critiques et elle dépend du seuil de signification choisi (α) et des degrés de liberté ν_1 et ν_2 qui correspondent et qui ont été utilisés pour le calcul du $F_{observé}$.

Faire les calculs et conclure

Vérifier la taille du $F_{observé}$ par rapport à la taille du $F_{critique}$ et appliquer la règle décisionnelle identifiée à l'étape précédente.

Quiz rapide 11.10
Vous avez 6 groupes et un total de 100 observations. À la suite de votre ANOVA, vous observez un F = 4,3. Pouvez-vous affirmer avec un niveau d'erreur de 5 % qu'au moins un groupe est différent des autres ? Et si vous voulez courir seulement 1 chance sur 1000 de vous tromper, maintiendrez-vous votre conclusion ?

Le tableau des sources de variance

Souvent, le détail des calculs est présenté suivant un format standard (utilisé entre autres par le logiciel SPSS) que l'on nomme le *tableau des sources de variance*. Ce tableau est pratique parce qu'il résume toutes les statistiques essentielles à notre interprétation, ce qui sera important lorsqu'il s'agira de poursuivre l'ANOVA avec des tests *a posteriori* ou pour le calcul de la taille de l'effet. Nous y reviendrons.

Le Tableau 11.3 est le tableau des sources de variance pour les données du Tableau 11.2, qui décrit la performance en algèbre des élèves dans trois pays. Il indique, à la deuxième colonne, la somme des carrés intergroupe et intragroupe et le total de ces deux quantités, la somme totale des carrés. Cette dernière quantité reflète l'ensemble de toutes les différences qui existent dans notre banque de données.

Tableau 11.3
Sources de variance

	Somme des carrés (SC)	*Degrés de liberté (dl)*	*Carrés moyens (CM)*	*F*	*Seuil de signification α*
Intergroupe	1 000,0	2	500,0	8,00	0,006
Intragroupe	750,0	12	62,5		
Total	1 750,0	14			

La colonne « carrés moyens » (CM) est obtenue en divisant la somme des carrés (SC) par les degrés de liberté (dl) correspondants. La statistique F est obtenue en divisant le carré moyen intergroupe par le carré moyen intragroupe.

La statistique obtenue, F = 8, semble indiquer une grande différence entre les groupes (8 fois plus de différences intergroupes moyennes que de différences intragroupes moyennes). Par conséquent, nous serions tentés de rejeter H_0 et de conclure que la compétence en algèbre n'est pas la même pour les élèves des trois pays. Mais est-ce vraiment le cas ?

Nous trouvons la valeur critique de F pour un seuil α de 5 % ($p < 0,05$) pour $\nu_1 = 2$ et $\nu_2 = 12$ dans le tableau de l'Annexe. Cette valeur critique est $F_{critique} = 3,885$, alors que le $F_{observé}$ est 8,0, une valeur qui lui est supérieure. Nous concluons alors au rejet de $H_{0:}$ il est peu probable (moins de 5 % des chances) d'obtenir une telle différence entre les élèves des différents pays si la performance en algèbre dans les pays est en réalité la même. La différence est statistiquement significative. Lorsqu'il faut décrire notre résultat, nous écrivons : « La différence entre la connaissance en algèbre des élèves des trois pays est statistiquement significative ($F(2,12) = 8,00$, $p < 0,05$). » Il est obligatoire d'indiquer le test statistique utilisé (F), ses degrés de libertés ν_1 et ν_2 (2 et 12 dans ce cas), la taille du F observé et enfin, le seuil alpha indiquant le risque d'une erreur de type I associé à notre conclusion.

Les tableaux de la distribution des valeurs critiques de F, comme ceux présentés dans l'Annexe, ont été conçus avant que les ordinateurs n'existent et afin de rendre le processus d'inférence statistique moins laborieux. Cependant, avec l'arrivée des ordinateurs, les tableaux de la distribution des valeurs critiques de F (ou de t) ne sont presque plus utilisés. Tous les logiciels professionnels d'analyses statistiques (SPSS, SAS, Systat, etc.) calculent la probabilité exacte d'une erreur de type I. Pour conclure, il suffit de vérifier la dernière colonne du Tableau 11.3 des sources de variance qui indique la probabilité exacte de commettre une erreur de type I (conclure au rejet de H_0, alors que les trois groupes proviennent de la même population). Dans ce cas, la probabilité est $p = 0,006$: il y a donc 6 chances sur 1 000 que nous fassions une erreur d'inférence en concluant que la compétence en algèbre n'est pas la même pour les élèves des trois pays. Cette probabilité (0,006) étant inférieure au seuil alpha minimal conventionnel (0,05) nous concluons à la signification statistique.

LES INFLUENCES SUR LA PROBABILITÉ DE REJETER H_0

Comme avec le test t, il est possible de réduire le risque d'une erreur de type I en réduisant le seuil de signification. Par exemple, si nous voulons tester notre hypothèse concernant la compétence en algèbre des élèves (F = 8), mais en n'acceptant qu'un risque très petit de commettre une erreur de type I (disons moins de 1/1000 ou $p < 0{,}001$), nous allons trouver que la valeur critique du F (voir le Tableau A.3.2 dans l'Annexe, pour $\alpha = 0{,}001$ et $\nu_1 = 2$ et $\nu_2 = 12$) est 12,973. Puisque le F observé de 8,0 est inférieur à la valeur critique F(2, 11) = 12,973, nous ne rejetterons pas l'hypothèse nulle, concluant que la différence entre les groupes n'est pas statistiquement significative. Dans ce dernier cas, nous sommes contraints de conclure que tous les groupes proviennent de la même population de connaissance en algèbre.

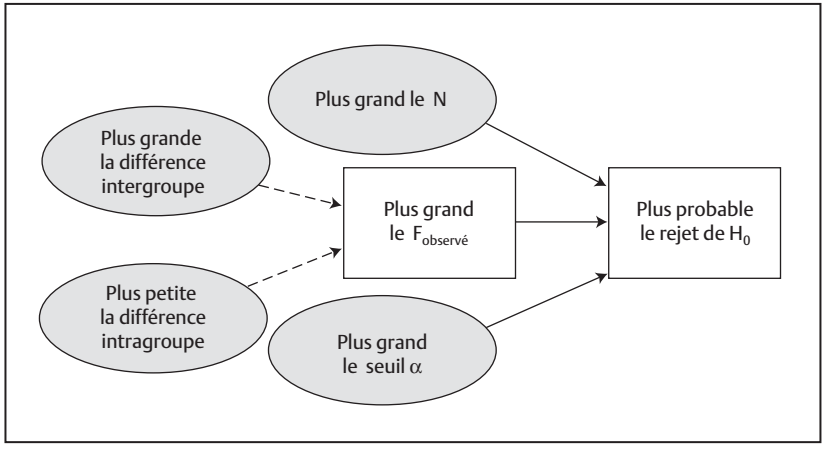
L'illustration précédente nous invite à comprendre qu'un même résultat ($F_{observé}$) pourrait être statistiquement significatif ou non, en fonction de plusieurs facteurs, dont les plus importants sont :

1) la taille du F trouvé ;
2) le nombre d'observations ;
3) le seuil α choisi pour tester l'hypothèse ;
4) l'homogénéité des observations.

La Figure 11.1 représente les principes et constats suivants.

1. *Plus la différence entre les moyennes des groupes est grande, plus la statistique F est grande et plus le rejet de H_0 est probable.* Lorsque la différence entre les moyennes des groupes est plus grande, la quantité CM_{inter} sera plus grande et le $F_{observé}$ résultant sera plus grand. Plus grand est le $F_{observé}$, plus grande est la probabilité qu'il soit égal ou supérieur au $F_{critique}$, et donc il est plus probable qu'il mène à une conclusion du rejet de H_0 (signification statistique).
2. *Plus le nombre d'observations (N) est grand, plus grande est la probabilité que le $F_{observé}$ soit plus grand que le $F_{critique}$, et plus probable est le rejet de H_0.* Le nombre d'observations a un effet direct sur la taille du $F_{critique}$ requis pour conclure à la signification statistique (plus grand est le N, plus petit est le $F_{critique}$). Un même $F_{observé}$ pourrait être non statistiquement significatif lorsqu'il est calculé à partir d'un N petit, et statistiquement significatif lorsqu'il provient d'une analyse comprenant un plus grand nombre d'observations.

FIGURE 11. 1 Quatre influences sur la probabilité de rejeter H_0 : les différences intergroupe et intragroupe, N et le seuil α choisi



3. *Plus le niveau α choisi est grand ($p < 0,05$ plutôt que $p < 0,01$), plus grande est la probabilité que le $F_{observé}$ soit égal ou plus grand que le $F_{critique}$ et qu'il y ait rejet de H_0.* Une analyse comprenant le même nombre de degrés de liberté inter et intragroupes pourrait être statistiquement significative à $p < 0,05$ mais ne pas l'être au seuil $\alpha = 0,01$.
4. Toutes choses étant égales par ailleurs, *plus la différence à l'intérieur des* groupes *est petite,* plus il est probable que les groupes ne proviennent pas de la même population (rejet de H_0). Plus il y a d'homogénéité dans les observations à l'intérieur des groupes, plus petite est la quantité CM_{intra}. Par conséquent, le rapport CM_{inter}/CM_{intra} sera plus grand, résultant en une statistique $F_{observé}$ de taille supérieure. Plus grand est le $F_{observé}$, plus grande sera la probabilité qu'il soit égal ou supérieur au $F_{critique}$ et donc, il sera plus probable qu'il mène à une conclusion du rejet de H_0 (signification statistique).

Quiz rapide 11.11
Vous pensez que le taux de criminalité est plus grand dans les villes plus grandes. Vous choisissez trois grandes villes, trois villes de taille moyenne et trois petites villes et vous mesurez, pour chacune d'entre elles, le niveau de criminalité. Vous ne trouvez pas de différences statistiquement significatives à $p < ,01$. N'étant pas convaincu que ce résultat est vrai, vous décidez de refaire l'étude. Changeriez-vous quelque chose dans votre nouvelle étude ? Quoi ?

Le choix du seuil α : l'erreur de type I versus l'erreur de type II

Le choix du seuil α est d'abord et avant tout une question d'acceptation du risque d'une erreur.

Le risque d'une erreur de type I est plus grand lorsque nous choisissons un seuil de signification plus grand ($p < 0,05$ plutôt que $p < 0,01$). À l'inverse, le risque d'une erreur de type II est plus grand lorsque nous choisissons un seuil de signification plus petit ($p < 0,01$ plutôt que $p < 0,05$). Ainsi, le choix d'α (petit ou grand) affecte le risque d'une erreur de type I et d'une erreur de type II. De plus, en choisissant le risque d'erreur que nous voulons minimiser (I ou II), nous augmentons l'autre risque d'erreur.

Quel est le risque le plus grave ? Il n'existe pas de règle nous permettant de trancher, ce choix étant totalement circonstanciel, en fonction de la situation. Examinez le scénario suivant.

Supposons que le cancer X est mortel et qu'il n'existe aucun traitement efficace pour le contrer. L'espérance de vie d'une personne atteinte de ce type de cancer est de 6 mois. Supposons qu'on propose un nouveau traitement médical qui est potentiellement bénéfique. Afin d'évaluer son efficacité, on administre le traitement à un groupe de patients, et on compare le nombre de mois que les personnes de ce groupe vivent comparativement à un groupe témoin de patients qui ne le reçoivent pas. On exécute une ANOVA qui compare le nombre de mois que les deux groupes de patients vivent. Afin de choisir le seuil alpha, il nous faudra prendre en considération, dans ce cas précis, les conséquences d'une erreur d'inférence. Quel est le risque d'erreur qu'il faut maintenant minimiser : type I ou type II ?

La conséquence d'une erreur de type I

La conséquence liée à une erreur de type I (conclure que le nouveau médicament est efficace, alors qu'il ne l'est pas) n'est pas particulièrement importante dans ce cas. La maladie étant incurable, l'utilisation d'un traitement qui, en réalité, n'est pas efficace n'aura aucun impact notable. Le risque de commettre une erreur d'inférence de type I n'est pas grave. Dans ce cas, il serait approprié de choisir un seuil α plus grand (disons $\alpha = 0,05$ plutôt que $\alpha = 0,01$). Le principe tient aussi bien à l'inverse. Lorsque le ris-

que associé à une erreur de type I est important, on choisira un seuil α plus petit ($\alpha = 0{,}01$ plutôt que $\alpha = 0{,}05$).

La conséquence d'une erreur de type II

Supposons que le F de l'ANOVA n'est pas statistiquement significatif au seuil choisi : alpha = 0,01. Nous n'allons pas prescrire le traitement parce que la règle décisionnelle que nous avons choisie ($p < ,01$) indique qu'il n'est pas efficace. Mais supposons que nous faisons erreur et qu'en réalité ce nouveau traitement est efficace. Nous commettons une erreur de type II. Mais si le traitement est en réalité efficace, en tirant une conclusion fausse, nous passons à côté d'un médicament très attendu. Donc, des malades qui auraient pu être sauvés périssent. Commettre une erreur de type II a des conséquences graves.

Dans ce scénario, il est donc plus important de limiter les risques d'une erreur de type II que ceux d'une erreur de type I. Le risque d'une erreur de type II est amoindri en choisissant un seuil α plus grand (disons $\alpha = 0{,}05$ plutôt que $\alpha = 0{,}01$). Ainsi, il sera plus facile de conclure à la signification statistique (traitement efficace) et plus difficile de conclure que la différence n'est pas significative (traitement inefficace).

Si, à l'inverse, la conséquence d'une erreur de type II est moins importante, il faudra opter pour un petit seuil α plutôt qu'un seuil plus grand ($\alpha = 0{,}01$ plutôt que $\alpha = 0{,}05$).

Comment réduire le risque d'erreur de type I et de type II ?

Ainsi, il est important de définir le risque d'erreur d'inférence que nous voulons minimiser. Cela se fera invariablement en étudiant le risque encouru pour chaque type d'erreur : I ou II.

S'il importe de réduire le risque d'une erreur de type I, nous pouvons faire appel à ces stratégies :

1. Réduire le seuil de signification α ($p < 0{,}001$ plutôt que $p < 0{,}05$).
2. Utiliser moins de sujets (d'observations) plutôt que plus.

S'il importe de réduire le risque d'une erreur de type II, nous faisons appel aux stratégies inverses :

1. Augmenter le seuil de signification α ($p < 0{,}05$ plutôt que $p < 0{,}001$).
2. Augmenter le nombre d'observations.

La logique est exactement celle décrite lors de la discussion au sujet de l'intervalle de confiance (chapitre 9). En augmentant le N, les bornes de l'intervalle de confiance se rétrécissent, augmentant la probabilité de conclure que les échantillons ne proviennent pas de la même population (rejet de H_0). À l'inverse, en réduisant le N, les bornes de l'intervalle de confiance s'accroissent, rendant moins probable la conclusion à la signification statistique (non-rejet de H_0).

LES TESTS DE COMPARAISONS MULTIPLES OU TESTS *A POSTERIORI*

La statistique F que l'ANOVA produit est le rapport entre les variabilités intergroupe et intragroupe. Supposons que nous avons trois groupes (groupe 1, groupe 2, groupe 3). On administre un nouveau médicament aux patients du groupe 1, alors que le groupe 2 reçoit un traitement placebo. Enfin, le groupe 3 est un groupe témoin qui ne reçoit ni médicament ni placebo. Nous comparons le niveau de symptômes qui existe dans les trois groupes avec l'ANOVA et nous concluons que le $F_{observé}$ est statistiquement significatif. Nous concluons que les trois groupes n'ont pas un niveau de symptômes égal. À partir de ce résultat statistiquement significatif, toutes les conclusions suivantes *sont potentiellement, mais pas nécessairement,* justes.

1) $\mu_1 \neq \mu_2 = \mu_3$. L'effet du nouveau médicament diffère de celui du placebo, qui produit un effet égal à celui du groupe témoin.
2) $\mu_1 = \mu_2 \neq \mu_3$. Le nouveau médicament est aussi efficace que le traitement placebo, mais les deux traitements diffèrent du groupe témoin.
3) $\mu_1 \neq \mu_2 \neq \mu_3$. Le nouveau médicament diffère du placebo, qui, lui, est différent du groupe témoin.

Ces résultats possibles mènent à des conclusions très différentes. Par exemple, une compagnie pharmaceutique serait ravie du résultat 1 ; par contre, elle serait profondément déçue du résultat 2. Il est donc utile de pouvoir distinguer entre ces diverses interprétations. L'ANOVA étant construite en comparant la différence intergroupe à la différence intragroupe, elle n'est pas en mesure de nous aider à résoudre le problème. Par

conséquent, des procédures statistiques additionnelles, les tests de comparaisons multiples (que l'on appelle parfois *tests de comparaison* a posteriori, *tests de comparaison* post hoc, ou encore plus simplement *tests* a posteriori), ont été conçues afin de nous aider à déterminer lequel ou lesquels des groupes se différencient des autres. Techniquement, lorsque nous avons obtenu une différence significative à la suite d'une ANOVA, le test de comparaisons multiples identifie avec précision la source de la différence. Par conséquent, les *tests de comparaisons multiples ne sont interprétables (et donc ne doivent être exécutés) que si l'ANOVA indique une différence statistiquement significative.*

Lorsque l'analyse de variance produit un F non statistiquement significatif, cela nous indique que tous les groupes proviennent de la même population. Par conséquent, il serait insensé d'analyser leurs différences alors que nous savons déjà qu'ils proviennent tous de la même population. Mais il faut faire attention : certains logiciels impriment automatiquement les tests *a posteriori*. Cela ne veut pas dire qu'il faille obligatoirement s'en servir pour tirer une conclusion.

Le test de comparaisons multiples de Scheffé

Il existe une grande diversité de tests de comparaison multiple mais nous n'en décrirons qu'un seul, le test de Scheffé. Ce test a été choisi parce qu'il est le plus conservateur (moins enclin à produire une erreur de type I[1]) et plus général dans ses applications. Le test de Scheffé est très flexible et permet la comparaison entre deux groupes ou entre deux ensembles de groupes. Par exemple, si nous avons 4 groupes, il nous permet de comparer chaque groupe avec les autres ou de faire une comparaison entre les groupes 1 et 2 versus les groupes 3 et 4, ou 1 versus 2, 3 et 4, etc. Qu'il y ait ou non le même nombre d'observations dans les divers groupes, on peut l'utiliser. Enfin, les informations requises par le test de Scheffé sont toutes

1. En contrepartie, et pour faire suite à la discussion portant sur les erreurs de type I et de type II, le test de Scheffé limite le risque d'une erreur de type I, mais il augmente le risque d'une erreur de type II. Le risque d'une erreur de type II est amoindri lorsque nous choisissons un seuil de signification plus grand. Par conséquent, il est généralement acceptable de choisir un seuil $\alpha = 0{,}10$ pour le test de Scheffé (plutôt que $p < 0{,}05$, qui est le critère minimal généralement requis).

disponibles à partir du tableau des sources de variance de l'ANOVA, ce qui facilite beaucoup les calculs requis.

En principe, le test de Scheffé suit la logique habituelle. Nous allons calculer une nouvelle statistique ($C_{observé}$) qui sera comparée à une valeur critique ($C_{critique}$). Dans cette application, la statistique $C_{critique}$ n'étant pas tabulée, il faudra la dériver à partir du tableau de la distribution de la statistique F.

Le calcul du test de comparaison multiple de Scheffé implique cinq étapes.

1. On exécute une ANOVA qui compare tous les groupes. Le test de Scheffé ne sera appliqué que si on obtient un F statistiquement significatif : il s'agit maintenant de déterminer où se trouvent les différences existantes.
2. On identifie la comparaison désirée (ex : groupe 1 vs groupe 2).
3. On calcule pour ces comparaisons la statistique $C_{observé}$ avec la Formule 11.13.
4. On calcule la valeur critique $C_{critique}$ avec la Formule 11.14.
5. On compare la statistique $C_{observé}$ à la valeur $C_{critique}$. Lorsque la statistique $C_{observé}$ est égale ou plus grande que la valeur $C_{critique}$, on conclut que les groupes examinés dans la comparaison sont statistiquement différents l'un de l'autre.

Les formules pour le calcul du test de Scheffé pour la comparaison multiple

Nous commençons par une présentation des formules pour le calcul des statistiques $C_{observé}$ et $C_{critique,}$ que nous illustrons à partir des données du Tableau 11.2. La construction du test de Scheffé exige d'abord le calcul de la statistique $C_{observé,}$ qui se donne avec la Formule 11.13

$$C_{observé} = \frac{M_1 - M_2}{\sqrt{CM_{intra}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{Formule 11.13}$$

où M_1 et M_2 sont les moyennes des groupes que l'on veut comparer, CM_{intra} est le carré moyen intragroupe, puisé directement dans le tableau des sources de variance de l'ANOVA, et n_1 et n_2 sont le nombre d'observations associées à chaque groupe comparé.

On remarquera la simplicité de la conception de Scheffé (Formule 11.13). Elle répond à la question suivante : la différence entre les deux groupes en question est-elle plus ou moins grande que la différence moyenne entre les personnes, tous groupes confondus ?

Il faut ensuite calculer la statistique $C_{critique}$:

$$C_{critique} = \sqrt{(k-1)F_{critique}} \qquad \text{Formule 11.14}$$

où le $F_{critique}$ est celui que l'on retrouve dans le tableau de la distribution des valeurs de F pour le nombre de degrés de liberté provenant de l'analyse de variance initiale.

Une illustration du $C_{observé}$ et du $C_{critique}$

Les données du Tableau 11.2 et les données du tableau des sources de variance (Tableau 11.3) sont requises pour le calcul des deux statistiques : le $C_{observé}$ et le $C_{critique}$. Puisque nous avons obtenu (Tableau 11.3) un résultat statistiquement significatif à la suite de l'ANOVA, nous pouvons maintenant déterminer si la différence entre, par exemple, le Pays 1 et le Pays 3 est statistiquement significative : les groupes 1 et 3 proviennent-ils de deux populations différentes ou de la même ? Les moyennes obtenues pour ces deux groupes sont $M_1 = 40$; $M_3 = 60$.

$$C_{observé} = \frac{40-60}{\sqrt{62{,}5\left(\frac{1}{5}+\frac{1}{5}\right)}} = \frac{20}{\sqrt{62{,}5(0{,}20+0{,}20)}} = \frac{20}{\sqrt{25}} = 4$$

Il faut maintenant calculer le $C_{critique}$, qui sera comparé avec le $C_{observé}$.

$$C_{critique} = \sqrt{(k-1)F_{critique}} = \sqrt{(3-1)3{,}885} = 2{,}79$$

La valeur $F_{critique}$ est celle qui provient du tableau de la distribution des valeurs de F, pour les degrés de liberté intergroupes (dans ce cas $\nu_1 = 3 - 1 = 2$), et pour les degrés de liberté intragroupes ($\nu_2 = 15 - 3 = 12$). Pour un seuil alpha de 0,05, cette valeur est 3,885 (voir le tableau pertinent dans l'Annexe).

Le $C_{observé}$ (4) étant supérieur au $C_{critique}$ (2,79), nous concluons que la performance en algèbre des élèves du Pays 1 est différente (inférieure) à celle des élèves du Pays 3. Les élèves du Pays 1 n'appartiennent pas à la même population de connaissance en algèbre que les élèves du Pays 3.

Le calcul du $C_{observé}$ peut sembler un peu long. Le numérateur de la Formule 11.13 est complètement et exclusivement déterminé par la moyenne des groupes que l'on compare ($M_1 - M_2$ ou $M_1 - M_3$, etc.). Cette quantité sera différente pour les diverses comparaisons. Mais lorsque nous analysons la différence entre des groupes de même taille ($n_1 = n_2 = n_3$), il est nécessaire de calculer le dénominateur de la Formule 11.13 une fois seulement et cette valeur sera valide pour la comparaison entre tous les groupes.

Pour les données du Tableau 11.2, chaque groupe a le même nombre d'observations ($n_1 = n_2 = n_3 = 5$). La quantité $CM_{intra} = 62{,}5$ est la même pour toutes les comparaisons. Par conséquent, le dénominateur de la Formule 11.13 est le même pour toutes les comparaisons. Il devient maintenant facile de déterminer où les différences se trouvent. Le Tableau 11.4 indique les résultats pour les données du Tableau 11.2.

Tableau 11.4
Comparaison multiple des moyennes avec la procédure de Scheffé

Comparaison (pays)	$C_{observé}$	$C_{critique}$	*Conclusion*
1 vs 2 (40 vs 50)	2	2,79	1 = 2 (non-rejet de H_0)
2 vs 3 (50 vs 60)	2	2,79	2 = 3 (non-rejet de H_0)
1 vs 3 (40 vs 60)	4	2,79	$1 \neq 3$ (rejet de H_0 ; $p < 0{,}05$)

À partir du Tableau 11.4, les conclusions suivantes sont justifiées. La différence de connaissance en algèbre des élèves des Pays 1 et 2 et des Pays 2 et 3 n'étant pas statistiquement significative, le test de Scheffé indique que seuls les élèves des Pays 1 et 3 diffèrent.

LA TAILLE DE L'EFFET ET LA STATISTIQUE ÊTA AU CARRÉ (η^2)

La taille de l'effet sert à indiquer si la différence entre les groupes est grande ou petite. Il s'agit du rapport de la différence entre les groupes (SC_{inter}) et la différence totale (SC_{total}), tel que décrit par la Formule 11.15. :

$$SC_{total} = SC_{inter} + SC_{intra} \qquad \text{Formule 11.15}$$

La logique de la taille de l'effet est simple : de toutes les différences qui existent dans nos données (SC_{total} à la Formule 11.15), quel est le pourcentage de ces différences qui provient de la différence entre les groupes (SC_{inter} à la Formule 11.15) ? La statistique êta au carré (η^2) est le nom que nous donnons à ce rapport[2] :

$$\eta^2 = SC_{inter}/SC_{total} \qquad \text{Formule 11.16}$$

Dans la Formule 11.16, on voit que la statistique η^2 définit le rapport entre les différences intergroupes par rapport à la différence totale, qui n'est rien de plus que la somme des différences intergroupes et des différences intragroupes (Formule 11.15). La statistique êta au carré peut prendre des valeurs variant entre 0 et 1,0. Mais on choisit souvent de l'exprimer en pourcentage, en multipliant sa valeur par 100 %. Ainsi, si on obtient $\eta^2 = 0{,}25$, on conclut que 25 % de la différence totale observée sur la variable dépendante est expliquée par la variable indépendante. Plus grande est la statistique êta au carré, plus grande est la différence entre les groupes. Lorsque la taille de l'effet est égale à 1,0 (ou 100 %), il faut comprendre que toutes les différences qui existent sont attribuables à (ou « expliquées » par) la différence entre les groupes.

Quiz rapide 11.12
Supposons que vous ne trouvez aucune différence entre les moyennes. Quelle sera obligatoirement la taille de l'effet ?

Une illustration de la taille de l'effet

Reprenons les données du Tableau 11.2 et le tableau de sources de l'ANOVA (le Tableau 11.3). Nous avons testé la différence de connaissance en algèbre des 15 (N = 15) étudiants dans nos trois échantillons (K = 3). Nous avons trouvé $F_{observé} = 8$, ce qui implique une différence statistique-

2. La statistique êta au carré est aussi connue sous le nom de « ratio de corrélation ». Êta est une corrélation non linéaire, car elle mesure le degré de changement sur la variable dépendante en fonction de la variable indépendante. Lorsque la relation entre la variable indépendante et la variable dépendante est linéaire, êta est exactement égal à r_{xy}, la corrélation (linéaire) de Pearson. Êta au carré s'interprète exactement comme le coefficient de détermination (voir le chapitre 6).

ment significative. La performance en algèbre n'est pas la même pour les élèves des trois pays. Mais cette différence est-elle grande ou petite ? Nous calculons alors êta au carré avec la Formule 11.16.

À partir du tableau des sources de variance (Tableau 11.3), nous voyons que la somme des carrés intergroupe = 1 000 et que la somme des carrés intragroupe = 750. La somme totale des carrés est donc $SC_{inter} + SC_{intra} = 1\,000 + 750 = 1750$. Pour le calcul d'êta au carré, nous utilisons la Formule 11.16 :

$$\begin{aligned}\eta^2 &= SC_{inter}/SC_{inter} + SC_{intra} \\ &= 1\,000/1\,750 \\ &= 0{,}57\end{aligned}$$

En exprimant êta au carré en pourcentage, nous obtenons 57 %. L'interprétation de ce résultat est directe : de toutes les différences de compétence en algèbre qui existent entre les élèves, 57 % de ces différences s'expliquent par le pays où l'élève reçoit son enseignement. On pourrait aussi dire que la connaissance du pays de l'élève réduit l'incertitude par rapport à sa compétence en algèbre de 57 %. Le pays est donc un élément important à prendre en considération pour comprendre la compétence en algèbre des élèves. Ce pourrait-il que certains pays utilisent des manuels d'instructions ou des approches pédagogiques qui ne facilitent pas l'apprentissage de l'algèbre ?

Formule simplifiée pour le calcul d'êta au carré

Il arrive souvent que les auteurs d'ouvrages scientifiques ou professionnels ne présentent pas la taille de l'effet ni le tableau des sources de variance dans leurs rapports. Cependant, ils incluent toujours le $F_{observé,}$ qui fait suite à leur ANOVA lorsque celle-ci est statistiquement significative. Pour évaluer à leur juste valeur ces résultats, le test de signification n'est pas suffisant. Il faut aussi calculer la taille de l'effet. Lorsque l'auteur ne présente pas cette taille de l'effet ni les données requises pour en faire le calcul (c'est-à-dire le tableau des sources de variance), on peut néanmoins calculer la taille de l'effet si on a accès au $F_{observé}$, au nombre de groupes (K) et au nombre d'observations (N). Ces informations sont presque toujours incluses dans les écrits. La Formule 11.17 est celle que nous utilisons :

$$\eta^2 = (K - 1)F/[(N - K) + (K - 1)\,F] \qquad \text{Formule 11.17}$$

où K = le nombre de groupes comparés dans l'ANOVA, N = le nombre total d'observations (tous les groupes confondus), et F = la statistique $F_{observé}$ produite par l'ANOVA.

Reprenons le calcul de la taille de l'effet pour la compétence en algèbre des élèves des trois pays en utilisant la Formule 11.17.

$$\begin{aligned}\eta^2 &= (K - 1)F/[(N - K) + (K - 1)\,F] \\ &= (3 - 1)8/[(15 - 3) + (3 - 1)8] \\ &= 16/(12 + 16) = 16 / 28 \\ &= 0{,}57, \text{ ou } 57\,\%\end{aligned}$$

L'interprétation de la taille de l'effet

Toutes choses étant égales par ailleurs, plus la différence intergroupe est grande, plus la taille de l'effet est grande.

Toutes choses étant égales par ailleurs, plus la différence intragroupe est petite, plus la taille de l'effet est grande.

La signification statistique indique si les groupes proviennent de populations différentes. La taille de l'effet indique cependant si la différence entre les groupes est de taille suffisante pour avoir un impact pratique : elle aide à faire la distinction entre une différence statistiquement significative et une différence pratique. Lorsque la taille de l'effet est une valeur extrême (0 ou 1,0), son interprétation est très facile : la différence entre les performances sur la variable dépendante est complètement ($\eta^2 = 1{,}0$) ou nullement ($\eta^2 = 0{,}0$) expliquée par la variable indépendante. Mais il est extrêmement rare que l'on obtienne en pratique de telles valeurs, particulièrement avec les variables qui sont utilisées en sciences sociales. Après tout, il est presque impossible qu'un ensemble de personnes réagissent exactement de la même manière à n'importe quel traitement ou qu'une seule variable indépendante soit parfaitement capable d'établir la distinction entre les observations. Il est quasi certain que la taille de l'effet sera plus petite que 1,0. Comment alors évaluer si elle est « grande » ou « petite » ?

Cohen (1988) propose que la taille de l'effet soit jugée « petite » lorsqu'elle est aux alentours de 1 % ($\eta^2 = 0{,}01$), « moyenne » lorsqu'elle se situe aux alentours de 6 % ($\eta^2 = 0{,}06$) et « grande » lorsqu'elle est aux alentours de

14 % ($\eta^2 = 0{,}14$). Bien que pratiques, ces critères pour définir la taille de l'effet sont parfaitement arbitraires. Le seuil de signification statistique, alpha = 0,05, par exemple, est lui aussi arbitraire.

Plutôt que de se borner aux critères de Cohen, la plupart des chercheurs et des intervenants évaluent la signification pratique en fonction du problème. Une taille de l'effet que l'on pourrait croire petite selon les critères de Cohen peut avoir une forte signification pratique. Par exemple, un médicament qui guérirait seulement 1 % des sidéens ($\eta^2 = 0{,}01$) serait fort important. Le sida étant incurable aujourd'hui, un médicament qui guérirait seulement quelques personnes représenterait un immense pas en avant.

Un dernier mot au sujet de la taille de l'effet et de la statistique η^2. Cette statistique ne doit être calculée et interprétée que lorsque l'ANOVA nous indique que la différence entre les groupes est statistiquement significative. Comme pour les tests *post hoc,* il serait insensé de calculer la taille d'une différence entre deux groupes alors que les deux groupes sont identiques !

SOMMAIRE DU CHAPITRE

L'analyse de variance fait partie des analyses statistiques les plus souvent utilisées. L'ANOVA généralise le test t et permet de déterminer si plusieurs groupes appartiennent à une ou à plus d'une population. L'ANOVA se sert de la statistique F, laquelle compare la différence entre les groupes relative à la différence entre les observations qui proviennent des mêmes groupes. L'interprétation de la statistique F se fait en comparant sa valeur à la valeur critique que l'on trouve dans le tableau des valeurs critiques. En faisant appel aux tests de comparaisons multiples, il est possible, après avoir fait une ANOVA, de déterminer avec plus de précision combien de populations sont représentées par les groupes. Enfin, la taille de l'effet nous donne une indication chiffrée, et en pourcentage, de la taille de la différence entre les groupes. Cette dernière statistique est essentielle pour estimer la signification pratique d'une différence détectée par l'analyse de variance.

EXERCICES DE COMPRÉHENSION

1. La statistique F est utilisée _____.
   a) lorsque nous devons comparer les moyennes de deux groupes ou plus
   b) lorsque nous devons comparer les moyennes de trois groupes ou plus
   c) lorsque nous n'avons pas d'échantillons indépendants
   d) lorsqu'il s'agit de vérifier la signification statistique

2. La signification statistique à la suite d'une ANOVA nous indique si les divers groupes _____, alors que la taille de l'effet nous indique si la différence entre les groupes _____.
   a) ont la même variabilité intergroupe ; est plus grande que zéro
   b) ont des moyennes différentes ; est vraie dans la population
   c) proviennent ou non de la même population ; est grande ou petite
   d) Toutes ces réponses sont justes.

3. Nous divisons la classe en 5 groupes, chacun composé de 21 étudiants. Nous administrons une version différente de l'examen à chaque groupe. Nous testons la différence entre les notes à l'aide d'une ANOVA à un facteur. Les degrés de liberté intergroupes sont _____ et les degrés de liberté intragroupes sont _____.
   a) 5 ; 105
   b) 5 ; 100
   c) 4 ; 105
   d) 4 ; 100

4. À la question 3, nous trouvons le résultat suivant : F = 3,52. D'après le tableau des valeurs critiques de F, les groupes appartiennent-ils à des populations différentes pour le risque d'erreur alpha = 0,05 ? _____ 0,01 ? _____ 0,001 ? _____
   a) Oui ; Oui ; Non.
   b) Oui ; Non ; Non.
   c) Non ; Non ; Oui.
   d) Oui ; Oui ; Oui.

5. En quoi le test de Scheffé est-il utile ?
   a) Pour identifier les groupes qui diffèrent des autres.
   b) Pour indiquer si la différence entre les groupes est de taille importante.
   c) Pour vérifier si les présomptions du test F sont respectées.
   d) Toutes ces réponses sont justes.

6. Lorsque nous trouvons un F = 4,0, cela indique
   a) que la moyenne des groupes diffère
   b) que la moyenne de chaque groupe diffère de la moyenne des autres groupes par un facteur de 4
   c) qu'il y a quatre fois plus de différence moyenne entre les groupes qu'il n'en existe en moyenne entre les individus
   d) que les chances de conclure que la différence est statistiquement différente sont de 1 sur 4

7. À la suite d'une ANOVA à un facteur, vous tirez la conclusion suivante : le test F étant statistiquement significatif à $p < 0,05$, les cinq groupes de l'étude n'appartiennent pas tous à la même population. Quel est le risque, dans ce cas, d'une erreur de type I ?
   a) 5 %
   b) 95 %
   c) Le risque ne peut pas être établi, car nous ne connaissons pas le groupe qui diffère des autres.
   d) Le risque peut être établi, mais il nous faudrait connaître le nombre d'observations pour ce faire.

8. La différence entre les 3 groupes est statistiquement significative à $\alpha = 0,001$. De toutes les différences qui existent dans cette banque de données, la différence entre les groupes en explique 50 %. Quelle est la taille de l'effet dans ce cas ?
   a) 99,99 % (1 – 0,001)
   b) 0,01 % (1 – 99,99)
   c) 50 %
   d) 1,5 (50 % de 3)

9. Nous avons à notre disposition la taille de toutes les femmes et de tous les hommes du Canada. Les femmes mesurent, en moyenne 1,50 m et les hommes mesurent, en moyenne, 1,5001 m. Dans ce cas, la différence entre les hommes et les femmes fait-elle qu'ils appartiennent à des populations de taille différente ?

a) Oui.

b) Non.

c) Probablement que non, mais il faudra faire un test de signification statistique pour en être certain.

d) Probablement que oui, mais il faudra faire un test de signification statistique pour en être certain.

Réponses

1. a
2. c
3. d
4. a
5. a
6. c
7. a
8. c
9. a (*Nota bene* : nous analysons des populations entières ; par conséquent, les tests statistiques ne sont pas pertinents.)

CHAPITRE 12
L'ANALYSE DE VARIANCE FACTORIELLE

CHAPITRE 12

L'ANALYSE DE VARIANCE FACTORIELLE

Le test t examine la différence entre deux petits groupes. L'analyse de variance à un facteur généralise la procédure, en examinant la différence entre plusieurs groupes de toutes tailles, chacun représentant un niveau différent de la même variable indépendante. Elle indique si les différents groupes, représentant les différents niveaux de la variable indépendante, proviennent ou non de la même population. L'ANOVA factorielle est une procédure statistique qui généralise l'ANOVA à un facteur. Elle permet d'examiner l'impact simple et conjoint sur une variable dépendante de plusieurs variables indépendantes, chacune détenant un nombre théoriquement illimité de niveaux, avec des échantillons de toutes tailles. Bien qu'il n'y ait pas de limite inhérente au nombre de variables indépendantes qui peuvent être analysées par l'ANOVA factorielle, ce chapitre se limite à l'analyse de deux variables indépendantes. Ainsi, l'ANOVA factorielle à deux facteurs examine l'impact de chacune des deux variables indépendantes, ainsi que leur impact conjoint sur une variable dépendante.

L'ANOVA À UN FACTEUR ET L'ANOVA FACTORIELLE : SIMILARITÉS ET DIFFÉRENCES

L'ANOVA factorielle est une construction statistique qui se sert de la totalité des concepts et des procédures utilisés pour l'ANOVA à un facteur (voir le chapitre 11) : dans les deux cas, des groupes définissent les niveaux des variables indépendantes et l'objectif consiste alors à comparer les moyennes obtenues par les groupes sur une seule variable dépendante

afin d'inférer les probabilités que les groupes proviennent ou non de la même population. Dans les deux cas, on calcule la même statistique F qui est construite et qui s'interprète de manière identique : le rapport entre les variabilités intergroupes et intragroupes. Ces estimations de variabilités se calculent avec les mêmes formules (les carrés moyens intra- et intergroupes) pour produire des statistiques $F_{observé}$ qui sont interprétées en les comparant avec les valeurs $F_{critique}$ qui, elles, se trouvent dans le même tableau des valeurs critiques. Les tests de comparaison multiple ainsi que la taille de l'effet se mesurent et s'interprètent de la même façon pour les deux formes d'analyse.

La distinction entre l'ANOVA factorielle et l'ANOVA à un facteur se situe sur le plan de la variable indépendante. Pour l'ANOVA à un facteur, on travaille avec une unique variable indépendante. Mais l'ANOVA factorielle est moins limitée : elle permet de vérifier, dans la même analyse, l'impact de plusieurs variables indépendantes sur l'unique variable dépendante, mais, ce qui est plus important encore, elle permet d'évaluer l'impact conjoint de ces variables indépendantes. Pour ce faire, elle introduit un nouveau concept statistique : *l'interaction*. L'interaction évalue l'influence conjointe de deux variables indépendantes (dans le cas de l'ANOVA factorielle à deux facteurs) sur la variable dépendante.

IMPORTANCE DE L'ÉTUDE DES INTERACTIONS

L'analyse de variance factorielle (à deux facteurs) est utilisée lorsque l'on soupçonne (ou postule) que l'effet d'une variable indépendante sur la variable dépendante n'est pas le même pour différentes valeurs d'une deuxième variable indépendante.

Par exemple, les pharmaciens s'inquiètent de plus en plus des effets « interactifs » des médicaments prescrits. En pratique clinique, le médecin choisirait de prescrire le médicament qui aurait la meilleure chance de soulager ou de guérir le malaise d'un patient. Mais il arrive parfois, particulièrement chez les personnes plus âgées, que le patient souffre simultanément de plusieurs malaises. En prescrivant le meilleur remède disponible pour chacun des malaises, on pourrait s'attendre à une guérison plus complète des maux du patient. Or, ce n'est pas toujours le cas. Parfois, deux médica-

ments parfaitement sécuritaires et efficaces lorsque pris séparément produisent des effets paradoxaux lorsqu'ils sont administrés simultanément : les effets bénéfiques d'un médicament peuvent être annulés (ou dans certains cas majorés) par la consommation de l'autre mais, parfois, la combinaison des deux médicaments peut créer des effets secondaires nocifs pour la santé des patients.

Avec la prolifération des médicaments disponibles, ce type de problème — ce que les pharmaciens et les statisticiens nomment une interaction — est devenu une préoccupation très importante pour ces milieux.

L'interaction entre les approches pour le traitement de la dépression

Imaginons la situation suivante : une compagnie pharmaceutique proclame, après avoir injecté des milliards de dollars dans la recherche et le développement, avoir trouvé un médicament miracle pour soigner la dépression. Pour appuyer ses dires, la compagnie a réalisé une vaste étude auprès de patients souffrant de dépression. La moitié d'entre eux ont pris le médicament en question et l'autre moitié, un médicament placebo. Par ailleurs, dans chaque groupe, la moitié des patients ont suivi simultanément une thérapie psychologique, l'autre moitié, non.

Supposons que la compagnie se limite à écrire dans son rapport : « Le nouveau médicament est efficace, les patients ayant reçu le médicament sont significativement moins déprimés ($F(1, 396) = 7{,}31$, $p < 0{,}05$) que ceux qui n'en ont pas pris. »

Il s'agit d'une affirmation générale : le médicament fonctionne. Or, qu'en est-il de la thérapie ? A-t-elle aidé la moitié des participants qui l'ont suivie ? Est-ce que la combinaison thérapie + médicament est le summum du traitement de la dépression ? En négligeant de prendre en considération l'impact potentiel de la thérapie, la conclusion à laquelle la compagnie est arrivée pourrait être suspecte.

Imaginons que les résultats sont tels que les patients ayant reçu le médicament et la thérapie ont vu leur état s'améliorer. Tel n'est pas le cas des participants ayant reçu le médicament mais pas de thérapie et dont l'état a même légèrement régressé. En faisant la moyenne de ces deux groupes, on trouve que la dépression s'est réduite, mais on ne peut pas clairement dire que cet effet bénéfique est exclusivement attribuable au médicament. La conclusion de la compagnie n'est pas fausse, mais elle pourrait être trompeuse. Elle oublie de mentionner que le médicament a un effet favorable uniquement lorsqu'il est pris en conjonction avec la thérapie. En l'absence de la thérapie, le médicament ne serait pas bénéfique et, pour certains patients, il serait même nuisible.

Ce médicament n'est donc efficace que pour les patients qui suivent une thérapie et l'effet de la thérapie ne se révèle que lorsque les patients prennent le médicament. L'effet bénéfique du médicament est différent lorsque les patients suivent ou ne suivent pas une thérapie. Il y a donc une interaction entre le médicament et la thérapie.

L'ANOVA factorielle est une procédure qui permet d'étudier simultanément les effets uniques et les effets conjoints de plusieurs variables indépendantes sur une seule variable dépendante. Cette procédure est utilisée en psychologie car les psychologues savent que le comportement est souvent tributaire de plusieurs causes. Par exemple, la satisfaction affichée envers son employeur est influencée, en partie, par le statut de l'employé, par la qualité de sa relation avec son superviseur mais aussi par la personnalité de l'employé. L'ANOVA factorielle sert à analyser l'impact de chacune de ces variables (statut, relation et personnalité) ainsi que leurs effets conjoints — l'interaction — sur la satisfaction au travail.

Le terme *facteur* est utilisé dans le même sens que les termes *traitement* ou *variable indépendante*. Comme pour l'ANOVA à un facteur, le facteur est défini par différents niveaux, chacun composé d'un groupe différent. Lorsque nous interprétons la différence entre les niveaux de chaque variable indépendante, nous parlons de l'*effet principal*. Lorsque nous analysons l'effet conjoint des variables, nous parlons de l'*effet d'interaction*, ou plus simplement, d'une *interaction*. Ainsi, pour l'analyse de variance factorielle à deux facteurs, nous aurons à faire trois calculs et trois interprétations :

- L'*effet principal A* se réfère à la différence entre les niveaux de la première variable indépendante. Dans le texte ci-dessus, l'effet principal A réfère à l'impact du médicament seulement : le médicament réduit-il ou non la dépression ?
- L'*effet principal B* analyse la différence entre les niveaux de la deuxième variable indépendante. Dans le texte ci-dessus, l'effet principal B analyse l'impact de la thérapie seulement : la thérapie réduit-elle ou non la dépression ?
- L'*effet d'interaction* $A \times B$ analyse la différence entre chacun des groupes définis par les deux variables indépendantes prises simultanément. Dans le texte ci-dessus, l'effet d'interaction analysera l'effet du médicament compte tenu de l'effet de la thérapie. L'effet du médicament est-il le même ou est-il différent pour les différents niveaux de la variable thérapie ?

Comme nous le verrons dans ce chapitre, l'interprétation de l'interaction est l'aspect le plus important de l'analyse, et celui sur lequel l'interprétation globale reposera en premier.

L'ORGANISATION D'UNE ANOVA FACTORIELLE

Supposons que l'on veuille évaluer l'effet de trois dosages différents d'un médicament sur le comportement de deux catégories de patients (dépressifs et schizophrènes). En termes analytiques, nous avons deux variables indépendantes (les facteurs). Ces deux facteurs sont *le dosage* (composé de trois niveaux : dosages 1, 2 et 3) et *l'état psychologique* du patient (composé de deux niveaux : dépression et schizophrénie).

La base de données comprend donc six groupes, tel qu'il est illustré au Tableau 12.1. Chaque groupe représente une combinaison des deux facteurs : dosage × catégorie de patient. Ainsi, un groupe est composé de dépressifs recevant le dosage 1, un deuxième groupe de dépressifs reçoit le dosage 2 et, enfin, un dernier groupe est composé de dépressifs auquel le troisième dosage est administré. Les trois groupes de schizophrènes sont formés de manière identique.

La variable dépendante est le comportement des patients que nous mesurons par le nombre de visites au médecin ou le nombre de journées d'absence au travail ou à l'école. Nous présumons un total de 10 patients dans chacun des six groupes. Le Tableau 12.1 décrit l'organisation des informations. Avec ce schème, nous postulons que le même médicament sera efficace pour les dépressifs et les schizophrènes à condition que le dosage soit différent pour chaque type de maladie.

Tableau 12.1
Moyennes de chaque groupe

Catégorie (B)	***Dosage (A)***		
	1	2	3
Dépressifs	groupe 1	groupe 2	groupe 3
Schizophrènes	groupe 4	groupe 5	groupe 6

L'ANOVA factorielle produit une statistique $F_{observé}$ pour chaque facteur ainsi qu'une autre pour l'interaction. Les trois résultats (et les trois statistiques F) sont les suivants.

Facteur A (effet principal du dosage) : la comparaison se fait entre les dosages 1, 2 et 3, combinant, pour chaque niveau de dosage les dépressifs et les schizophrènes. Ainsi, au Tableau 12.1, l'analyse combine les groupes 1 et 4 (dosage 1, $N_{A1} = 20$ patients), les groupes 2 et 5 (dosage 2, $N_{A2} = 20$ patients) et les groupes 3 et 6 (dosage 3, $N_{A3} = 20$ patients). Ensuite, elle calcule la moyenne de chacun de ces trois regroupements et produit une statistique $F_{observé}$ qui est le rapport entre les carrés moyens intergroupes et intragroupes. Conceptuellement, il s'agit d'une ANOVA à un facteur ayant trois niveaux, chacun composé de 20 observations. Si le $F_{observé}$ est statistiquement significatif, il faudra conclure que les trois dosages ne produisent pas le même effet sur les patients, peu importe leur maladie.

Facteur B (effet principal de catégorie) : la comparaison se fait entre les deux catégories de patients, les dépressifs et les schizophrènes. Au Tableau 12.1, l'analyse combine les groupes dépressifs recevant les dosages 1, 2 et 3 (dosage 1, $N_{B1} = 30$ patients) et elle combine les groupes de schizophrènes qui reçoivent les dosages 1, 2 et 3 (dosage 1, $N_{B2} = 30$ patients). La statistique F comparera alors deux groupes (les catégories de patients), ce qui réduit cette partie de l'analyse à une ANOVA à un facteur composé de deux niveaux, chacun comportant 30 patients. Si le $F_{observé}$ est statistiquement significatif, il faudra conclure que (indépendamment du dosage) le médicament ne produit pas le même effet sur les patients dépressifs que sur les schizophrènes.

Interaction A × B : la comparaison finale porte sur l'interaction. Ici, une statistique $F_{observé}$ est calculée à partir des variabilités intergroupes et intragroupes des six groupes décrits au Tableau 12.1. La statistique F pour l'interaction comparera alors les six groupes, chacun composé de 10 personnes. Si le $F_{observé}$ pour l'interaction est statistiquement significatif, il faudra conclure que, selon le dosage, le médicament ne produit pas le même effet sur les patients dépressifs que sur les schizophrènes. Le dosage prescrit sera différent selon la maladie.

Chacun de ces résultats statistiques se doit d'être étudié et interprété, mais l'ordre dans lequel cela se fait est important. Nous y reviendrons plus loin dans le chapitre.

LE FONCTIONNEMENT DE L'ANOVA FACTORIELLE

Le fonctionnement interne de l'ANOVA factorielle est quasi identique à celui de l'ANOVA à un facteur. Tout comme pour l'ANOVA à un facteur, l'ANOVA factorielle compare la différence moyenne entre les groupes (les carrés moyens intergroupes) avec la différence moyenne entre les observations (les carrés moyens intragroupes). Ainsi, une statistique $F_{observé}$ est produite pour chaque variable indépendante ainsi que pour l'interaction. Au total, l'ANOVA factorielle produira donc trois statistiques F. En comparant chacune de ces statistiques F avec les valeurs critiques de la statistique F (Tableau A.3 dans l'Annexe), il est possible de conclure si l'hypothèse nulle associée à chacune des trois comparaisons doit être rejetée ou non. Dans tous les cas, le même tableau des valeurs critiques de F est utilisé. Les degrés de liberté intergroupes et intragroupes nécessaires pour ces comparaisons sont décrits plus loin dans le chapitre.

L'interprétation des effets principaux

L'interprétation des effets principaux est identique à celle faite pour l'ANOVA à un facteur. Dans tous les cas, la statistique $F_{observé}$ calculée pour chaque effet principal sera déclarée statistiquement significative lorsqu'elle est égale ou supérieure au $F_{critique}$. Ainsi, un effet principal qui est statistiquement significatif indique qu'au moins un des groupes provient d'une population différente des autres. Comme pour l'ANOVA à un facteur, un test de comparaison multiple (tel le test de Scheffé, voir le chapitre 11) peut être appliqué aux données, permettant de déterminer quel groupe diffère des autres. Enfin, la taille d'effet de chaque effet principal peut être calculée en utilisant la statistique η^2 telle que décrite au chapitre 11.

Par exemple, supposons que, à la suite de l'ANOVA exécutée sur des données d'une expérience de dosage (maladie [voir le Tableau 12.1]), nous trouvons un F statistiquement significatif pour l'effet principal de dosage. Nous pouvons alors déterminer avec un test de Scheffé quel groupe diffère significativement et nous pouvons calculer la taille de cette différence avec la statistique η^2. La même chose pourra être faite pour l'effet principal du type de maladie.

Quiz rapide 12.1
Dans l'exemple du Tableau 12.1, il ne sera pas nécessaire d'exécuter un test de Scheffé à la suite d'un effet statistiquement significatif pour le facteur « type de maladie ». Pouvez-vous expliquer pourquoi ?

Les Formules 12.1 et 12.2 présentent le calcul des statistiques F décrivant les effets principaux.

$$\text{Effet principal A : } F_{\text{intergroupes facteur A}} = \frac{CM_{\text{inter facteur A}}}{CM_{\text{intra}}} \qquad \text{Formule 12.1}$$

$$\text{Effet principal B : } F_{\text{intergroupes facteur B}} = \frac{CM_{\text{inter facteur B}}}{CM_{\text{intra}}} \qquad \text{Formule 12.2}$$

Les formulations exactes requises pour le calcul des écarts moyens inter- et intragroupes sont identiques à celles décrites au chapitre 11 et utilisées pour l'ANOVA à un facteur.

$$\text{Effet d'interaction A} \times \text{B : } F_{\text{interaction}} = \frac{CM_{\text{interaction}}}{CM_{\text{intra}}} \qquad \text{Formule 12.3}$$

L'interaction se calcule avec la Formule 12.3. Le terme $CM_{\text{interaction}}$ est la différence moyenne (au carré) entre la moyenne de chacun des groupes et la moyenne globale. Les carrés moyens intragroupes dans les Formules 12.1 et 12.2 sont une seule et même quantité. Puisque chaque effet principal ainsi que l'interaction incluent toutes les observations, la quantité « intragroupe » est la même pour le test de tous les effets principaux et pour l'interaction.

Un exemple d'ANOVA factorielle à deux facteurs

Supposons que l'on désire étudier, chez les cadres en milieu de carrière, l'impact sur la satisfaction de vie de deux facteurs : la richesse personnelle et l'état de santé. Nous désirons répondre à trois questions :

a) Existe-t-il un effet de la *richesse personnelle* ? Comparativement aux cadres pauvres, les cadres riches jouissent-ils d'une satisfaction de vie différente ? Nous désirons alors tester l'effet principal de la richesse personnelle sur la satisfaction de vie.
b) Existe-t-il un effet de l'*état de santé* ? Comparativement aux cadres malades, les cadres en santé jouissent-ils d'une satisfaction de vie

différente ? Il s'agit alors d'analyser l'effet principal de l'état de santé sur la satisfaction de vie.

c) *Effet d'interaction :* L'état de santé et la richesse personnelle interagissent-ils dans leurs effets sur la satisfaction de vie ? Les cadres qui sont simultanément riches et en santé, riches et malades, pauvres et malades, ou pauvres et en santé proviennent-ils de populations différentes de satisfaction de vie ? Il s'agit alors de tester l'interaction.

Le Tableau 12.2 présente la satisfaction de vie moyenne obtenue dans chaque groupe. La dernière colonne et la dernière rangée du Tableau 12.2 présentent *les moyennes marginales*. Les moyennes marginales sont importantes pour l'interprétation des effets principaux.

La satisfaction de vie des cadres riches mais malades est de 20 alors que celle des cadres riches mais en santé est de 35. Ces deux groupes sont composés uniquement de cadres riches. En prenant la moyenne de ces deux groupes — la moyenne marginale des riches — nous obtenons leur satisfaction moyenne : 27,5. Pour les cadres qui sont pauvres, nous trouvons une moyenne marginale de 20. Nous pouvons anticipé que la comparaison statistique entre la satisfaction de vie des riches et des pauvres impliquera l'analyse de la différence entre 27,5 et 20,0, leurs deux moyennes.

De façon équivalente nous examinons la moyenne marginale pour les cadres malades (20) ainsi que celle pour ceux qui sont en santé (27,5).

Enfin, la moyenne globale (23,75) est la moyenne de satisfaction de vie de tous les participants à l'étude. Elle se calcule par la somme des moyennes pour chaque groupe divisée par le nombre de groupes.

Tableau 12.2
Effet de la richesse et de l'état de santé sur la satisfaction de vie moyenne des cadres en milieu de carrière

	Riches	***Pauvres***	***Moyenne***
Malades	20	20	(20 + 20)/2 = 20
En santé	35	20	(35 + 20)/2 = 27,5
Moyenne	(20 + 35)/2 = 27,5	(20 + 20)/2 = 20	(20 + 20 + 35 + 20)/ 4 = 23,75

Le Tableau 12.2 explicite les comparaisons qui seront faites en indiquant la satisfaction de vie moyenne pour chaque groupe. Les données brutes se retrouvent au Tableau 12.3.

L'effet principal A (la situation financière) : la comparaison sera faite entre les cadres riches, qu'ils soient en santé ou malades (moyenne marginale de 27,5) et ceux qui sont pauvres, nonobstant leur état de santé (moyenne marginale de 20).

L'effet principal B (l'état de santé) : nous comparons maintenant les cadres qui sont malades (moyenne marginale de 20) à ceux qui sont en santé, nonobstant leur situation financière (moyenne marginale de 27,5).

L'effet d'interaction (situation financière × état de santé) : pour l'interaction, nous comparons les moyennes obtenues dans chacun des groupes. Dans ce cas, les moyennes suivantes sont comparées : 20, 20, 35 et 20.

LES HYPOTHÈSES DE L'ANOVA FACTORIELLE

Le jeu d'hypothèses suit la forme habituelle. Nous établissons une hypothèse H que nous comparons avec son hypothèse nulle.

Effet principal pour le facteur A

$H_1 : \mu_{riche} \neq \mu_{pauvre}$: les cadres qui sont riches ont un niveau de satisfaction de vie différent de celui des cadres qui sont pauvres.

$H_{01} : \mu_{riche} = \mu_{pauvre}$: les cadres qui sont riches et ceux qui sont pauvres n'ont pas un niveau de satisfaction de vie inégal.

Effet principal pour le facteur B

$H_2 : \mu_{santé} \neq \mu_{malade}$: les cadres qui sont en santé ont un niveau de satisfaction de vie différent de celui des cadres qui sont malades.

$H_{02} : \mu_{santé} = \mu_{malade}$: les cadres qui sont en santé et les cadres qui sont malades n'ont pas un niveau de satisfaction de vie inégal.

Effet d'interaction A × B

H_3 : la satisfaction de vie des cadres qui sont riches ou pauvres est différente selon qu'ils sont malades ou en santé.

H_{03} : la satisfaction de vie des cadres qui sont riches ou pauvres n'est pas différente selon qu'ils sont malades ou en santé.

Le Tableau 12.3 présente les résultats obtenus pour les quarante cadres participant à cette étude hypothétique. Chaque groupe est composé de 10 cadres. Chaque ligne du Tableau 12.3 définit la situation financière de chaque cadre (riche ou pauvre), son état de santé (en santé ou malade) ainsi que sa satisfaction de vie mesurée sur une échelle allant de 0 à 45 : plus la valeur est forte, plus la satisfaction de vie est forte. Ce tableau représente la façon dont il faut organiser la banque de données afin de procéder à son analyse avec un logiciel, tel que SPSS.

Tableau 12.3
Situation financière, état de santé et satisfaction de vie des cadres

Situation financière	*État de santé*	*Satisfaction de vie*	*Situation financière*	*État de santé*	*Satisfaction de vie*
1 = riche ; 2 = pauvre	1 = en santé ; 2 = malade		1 = riche ; 2 = pauvre	1 = en santé ; 2 = malade	
1	1	25	2	1	10
1	1	30	2	1	15
1	1	35	2	1	20
1	1	40	2	1	25
1	1	45	2	1	30
1	1	25	2	1	10
1	1	30	2	1	15
1	1	35	2	1	20
1	1	40	2	1	25
1	1	45	2	1	30
1	1	10	2	2	10
1	2	15	2	2	15
1	2	20	2	2	20
1	2	25	2	2	25
1	2	30	2	2	30
1	2	10	2	2	10
1	2	15	2	2	15
1	2	20	2	2	20
1	2	25	2	2	25
1	2	30	2	2	30

LA DÉCOMPOSITION DE LA SOMME TOTALE DES CARRÉS

Dans le cas de l'ANOVA factorielle à deux facteurs, la somme totale des différences inclut les éléments suivants : l'interaction, chacun des deux facteurs et la variabilité intragroupe. La Formule 12.4 représente cette sommation :

$$SC_{total} = SC_{inter\ facteur\ 1} + SC_{inter\ facteur\ 2} + SC_{interaction} + SC_{intra} \qquad \text{Formule 12.4}$$

Chacune des quatre sources de variabilité est indépendante des autres. L'indépendance veut dire que chaque élément est libre de varier sans être influencé par les autres. Ainsi, il est tout à fait possible de conclure ou non à une différence statistiquement significative pour le facteur A et/ou le facteur B et/ou pour l'interaction A × B. Ces quatre sources de variabilité se retrouvent au tableau des sources de variance pour l'ANOVA factorielle.

Le tableau des sources de variance pour l'ANOVA factorielle

Le Tableau 12.4 montre le tableau de sources de variance pour cette ANOVA factorielle hypothétique. Pour chaque effet principal et pour l'interaction, nous y retrouvons la somme des écarts au carré (SC), les degrés de liberté (dl), les carrés moyens (CM), la statistique F et la probabilité (p) qu'une telle différence puisse exister lorsque tous les échantillons sont extraits de la même population. Le nombre d'observations dans chaque groupe est n = 10, pour un total de 40 personnes.

Tableau 12.4
Tableau des sources de variance pour les données du Tableau 12.3

Source	*SC*	*dl*	*CM*	*F*	*p*
Finance	562,5	1	562,5	10,13	0,003
Santé	562,5	1	562,5	10,13	0,003
Finance × Santé	562,5	1	562,5	10,13	0,003
Intragroupe	2 000,0	36	55,6		

Le tableau des sources de variance indique dans le cas présent que les effets principaux plus l'interaction sont statistiquement significatifs ($p < 0{,}003$).

L'interprétation de ces différences significatives se fait plus facilement par l'entremise de graphiques. Un graphique est généralement construit pour visualiser chacune des différences significatives qui émergent du tableau des sources de variance. Lorsqu'une différence n'est pas statistiquement significative, il n'est pas utile de la décrire graphiquement.

La signification statistique des statistiques F pour l'ANOVA factorielle

La signification statistique de chacun des effets, quantifiée par les statistiques F, est établie en se référant au tableau des valeurs critiques de F. Il s'agit du même tableau que celui utilisé pour l'ANOVA simple et sa lecture est identique. Après avoir calculé la statistique F pour chaque facteur et pour l'interaction et les degrés de liberté associés à chaque comparaison, on cherche dans le tableau la valeur critique $F(dl_{inter}, dl_{intra})$ correspondant au seuil α désiré (0,05, 0,01, etc.). On compare le $F_{observé}$ au $F_{critique}$. Lorsque le $F_{observé}$ est égal ou supérieur à la valeur critique, l'effet est statistiquement significatif, avec le risque d'une erreur de type I qui correspond au seuil α choisi.

Les degrés de liberté pour l'ANOVA factorielle

Les degrés de liberté pour l'ANOVA factorielle suivent étroitement la logique décrite pour l'ANOVA simple. Il faut calculer les degrés de liberté pour chaque comparaison, y compris celle se rapportant à l'interaction, et faire le calcul des degrés de liberté pour la différence intragroupe. Comme pour l'ANOVA simple, nous perdons un degré de liberté pour chaque moyenne calculée.

Les degrés de liberté intergroupes pour les effets principaux

Les degrés de liberté pour les effets principaux (les différences intergroupes) sont donnés par K – 1, où K est le nombre de niveaux pour le facteur considéré. Dans le cas de l'ANOVA, qui teste l'effet de la richesse et de la santé sur la satisfaction de vie, nous avons deux niveaux (groupes) pour chacune des deux variables indépendantes (facteur A : riches vs pauvres ;

facteur B : malades vs en santé). Nous pouvons alors calculer le nombre de degrés de liberté pour le facteur A ($K_A - 1 = 2 - 1 = 1$) et pour le facteur B ($K_B - 1 = 2 - 1 = 1$).

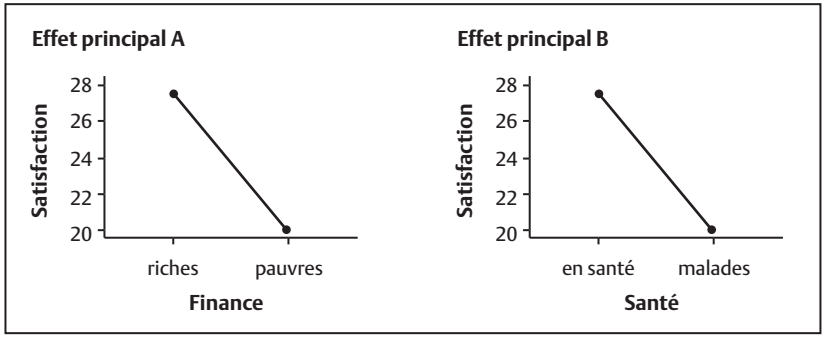
Les degrés de liberté pour l'interaction

Les degrés de liberté pour l'interaction s'obtiennent par le produit du nombre de groupes moins 1 pour chaque effet principal impliqué dans l'interaction. Pour l'ANOVA factorielle avec deux facteurs, les degrés de liberté pour l'interaction deviennent : $dl_{interaction} = (K_A - 1) \times (K_B - 1) = (2 - 1) \times (2 - 1) = 1 \times 1 = 1$. Si nous avions six groupes pour le facteur A et trois groupes pour le facteur B, les degrés de liberté pour l'interaction deviendraient $dl_{interaction} = (K_A - 1) \times (K_B - 1) = (6 - 1) \times (3 - 1) = 5 \times 2 = 10$.

Les graphiques d'interprétation pour les ANOVA factorielles

On construit généralement des graphiques pour faire l'interprétation des résultats obtenus à la suite d'une analyse de variance factorielle. Les Graphiques 12.1 et 12.2 illustrent les moyennes marginales décrites au Tableau 12.2. Chacun des graphiques présente un effet principal statistiquement significatif du Tableau 12.4. En principe, en l'absence d'une différence statistiquement significative, il n'y a pas lieu de construire un graphique d'interprétation.

FIGURE 12.1 Graphiques décrivant les deux effets principaux statistiquement significatifs détectés au Tableau 12.4



Pour les effets principaux, l'élaboration des graphiques est facile. Pour chacun des facteurs, il faut faire un graphique avec les niveaux de ce facteur en abscisse. Par exemple, pour la variable indépendante « richesse », nous aurions deux catégories le long de l'abscisse : la première représentant les cadres riches ; la deuxième, les cadres qui sont pauvres. L'ordonnée représente la variable dépendante. Il s'agit ici de la satisfaction de vie. Nous mettons un point représentant la moyenne de satisfaction de vie (sur l'ordonnée) pour les cadres riches et nous y plaçons un deuxième point définissant la satisfaction de vie moyenne pour les cadres pauvres. Nous faisons la même chose, sur un graphique séparé, pour représenter les états de santé (satisfaction de vie pour les personnes malades ou en santé).

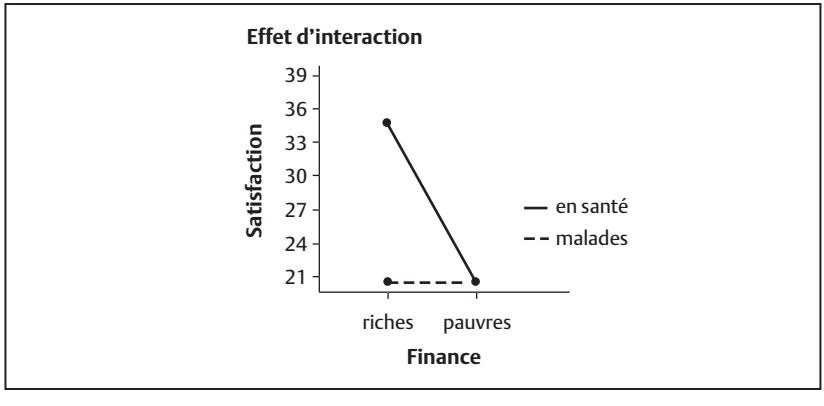
La Figure 12.1 reprend les moyennes marginales du Tableau 12.2. À gauche (effet principal des finances, A), on voit une nette chute entre la satisfaction de vie des personnes riches et celle des personnes pauvres. À droite (effet principal de la santé, B), on voit une chute lorsque la comparaison est faite entre les cadres en santé et ceux qui sont malades.

La représentation visuelle d'une interaction est un peu plus complexe à réaliser. Comme pour les effets principaux, elle n'est requise que lorsque l'interaction est statistiquement significative. Puisque l'interaction représente l'effet conjoint des deux variables indépendantes, les deux variables doivent être représentées sur le même graphique. La Figure 12.2 montre le graphique décrivant l'interaction. Les moyennes placées sur le graphique sont celles obtenues par chacun des groupes (voir Tableau 12.2). Dans ce cas, nous avons un total de quatre groupes et il faut alors mettre quatre points sur le graphique.

L'abscisse représente l'une des deux variables indépendantes (celle de votre choix). Nous utilisons une ligne différente pour chaque niveau de l'autre variable indépendante. À la Figure 12.2, l'abscisse représente la variable indépendante Finance composée de deux niveaux : riche et pauvre. La deuxième variable indépendante est l'état de santé, elle-même composée de deux niveaux (en santé et malades). Une ligne décrivant les personnes en santé (mais qui sont soit riches soit pauvres) et une deuxième ligne représentant les cadres qui sont malades (mais qui sont soit riches soit pauvres) sont placées sur le graphique.

Dans le cas de la Figure 12.2, nous notons que tous les groupes sauf un obtiennent exactement la même satisfaction de vie moyenne : cet unique groupe est composé de personnes qui sont riches et en santé. En examinant cette interaction, nous comprenons maintenant que seules les personnes qui sont simultanément riches et en santé jouissent d'un niveau de satisfaction de vie plus élevé.

FIGURE 12.2 Graphique décrivant l'interaction statistiquement significative détectée au Tableau 12.4



L'interprétation préliminaire des résultats statistiquement significatifs

En se référant aux hypothèses, les conclusions *préliminaires* suivantes peuvent être faites.

H_1 (effet principal A, l'état financier). La statistique $F_{observé} = 10{,}13$, $p < 0{,}003$ indique que la moyenne de satisfaction de vie des personnes riches (27,5) est significativement différente que celle des personnes pauvres (20). La probabilité de trouver une telle différence de moyenne si les riches et les pauvres provenaient de la même population de satisfaction de vie est $p = 0{,}003$. Cette probabilité étant inférieure au seuil alpha de 0,05, la conclusion préliminaire, en se basant sur cette différence statistiquement significative, est qu'il est préférable d'être riche plutôt que pauvre pour être satisfait de sa vie.

H_2 (effet principal B, l'état de santé). La satisfaction de vie moyenne pour les cadres en santé est de 27,5 et celle des cadres malades est de 20. Cette différence produit un $F_{observé}$ de 10,13, qui lui aussi est statistiquement significatif ($p < 0{,}003$). Cette probabilité étant plus petite que le seuil de signification habituel ($p < 0{,}05$), et en notant que la moyenne de satisfaction de vie est plus élevée pour les personnes en santé, nous concluons, de manière préliminaire, qu'il est préférable d'être en santé plutôt que malade.

H_3 (l'interaction, finance × santé). L'interaction est statistiquement significative ($p < 0{,}003$) et indique que la différence de satisfaction de vie entre les cadres riches et les cadres pauvres n'est pas la même pour ceux qui sont en santé et ceux qui sont malades.

Ces interprétations sont *préliminaires* car les résultats qui décrivent les effets principaux ne peuvent être correctement compris que lorsqu'on a préalablement considéré et interprété l'interaction statistiquement significative. Lorsque l'interaction n'est pas significative, chaque effet principal s'interprète de la même manière qu'avec l'ANOVA à un facteur.

L'interprétation définitive des résultats de l'ANOVA factorielle

Le Tableau 12.2 donne la moyenne obtenue par chacun des quatre groupes. Les interprétations préliminaires indiquent que les personnes riches sont plus satisfaites que les pauvres et que les personnes en santé le sont plus que les personnes malades (les deux effets principaux sont statistiquement significatifs). Mais en étudiant l'interaction, en particulier le graphique de l'interaction (Figure 12.2), il est clair que seul un groupe diffère des autres : les personnes riches et en santé. Les personnes qui sont malades sont moins satisfaites, qu'elles soient riches ou pauvres. En ce qui concerne l'inférence, si nous concluons qu'il existe quatre populations de satisfaction de vie — une pour les riches, une autre pour les pauvres, une troisième pour les personnes en santé et une dernière pour les malades —, nous faisons une erreur. En réalité, il n'existe que deux populations : une pour les cadres riches et en santé, et une deuxième pour les trois autres groupes.

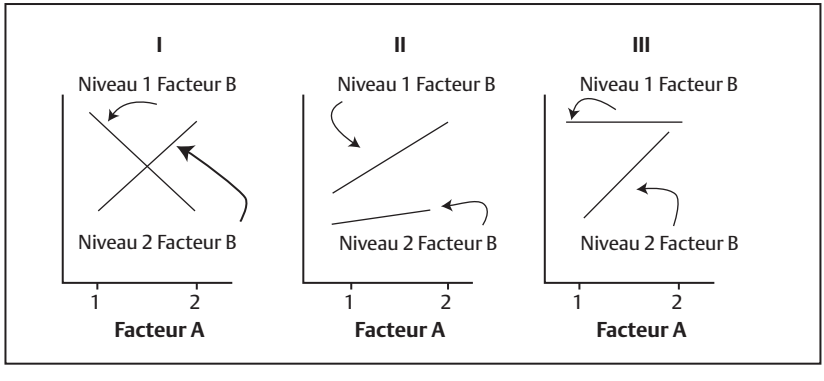
L'interprétation des résultats produits par l'ANOVA factorielle débute invariablement par l'interaction. Lorsque l'interaction est significative, il n'est pas toujours possible de tirer une interprétation valide des effets prin-

cipaux. En effet, l'interaction significative indique que les effets d'un facteur dépendent du niveau de l'autre facteur.

Il s'agit d'une situation où les effets principaux statistiquement significatifs sont trompeurs. Autrement dit, une déclaration générale telle que « La richesse augmente la satisfaction de vie » sera erronée.

Quiz rapide 12.2
Pouvez-vous inventer des moyennes qui correspondraient à la description des quatre groupes de l'encadré du début de ce chapitre ? Prenez pour point de départ que les deux groupes sans médicament ont une moyenne de 100 comportements morbides. Calculez les moyennes marginales.
La compagnie pharmaceutique a-t-elle utilisé dans son rapport les moyennes marginales ou les moyennes de chaque groupe individuellement ?

FIGURE 12.3 Trois formes possibles d'interaction à la suite d'une ANOVA factorielle



La Figure 12.3 décrit trois formes d'interaction, laquelle, par ailleurs, peut en prendre d'autres. Une interaction est détectée lorsque les lignes ne sont pas parallèles. Dans tous les cas, l'interaction n'est à interpréter que lorsque le tableau des sources de variance indique qu'elle est statistiquement significative. Lorsque les lignes sont parallèles, il n'y a pas d'interaction.

En général, une interaction statistiquement significative empêche l'interprétation des effets principaux (les panneaux I et III de la Figure 12.3 illustrent de telles situations). Par contre, le panneau II présente une situa-

tion où l'interaction est détectée (les lignes ne sont pas parallèles), mais où il existe néanmoins un effet principal. Dans ce cas, il existe une différence entre les niveaux 1 et 2 du facteur B. L'interaction significative, pour le panneau II, offre une interprétation supplémentaire : bien que le niveau 1 du facteur B produise une moyenne significativement plus élevée que le niveau 2 de ce même facteur, l'effet du niveau 1 est encore plus fort pour le groupe 2 du facteur A. Une telle situation se produit lorsque les effets d'un traitement sont amplifiés par l'autre traitement. Par exemple, la thérapie peut aider les patients dépressifs et les médicaments le peuvent aussi. Mais les patients qui prennent les médicaments et qui suivent une thérapie progressent encore plus.

Les effets simples

Dans l'éventualité où l'interaction est significative, il est recommandé de décomposer les données en sous-groupes et de tester la différence entre chacun des groupes en utilisant des procédures statistiques supplémentaires que l'on appelle les *effets simples*. Par exemple, dans le cas des cadres, on pourrait choisir de décomposer suivant le niveau de richesse. On pourra donc dire si la santé affecte la satisfaction quand le cadre est riche (effet simple 1) et si la santé affecte la satisfaction quand le cadre est pauvre (effet simple 2). Les hypothèses sont donc :

$$H_1 : \mu_{\text{santé}} \neq \mu_{\text{malade}} \text{ chez les cadres riches}$$
$$H_{01} : \mu_{\text{santé}} = \mu_{\text{malade}} \text{ chez les cadres riches}$$

et

$$H_2 : \mu_{\text{santé}} \neq \mu_{\text{malade}} \text{ chez les cadres pauvres}$$
$$H_{02} : \mu_{\text{santé}} = \mu_{\text{malade}} \text{ chez les cadres pauvres}$$

Encore une fois, l'ANOVA retourne une statistique F pour chaque hypothèse (il y en a autant que le facteur richesse a de niveaux). Un logiciel va obtenir les F pour chacun des effets simples en décomposant la somme des carrés totaux d'une façon différente :

$$SC_{\text{total}} = SC_{\text{inter facteur 1 quand facteur 2 vaut « riches »}} + SC_{\text{inter facteur 1 quand facteur 2 vaut « pauvres »}} + SC_{\text{interaction}} + SC_{\text{intra}} \qquad \text{Formule 12.5}$$

Dans cette nouvelle analyse, seules la SC de l'interaction et la SC intra restent inchangées.

Avec les données du Tableau 12.3, on obtient le tableau d'ANOVA qui suit :

Tableau 12.5
Tableau des sources de variance pour les effets simples basées sur les données du Tableau 12.3

Source	*SC*	*dl*	*CM*	*F*	*p*
Finances quand les cadres sont en santé	1 125,0	1	1 125,0	20,25	0,001
Finances quand les cadres sont malades	0,0	1	0,0	0,00	1,00
Finances × Santé	562,5	1	562,5	10,13	0,003
Intragroupe	2 000,0	36	55,6		

Le F observé lorsque les cadres sont en santé est de 20,25. Si on vérifie dans une table de valeurs critiques, on trouve que la valeur critique pour 1 degré de liberté au numérateur et 36 degrés de liberté au dénominateur est 4,11. Comme le $F_{observé}$ dépasse le $F_{critique}$, on rejette l'hypothèse H_{01} au profit de H_1. Concernant les cadres malades, on ne peut pas rejeter l'hypothèse nulle H_{02}. Il n'y a pas de différence observable due au niveau de richesse sur la satisfaction au travail quand les cadres sont malades.

SOMMAIRE DU CHAPITRE

L'ANOVA factorielle est la généralisation de l'ANOVA à un facteur. Elle permet d'analyser l'effet de plusieurs variables indépendantes et leurs effets conjoints sur une seule variable dépendante. L'ANOVA factorielle se sert des mêmes statistiques et des mêmes formules que l'ANOVA à un facteur, et la signification statistique est éprouvée en se servant des mêmes tableaux des valeurs critiques de la statistique F. La grande différence entre l'ANOVA factorielle et l'ANOVA à un facteur repose sur l'interaction. L'interprétation de l'interaction est facilitée par la construction de graphi-

ques qui présentent simultanément les moyennes obtenues par tous les groupes. L'interprétation des effets principaux se fait exactement comme pour l'ANOVA à un facteur, sauf dans le cas où l'interaction est statistiquement significative. Dans ce cas, il faudra d'abord interpréter l'interaction pour ensuite, si nécessaire, passer à l'interprétation des effets principaux.

EXERCICES DE COMPRÉHENSION

1. L'ANOVA factorielle se distingue de l'ANOVA à un facteur principalement, car L'ANOVA factorielle _____.
   a) ne peut être appliquée qu'avec des échantillons contenant de grands effectifs
   b) ne compare pas la variabilité inter à l'intra
   c) analyse plusieurs variables dépendantes simultanément
   d) analyse l'effet conjoint de plusieurs variables indépendantes

2. Nous avons deux facteurs, A et B, chacun composé de trois niveaux, et nous avons 20 observations par groupe indépendant. Au total, nous avons donc _____ observations, les degrés de liberté intergroupes pour les facteurs A et B sont respectivement de _____ et de _____, et le nombre de degrés de liberté pour l'interaction est de _____.
   a) 180 ; 2 ; 2 ; 4
   b) 60 ; 3 ; 3 ; 6
   c) 180 ; 3 ; 2 ; 4
   d) 60 ; 20 ; 20 ; 20

3. Nous trouvons, à la suite d'une ANOVA à deux facteurs, un effet statistiquement significatif pour l'interaction. Cela veut dire _____.

4. Nous trouvons une interaction statistiquement significative, ainsi que des différences statistiquement significatives pour chacun des deux facteurs de cette ANOVA factorielle à deux facteurs. Pour interpréter ces résultats, il faut _____.
   a) commencer par l'interaction puis passer aux effets principaux ou aux effets simples

b) commencer par les effets principaux, puis passer à l'effet d'interaction
c) interpréter l'interaction et les effets principaux simultanément
d) Toutes ces stratégies sont équivalentes.

5. L'analyse des « effets simples » consiste à déterminer _____.
a) les groupes qui diffèrent des autres, à la suite d'une interaction significative
b) la différence simple entre les moyennes
c) la taille de l'interaction relative à la taille des effets principaux
d) la taille de la statistique F

6. La taille de l'effet à la suite de l'analyse de variance factorielle _____.
a) se calcule et se définit de la même façon que pour l'ANOVA à un facteur
b) n'est interprétable que si la différence est statistiquement significative
c) devient un indice de l'importance de la différence
d) Toutes ces réponses sont justes.

7. Nous construisons un graphique des résultats à la suite d'une ANOVA factorielle à deux niveaux. Nous observons que les lignes du graphique sont parallèles. Alors, _____.
a) aucun des effets principaux n'est « statistiquement significatif »
b) l'interaction est statistiquement significative
c) l'interaction n'est pas statistiquement significative
d) Tous ces résultats sont possibles.

8. Il y a 100 degrés de liberté intra dans cette ANOVA factorielle à deux facteurs. Pour le facteur A, nous avons trois groupes, et quatre groupes pour le facteur B. Nous trouvons les F suivants : interaction F = 4,0 ; facteur A, F = 4,0 ; facteur B, F = 4. Indiquez pour chaque effet s'il est statistiquement significatif ou non, au niveau $\alpha = 0{,}01$.

9. Nous faisons une ANOVA factorielle comparant deux variables indépendantes, chacune composée de cinq niveaux. Combien de tests F le tableau des sources de variance contiendra-t-il ?

a) 10

b) 5

c) 3

d) Cela dépend du nombre d'observations.

Réponses

1. d
2. a
3. La différence observée sur un facteur dépend de la différence observée sur l'autre facteur.
4. a
5. a
6. d
7. c
8. Interaction : oui ; facteur A : non ; facteur B : oui.
9. c

CHAPITRE 13
LES STATISTIQUES NON PARAMÉTRIQUES

Page laissée blanche

CHAPITRE 13

STATISTIQUES NON PARAMÉTRIQUES

La plupart des statistiques vues dans les chapitres précédents ne sont utilisables que lorsque nous sommes en mesure de présumer que le phénomène ou la variable à l'étude se distribue normalement dans la population (voir le chapitre 9). De plus, la distribution normale présume que la variable à l'étude est continue, ce qui revient à dire qu'elle est de type II — c'est-à-dire construite à partir des échelles à intervalles ou de rapport (voir le chapitre 1). Or, ce n'est pas toujours le cas. Prenez l'exemple d'une compagnie qui décide de se doter d'un nouveau logo corporatif. Une firme de publicité lui propose trois logos potentiels. N'arrivant pas à faire un choix final, le Conseil de direction demande à une firme de recherche de trancher en sondant la préférence des clients, ce qu'elle fait en demandant à un échantillon approprié de choisir parmi les trois logos celui qu'il préfère.

La variable indépendante est les logos et la variable dépendante est la préférence qui, elle, est mesurée sur une échelle nominale : la taille des effectifs pour chacune des trois catégories de la variable indépendante. La question statistique à laquelle il faut répondre devient : « La taille des effectifs est-elle la même ou différente pour les trois catégories ? »

Nous ne pouvons pas, dans ce cas, comparer la préférence accordée à chacun des trois logos par l'entremise de l'ANOVA car cette dernière statistique présume que la variable dépendante est continue, ce qui n'est pas le cas pour une variable nominale. Alors comment déterminer la préférence ? En faisant appel à des procédures ***non paramétriques*** qui, elles, n'exigent pas le respect de la présomption de continuité.

De manière similaire, certaines variables — même si elles sont continues — ne sont pas distribuées normalement dans la population. La distribution des revenus est typiquement très asymétrique. Au Québec, par exemple, le revenu modal se situe près de 30 000 $ par année alors que le revenu moyen est plutôt de 60 000 $ par année. La moyenne et le mode n'étant pas identiques, la distribution ne peut pas être normale (voir le chapitre 5). Nous pouvons faire un constat similaire pour le salaire des joueurs de hockey (voir le chapitre 2). Clairement, cette distribution n'est pas normale. Encore une fois, une analyse statistique valide peut être faite, mais seulement en faisant appel à des procédures non paramétriques.

Enfin, dans certaines écoles, les élèves ne reçoivent pas de notes, l'école indiquant uniquement leur classement : la position de chaque élève par rapport aux autres élèves (1er, 2e, etc.). Il s'agit alors d'une échelle ordinale. Nous aimerions savoir si les élèves qui sont « forts » dans une matière — c'est-à-dire qui se situent vers la partie supérieure du classement dans un cours — se classent de manière similaire dans d'autres matières. En principe, il s'agit de calculer la corrélation entre le classement obtenu dans les diverses matières. Mais, comme nous le verrons, la corrélation de Pearson (voir le chapitre 6) n'est pas une statistique appropriée dans ce cas car elle exige, entre autres, que les variables soient de type II. Or, le classement est une variable ordinale, de type I. Encore une fois, une approche non paramétrique est celle qu'il faudra mettre en marche pour analyser ces données.

Dans ce chapitre, nous allons voir les procédures statistiques permettant l'analyse des variables nominales, aussi bien qu'une alternative au coefficient de corrélation de Pearson — le coefficient de corrélation de Spearman —, au test t sur deux groupes indépendants, au test t sur des données appareillées et à l'analyse de la variance à un facteur. Ces analyses statistiques sont le pendant des procédures paramétriques décrites dans les chapitres antérieurs. Plusieurs de ces méthodes non paramétriques nécessitant de calculer des rangs (avec ou sans ex æquo) ; il est important d'avoir assimilé la section 1 du chapitre 4.

L'ANALYSE DES VARIABLES NOMINALES : LE TEST CHI DEUX

Reprenons l'exemple hypothétique des logos : il y en a trois et nous voulons déterminer s'il existe une préférence pour l'un d'eux. Nous présentons les trois logos à 90 personnes et nous demandons à chacune de choisir celui qu'elle préfère. Le Tableau 13.1 présente les données recueillies : 50 personnes ont préféré le logo 2, alors que les 40 autres personnes ont préféré les logos 1 ou 3. L'analyse requise devra répondre à la question suivante : les trois logos sont-ils également préférés ou pas ?

Tableau 13.1
La préférence relative pour les trois logos

	Logo 1	*Logo 2*	*Logo 3*	*Total*
Préférence (la fréquence observée)	24	50	16	90
La fréquence attendue si la préférence pour chacun des logos est égale (H_0)	30	30	30	90

Comme pour n'importe quelle inférence statistique, il faut établir une hypothèse (H) et une hypothèse nulle (H_0). Dans ce cas, l'hypothèse serait qu'il existe effectivement une différence entre les effectifs alors que l'hypothèse nulle devrait être qu'il n'y a pas de différence en ce qui concerne la préférence pour les logos. Si l'hypothèse nulle n'est pas fausse, nous devrions observer une préférence égale pour les trois choix : chaque logo serait choisi par 30 personnes. La deuxième ligne du Tableau 13.1 montre les résultats attendus si l'hypothèse nulle est vraie.

La première ligne du Tableau 13.1 indique la fréquence avec laquelle chaque logo est effectivement choisi. Nous lui donnons le nom de *fréquence observée* (f_o). À première vue, la préférence pour les logos ne semble pas égale, ce qui nous encourage à rejeter l'hypothèse nulle. Mais, comme pour tous les tests statistiques, la différence concernant la préférence pour les logos (f_o) pourrait être le fruit du hasard. Il faut donc déterminer la probabilité d'obtenir la différence observée si, en réalité, il n'y avait pas de différence concernant la préférence pour les logos (c'est-à-dire l'hypothèse nulle). Cette deuxième fréquence prend le nom de *fréquence attendue* (f_a).

En principe la comparaison entre ces deux fréquences sera la base de l'analyse statistique requise.

Pour réaliser cette analyse, il faut faire appel à une nouvelle statistique, la statistique *chi deux*, que l'on nomme parfois le *chi carré* et qui est symbolisée par la lettre grecque χ^2. La statistique χ^2 produit un indice mathématique qui compare la taille de la différence entre la fréquence observée et celle prédite par l'hypothèse nulle. Si la différence est grande, la conclusion sera le rejet de l'hypothèse nulle. Sinon, il faudra conclure que la différence observée est attribuable à l'aléa et, par conséquent, il ne sera pas possible de rejeter H_0. Nous verrons plus loin ce que l'on veut dire par « une grande différence », mais d'abord, examinons la Formule 13.1 qui calcule le χ^2.

$$\chi^2 = \sum \left[\frac{\left(f_o - f_a\right)^2}{f_a} \right] \qquad \text{Formule 13.1}$$

où f_o est la fréquence observée et f_a est la fréquence attendue (sous H_0).

Le numérateur calcule la différence (au carré) qui existe entre les valeurs observées (f_o) et attendues (f_a) pour chaque catégorie de la variable nominale et nous établissons le rapport de cette différence avec la fréquence attendue. Lorsque les fréquences observées et attendues sont les mêmes pour une ou plusieurs catégories, le rapport établi entre ces fréquences sera de zéro. Mais, au fur et à mesure que la différence augmente, le rapport prend des valeurs positives de plus en plus importantes, et ce faisant, la sommation finale produira un χ^2 de plus en plus grand.

La mise au carré élimine le signe de la soustraction. En l'absence de cette précaution, les différences négatives et positives pourraient s'éliminer, créant la conclusion erronée qu'il y a peu de différences entre les fréquences observées et attendues.

Calculons la statistique χ^2 pour le problème des logos. Chaque cellule du Tableau 13.1 doit être prise en considération. Nous calculons la différence entre la fréquence observée f_o et la fréquence attendue f_a (si H_0 est vrai) que nous mettons au carré et nous divisons cette différence au carré par la fréquence attendue. Nous faisons cela pour chaque cellule, puis nous additionnons tous ces résultats pour obtenir la statistique χ^2. Pour les données du Tableau 13.1, $\chi^2 = 21{,}03$.

$$\chi^2 = \sum\left[\frac{\left(f_o - f_a\right)^2}{f_a}\right] = \frac{(24 - 30)^2}{30} + \frac{(50 - 30)^2}{30} + \frac{(16 - 30)^2}{30}$$

$$= \frac{(6)^2}{30} + \frac{(20)^2}{30} + \frac{(14)^2}{30} = \frac{(36)}{30} + \frac{(400)}{30} + \frac{(196)}{30} = 1{,}2 + 13{,}33 + 6{,}53 = 21{,}03$$

$\chi^2 = 21{,}03$

La statistique chi deux est un indice de la taille de la différence entre les fréquences observées et celles réellement obtenues dans notre expérience. Remarquons que la valeur du chi deux ne peut jamais être négative puisque la différence entre les valeurs obtenues (f_o) et attendues (f_a) est mise au carré, ce qui a comme effet d'éliminer les signes négatifs. La valeur minimale de χ^2 est donc 0,0 (lorsque les valeurs observées et attendues sont toutes identiques) et sa valeur maximale est indéterminée. Plus grande est la valeur de χ^2, plus il est probable que les fréquences obtenues diffèrent de celles que nous aurions dû obtenir si l'hypothèse nulle était celle à retenir. Vous remarquerez aussi que plus nous avons de catégories à analyser, plus grande est la somme des différences entre les valeurs obtenues et attendues et plus grande est la quantité chi deux. Si nous avions comparé cinq logos au lieu de trois, la valeur numérique de χ^2 aurait eu de bonnes chances d'être plus grande. Par conséquent, pour faire l'interprétation du chi deux, il faut prendre en considération non seulement sa taille mais aussi le nombre de catégories sur lequel il a été calculé.

L'interprétation de la statistique chi deux

Pour faire une interprétation valide de la statistique chi deux, il faut comparer le résultat obtenu (par exemple $\chi^2 = 21{,}03$) à un tableau des valeurs critiques de la distribution de la statistique χ^2. Le tableau des valeurs critiques de χ^2 se trouve dans l'Annexe A4.

Comme pour toutes les statistiques, il faut prendre en considération les degrés de liberté qui, dans ce cas, seront déterminés à partir du nombre de catégories dans l'étude. La logique du degré de liberté qui s'applique dans ce cas est identique à celle que nous connaissons déjà (chapitre 10). Revenons au Tableau 13.1. Nous savons que nous avons un total de 90 personnes

dans notre étude. Nous savons aussi que 24 personnes ont choisi le logo 1 et que 50 ont choisi le logo 2. Il y a donc 16 personnes qui ont choisi le logo 3 [90 − (24 + 50) = 16]. La taille de l'effectif pour cette dernière catégorie est parfaitement déterminée par la taille des effectifs obtenue pour les autres. Elle ne peut varier librement, et cela implique la perte d'un degré de liberté. De la même façon, en connaissant le nombre total de personnes et le nombre de celles qui préfèrent la catégorie « logo 1 » et la catégorie « logo 3 », nous pouvons déduire exactement le nombre de personnes qui ont choisi la catégorie « logo 2 » [90 − (24 + 16) = 50]. Ainsi le nombre de personnes qui préfèrent une des catégories est parfaitement déterminé par la somme des fréquences observées pour les autres catégories. Une des catégories ne pouvant pas varier librement, le nombre de degrés de liberté devient donc K-1, où K représente le nombre de catégories qui sont comparées. Ainsi, au Tableau 13.1, parce que nous avons trois catégories, le nombre de degrés de liberté est K − 1= 3 − 1= 2.

Quiz rapide 13.1
Ce restaurant offre 10 choix au menu du jour et 99 clients mangent dans ce restaurant aujourd'hui. À la fin de la journée, nous comptons le nombre de personnes qui ont choisi chaque plat. Si nous calculons un χ^2 sur ces données, combien de degrés de liberté avons-nous ?

Le Tableau 13.2 est un extrait du tableau de la distribution des valeurs critiques du chi deux qui se trouvent dans l'Annexe A4. D'abord, nous choisissons la rangée qui correspond au nombre de degrés de liberté dans notre étude. Il s'agit de la quantité ν (« nu »). Puisque nous avons K = 3 catégories au Tableau 13.1, nous obtenons deux degrés de liberté (dl = K − 1 = 3 − 1 = 2) pour ces données. Ensuite, nous choisissons, dans les colonnes du tableau, le seuil alpha. Au Tableau 13.2 (tout comme à l'Annexe A4), nous pouvons établir notre risque d'erreur de type I à moins de 10 %, 5 % ou 1 %. Choisissons alpha < 0,05. La valeur critique notée à l'intersection de la rangée $\nu = 2$ et $\alpha = < 5\,\%$ est 9,488. Le chi deux obtenu est 21,03. Cette valeur étant plus grande que la valeur critique (9,488), nous concluons au rejet de l'hypothèse nulle voulant qu'il n'existe pas de différence de préférence pour les logos : les clients préfèrent le logo 2 et cette conclusion est valide compte tenu que nous avons accepté un risque d'erreur inférieur à 5 %. Mais sup-

posons que l'on désire réduire notre risque d'erreur en choisissant un seuil alpha de < 1 %. La valeur critique pour alpha < 1 % est 21,666. La valeur du chi deux que nous avons obtenue (21,03) étant inférieure à celle de la valeur critique (21,666), nous ne pouvons pas conclure que la préférence diffère statistiquement. Dans ce cas, il serait erroné de rejeter l'hypothèse nulle, et nous serions contraints de conclure que les trois logos jouissent d'une préférence égale.

Tableau 13.2
Extrait du tableau de la distribution des valeurs critiques du χ^2 (Annexe A4)

Degrés de liberté	***Seuil alpha***		
ν	**< 10 %**	**< 5 %**	**< 1 %**
1	2,706	5,991	15,086
2	**4,605**	**9,488**	**21,666**
3	6,251	12,592	26,217

L'ANALYSE DES VARIABLES NOMINALES POUR DEUX VARIABLES INDÉPENDANTES

Maintenant que nous avons en main les éléments requis pour la compréhension et l'interprétation du test chi deux, étendons cette logique à un problème un peu plus complexe. Le Tableau 13.3 revient au problème de la préférence pour les logos sauf que maintenant nous voulons aussi examiner si la préférence est la même pour les hommes et les femmes. Nous refaisons l'analyse mais, cette fois, nous séparons l'échantillon en fonction du genre. Des 90 personnes dans notre étude, 50 sont des hommes et 40 sont des femmes. Le Tableau 13.3 présente la préférence des hommes et celle des femmes pour les logos.

Tableau 13.3
La préférence relative des hommes et des femmes pour les trois logos

	Logo 1	*Logo 2*	*Logo 3*	*Total*
Hommes	12	28	10	**50**
Femmes	12	22	6	**40**
Total	**24**	**50**	**16**	**90**

Le calcul du chi deux exige le calcul des fréquences attendues (f_a). Pour ce faire, il s'agit de calculer, pour chaque cellule du Tableau 13.3, les fréquences marginales qui lui sont associées, que nous divisons par la fréquence totale (toutes les observations). Ici, le total général est 90 et indique le nombre total de personnes dans notre étude. Le Tableau 13.4 présente les calculs.

Tableau 13.4
Le calcul des fréquences attendues pour le Tableau 13.3

	Logo 1	*Logo 2*	*Logo 3*	*Total*
Hommes	$(50 \times 24)/90$ = **13,33**	$(50 \times 50)/90$ = **27,78**	$(50 \times 16)/90$ = **8,89**	**50**
Femmes	$(40 \times 24) / 90$ = **10,67**	$(40 \times 50)/90$ = **22,22**	$(40 \times 16)/90$ = **7,11**	**40**
Total	**24**	**50**	**16**	**90**

Prenons la cellule représentant les hommes qui préfèrent le logo 1. Au total, nous avons 50 hommes dans notre échantillon d'hommes et, au total, 24 personnes (hommes et femmes confondus) préfèrent le logo 1. Pour calculer la fréquence attendue, nous multiplions ces deux fréquences marginales (50 × 24) que nous divisons par 90, le nombre total de personnes dans l'étude. Pour cette cellule, la fréquence attendue est (50 × 24) / 90 = 13,33. Ainsi, 12 personnes ont réellement choisi le logo 1 (Tableau 13.3) mais, par pur hasard, nous nous serions attendus à ce que (exactement) 13,33 personnes le choisissent. Pour la cellule décrivant les femmes qui préfèrent le logo 3, nous obtenons une fréquence attendue (40 × 16) / 90 = 7,11,

alors qu'en réalité 6 femmes l'ont préféré. Nous faisons cela pour chaque cellule du tableau. Nous avons maintenant tous les éléments pour appliquer la formule pour la computation du chi carré avec la Formule 13.1 $\chi^2 = [(12 - 13{,}33)^2/13{,}33] + [(28 - 27{,}78)^2/27{,}78] + \ldots + [(6 - 7{,}11)^2/7{,}11] = 0{,}614$.

Pour être interprétée, cette valeur du chi deux ($\chi^2 = 0{,}614$) doit être confrontée à une valeur critique de la statistique χ^2. Pour choisir les degrés de liberté, nous devons prendre en considération le nombre de catégories de logo (C = 3) ainsi que le nombre de genres dans notre étude (R = 2). À titre mnémonique, nous utilisons la lettre *C* pour indiquer le nombre de colonnes (qui, dans ce cas, reflète les logos) et la lettre *R* pour indiquer le nombre de rangées (représentant les deux genres, hommes et femmes). Nous perdons un degré de liberté pour les colonnes et un degré pour les rangées. Le calcul final devient le produit des deux degrés de liberté. Formellement, les degrés de liberté pour le chi deux ayant deux variables s'obtiennent par la Formule 13.2.

$$dl = (R - 1) \times (C - 1). \qquad \text{Formule 13.2}$$

où R est le nombre de rangées et C est le nombre de colonnes.

Puisque nous avons trois logos (C = 3) et deux genres (R = 2), les degrés de liberté sont $(3 - 1) \times (2 - 1) = 2 \times 1 = 2$.

Au seuil $\alpha < 0{,}05$, et pour 2 degrés de liberté (Tableau A4 dans l'Annexe ou Tableau 13.2 dans le texte), nous trouvons la valeur critique de 9,488. Le chi deux observé n'est que de 0,614, une valeur moindre que la valeur critique impliquant qu'il n'est pas statistiquement significatif. Ne pouvant pas rejeter l'hypothèse nulle, nous concluons que les hommes et les femmes ne diffèrent pas entre eux quant à leurs préférences pour les logos. Nous savons déjà que « généralement » les gens préfèrent le logo 2 et, grâce à cette dernière analyse, il est maintenant établi qu'il n'y a pas de différence de préférence entre les hommes et les femmes. Nous pouvons dès lors recommander au Conseil de direction l'adoption du logo 2.

Quiz rapide 13.2
Pour le lunch, les étudiants de 1re, 2e et 3e année peuvent apporter leur repas de la maison, manger à la cafétéria de l'université ou se rendre au restaurant. Nous demandons à une centaine d'étudiants choisis au hasard ce qu'ils feront pour le lunch aujourd'hui. Vous désirez tester l'hypothèse selon laquelle leurs choix pour le lunch seront différents. Combien de degrés de liberté avons-nous dans cette étude ? Nous avons trouvé $\chi^2 = 13$. La différence est-elle statistiquement significative ?

LA CORRÉLATION ENTRE LES VARIABLES ORDINALES : LE COEFFICIENT DE CORRÉLATION DE SPEARMAN

Pour qu'il soit valable de calculer un r_{xy} de Pearson, il y a deux prérequis :

a) les données doivent être mesurées avec une échelle de type II ;
b) les données doivent être **homoscédastiques** (voir l'encadré).

Lorsque les variables à mettre en corrélation ne sont pas homoscédastiques ou qu'elles ne sont pas de type II, le coefficient de corrélation approprié pour l'analyse est le coefficient de corrélation par rang, généralement appelé le *coefficient de corrélation de Spearman* en l'honneur de son inventeur. La corrélation de Spearman est généralement identifiée par la lettre grecque ρ_{xy} (rho) pour la distinguer du r_{xy} de Pearson (voir le chapitre 6). Le coefficient de corrélation par rang ressemble beaucoup au r_{xy} de Pearson. La valeur numérique de ces deux coefficients (r_{xy} et ρ_{xy}) varie entre −1 et +1.

- Une valeur de +1 (−1) indiquant une corrélation positive (négative) parfaite.
- Une valeur de 0 indiquant une absence de corrélation.

Mis à part la nature exacte des calculs requis, la grande différence entre le coefficient de Pearson et le coefficient de Spearman est que ce dernier exige que les données mises en corrélation soient des rangs. S'il y a N données, la plus petite reçoit le rang 1 et la plus grande, le rang N (s'il n'y a pas d'ex aequo bien entendu). Lorsque les personnes qui obtiennent un rang élevé (ou faible) sur une variable obtiennent aussi un rang élevé (ou faible) sur l'autre variable, la corrélation de Spearman est élevée. La corrélation de Spearman est négative lorsque ceux qui obtiennent de forts rangs sur une variable tendent à obtenir des rangs faibles sur l'autre (et vice versa). Enfin, s'il n'y a ni tendance positive ni tendance négative, la corrélation de Spearman sera pro-

che de zéro. *Ainsi, la corrélation de Spearman indique le degré avec lequel les personnes de l'échantillon occupent le même rang sur les deux variables.*

Le calcul du ρ de Spearman est très simple, s'effectuant en quatre étapes.

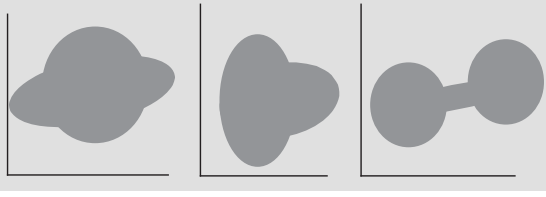
1) Nous mettons en rang croissant chacune des valeurs de la variable X et, séparément, chacune des valeurs de la variable Y. Par exemple l'étudiante qui est première de classe en mathématique mais 15e en français, serait attribuée le rang 1 à la variable X (mathématique) et le rang 15 à la variable Y (Français).
2) Nous calculons pour chaque observation la différence (d_i) entre les rangs obtenus aux deux variables (d_i = rangX – rangY).
3) Nous élevons cette différence au carré pour chaque observation (di^2).
4) Enfin, nous appliquons la Formule 13.3 pour calculer la corrélation par rang.

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N} \qquad \text{Formule 13.3}$$

où N est le nombre d'observations et Σ est la somme des différences au carré.

Les données homoscédastiques versus hétéroscédastiques

Le coefficient de corrélation de Pearson produit un résultat valide à condition que les variables sur lesquelles il est calculé soient continues et que leurs relations soient *homoscédastiques*. L'homoscedasticité des données réfère à la forme du nuage de points dans un graphique de dispersion. Si l'épaisseur du nuage de point est constante pour toutes les valeurs, nous disons que les données sont homoscédastiques et calculer un r_{xy} de Pearson est tout à fait légitime. Mais cela n'est pas toujours le cas. Pensez au nuage de points indexant la relation entre le nombre de mois de chômage au cours des cinq dernières années et le nombre d'années de scolarité. On peut présumer que les personnes peu scolarisées varieront beaucoup quant au nombre de mois de chômage alors que celles très scolarisées varieront peu. Si on fait un graphique de dispersion, le nuage de points aura la forme d'un triangle plutôt que d'un ovale (ici, avec une pointe du côté de la scolarité élevée tel qu'identifié par le pointillé de la Figure 13.1). Ici, nous aurons un cas d'hétéroscédasticité et le coefficient de corrélation de Pearson ne sera pas la technique pertinente pour établir la corrélation entre les mois de chômage et le niveau de scolarité. Les trois nuages de points ci-dessous représentent trois cas d'hétéroscédasticité. Dans ces cas, la corrélation de Pearson n'est pas la forme de la corrélation qui est appropriée.



Un exemple de la corrélation par rang de Spearman

On souhaite savoir si le revenu d'un joueur de hockey au sommet de sa carrière prédit bien son revenu après son retrait en fin de carrière. On consulte donc un échantillon de 10 joueurs à la retraite et on obtient pour chacun son revenu maximal en carrière puis son revenu 10 ans plus tard. Les données (fictives) pour ces 10 personnes (en milliers de $) sont inscrites au Tableau 13.5.

Tableau 13.5
La relation entre le revenu avant et après la retraite

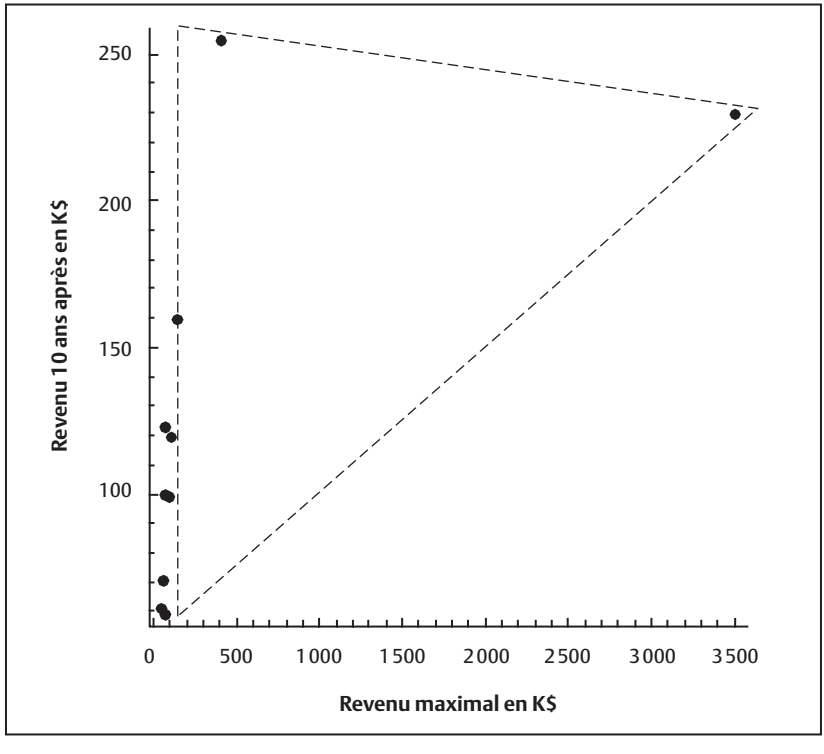
Col 1	*Col 2*	*Col 3*	*Col 4*	*Col 5*	*Col 6*	*Col 7*
Joueurs de hockey	*Revenu pré-retraite (K$)*	*Revenu post-retraite (K$)*	*Rang du revenu pré-retraite*	*Rang du revenu post-retraite*	*d = différence entre les rangs*	*d^2 = différence au carré entre les rangs*
A	43	61	1	2	1	$1^2 = 1$
B	52	71	2	3	1	$1^2 = 1$
C	61	59	3	1	–2	$-2^2 = 4$
D	62	100	4	5	1	$1^2 = 1$
E	72	123	5	7	2	$2^2 = 4$
F	88	99	6	4	–2	$-2^2 = 4$
G	102	120	7	6	–1	$-1^2 = 1$
H	133	160	8	8	0	$0^2 = 0$
I	400	255	9	10	1	$1^2 = 1$
J	3500	4200	10	9	–1	$-1^2 = 1$

Comme on peut le voir en consultant les colonnes 2 et 3 du Tableau 13.5, les revenus les plus élevés sont radicalement différents des revenus plus faibles. Le nuage de point est donné à la Figure 13.1. Techniquement, cette distribution n'est pas homéoscédastique (voir l'encadré).

On voit une certaine tendance positive, mais le fait que le nuage de points ne soit pas rond (loin de là) mais plutôt triangulaire (comme l'indiquent les pointillés à la Figure 13.1) nous empêche de calculer le coefficient

de corrélation de Pearson. On recode donc les données pour ne conserver que leurs rangs absolus tel qu'indiqué au Tableau 13.5 dans les colonnes 4 et 5. À la colonne 6, nous trouvons la différence entre le rang pré et le rang post-retraite obtenus pour chaque joueur et cette différence est mise au carré à la colonne 7.

FIGURE 13.1 Revenu de 10 joueurs de hockey maximal en cours de carrière, puis 10 ans plus tard



Nous avons maintenant en main toutes les informations requises pour calculer la corrélation de Spearman en nous servant de la Formule 13.3.

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N}$$

Nous calculons d'abord la somme des différences au carré entre les rangs : $\Sigma d_i^2 = 1+ 1+ 4+ 1+ 4+ 4+ 1+ 0+ 1+ 1 = 18$, puis nous entrons ce résultat dans le reste de la Formule 13.3.

$$\begin{aligned} 1 - \frac{6 \times 18}{10^3 - 10} &= 1 - \frac{108}{1000 - 10} \\ &= 1 - \frac{108}{990} \\ &= 1 - 0,109 \\ &= 0,890 \end{aligned}$$

La corrélation de Spearman entre les revenus pré et post-retraite est de 0,89, une corrélation substantielle et positive. De toute évidence, les joueurs qui ont eu un meilleur revenu pendant leur vie active finissent 10 ans plus tard avec un revenu plus important que ceux qui ont bénéficié d'un revenu moindre en carrière. Notons que la corrélation obtenue dans ce cas est positive. L'interprétation que l'on fait d'une corrélation de Spearman est identique à celle que l'on fait de la corrélation de Pearson (réduction de l'incertitude, pas de démonstration de causalité, etc.), telle que décrite au chapitre 6[1].

Quiz rapide 13.3
Revenez au Tableau 13.5. Inversez le revenu post-retraite (Col 3) de manière à ce que ce revenu de la personne J soit attribué à la personne A, celui de la personne I à la personne B, etc. Recalculez le rang attribué à chaque revenu post-retraite et calculez la corrélation de Spearman. Quelle est la corrélation que vous obtenez ? Quelle est la conclusion qui s'impose maintenant ?

UN TEST SUR DEUX ÉCHANTILLONS INDÉPENDANTS : LE WILCOXON-MANN-WHITNEY

Le test t sur deux échantillons indépendants sert à déterminer si les échantillons proviennent de la même population ou de deux populations différentes (voir le chapitre 10). Ce test présume que les données dans chaque

1. Cette formulation du coefficient de corrélation par rang de Spearman présume qu'il n'y a pas (ou très peu) de rangs qui soient ex æquo. Sinon, la formule traditionnelle pour le calcul du r de Pearson fournira une réponse plus juste.

Quiz rapide 13.4
Le problème suivant nous vient de la NASA. Vous faites partie d'une mission d'exploration voyageant d'une station spatiale qui est en orbite autour de la Lune vers la base qui se trouve sur la Lune. Une panne technique vous oblige à faire un atterrissage forcé, sur la face ensoleillée de la Lune, à 120 km de la base lunaire. Votre vaisseau étant très endommagé, vous devez vous rendre à pied à la base lunaire aussi vite que possible car votre survie en dépend. Quinze objets non endommagés à l'atterrissage d'urgence sont récupérables de votre vaisseau spatial. Ce sont : *une boîte d'allumettes ; de la nourriture sèche ;18 m de corde de nylon ; une toile de parachute ; une chaufferette solaire ; deux revolvers ; une boîte de lait en poudre ; deux réservoirs d'oxygène ; une carte lunaire ; un radeau autogonflable par sa cartouche de CO_2 ; un compas magnétique ; 15 litres d'eau ; des fusées de signalement ; une trousse de premiers soins ; un walkie-talkie à pile solaire.*

a) Classez ces 15 objets par ordre d'importance pour assurer votre survie (1 = le plus important et 15 = le moins important).

b) Demandez à un collègue de faire la même chose.

c) Comparer vos deux réponses à la solution produite par les experts de la NASA en faisant appel à la procédure statistique appropriée (la réponse de la NASA se trouve dans la réponse au Quiz rapide 13.4).

d) Qui a la meilleure chance de survie : votre collègue ou vous ?

échantillon sont mesurées sur une échelle de type II et que la distribution de la variable dans la population est à peu près normale. Ce test est donc inapproprié si, par exemple, on désire comparer le revenu moyen des hommes avec le revenu moyen des femmes, puisque la distribution des revenus, dans la population, est très asymétrique.

Une alternative au test t qui est appropriée pour cette situation est le test de Wilcoxon-Mann-Whitney (du nom de ses inventeurs). Pour être applicable, l'unique prérequis de ce test est qu'il soit possible de déterminer le rang des observations, nonobstant le groupe auquel chaque observation appartient. Si nous avons six femmes et six hommes, par exemple, ce test peut être appliqué à la condition que les 12 observations puissent être mises en rang, de 1 à 12, sans ex æquo.

La logique de ce test est élégante. Supposons que le revenu des femmes et des hommes soit similaire. Il devrait donc y avoir des individus (femmes et hommes) détenant un rang (salaire) élevé et des individus (femmes et hommes) détenant un rang (salaire) faible dans chaque groupe. Si on examine les rangs dans chaque groupe, on devrait donc y voir un nombre semblable de rangs élevés et de rangs faibles parmi les deux genres. Si on

additionne les rangs attribués aux femmes et ceux attribués aux hommes, ces deux sommes devraient être identiques lorsqu'il n'y a pas de différence entre le revenu des membres de ces deux groupes.

Le test non paramétrique Wilcoxon-Mann-Whitney permet de vérifier si c'est ou non le cas. Si cela est vrai, nous concluons à l'hypothèse nulle (les femmes et les hommes détiennent des niveaux de revenus équivalents). Sinon, l'hypothèse nulle est à rejeter et nous pouvons ainsi conclure que les deux groupes n'ont pas le même niveau de revenus.

Le Tableau 13.6 présente le salaire de N = 12 personnes, dont 6 sont des femmes (N_1 = 6) et 6 sont des hommes (N_2 = 6). Toutes ces personnes reçoivent des salaires différents (il n'y a pas d'ex æquo). Nous mettons en rang le salaire de ces douze personnes, sans nous préoccuper pour l'instant du groupe auquel elles appartiennent (femmes ou hommes). L'individu dont le salaire est le plus bas obtient le rang 1 alors que celui qui a le salaire le plus élevé se classe au douzième rang. L'addition des rangs (1 + 2 + 3 + … + 11 + 12) donne 78. Si les revenus des personnes dans les deux groupes sont comparables, les rangs devraient être répartis aléatoirement entre les deux groupes et, donc, la somme des rangs pour les femmes devrait être 39 (la moitié de 78) tout comme pour les hommes (39). L'encadré présente les formules qui facilitent le calcul de la somme des rangs.

Quelques formules de computation pour le Wilcoxon-Mann-Whitney

1. La somme de tous les rangs s'obtient avec la formule $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$, où N est le nombre d'observations totales. Avec 12 personnes, la somme de tous les rangs est $\sum_{i=1}^{N} i = \frac{12\,(13)}{2} = 78$.
2. La somme des rangs (sous H_0) dans un groupe lorsque les groupes contiennent un nombre identique d'observations = $\frac{N(N+1)}{4}$.
   Avec six personnes dans chaque groupe, la somme des rangs pour chaque groupe est = $\frac{12\,(13)}{4} = 39$.
3. La somme des rangs d'un groupe (sous H_0) lorsque le N des groupes est inégal s'obtient avec la formule $\frac{N_1(N+1)}{2}$ où N_1 est le nombre d'observations dans un groupe, N le nombre d'observations des deux groupes. Si N_1=4 et N_2=8, N=12 et la somme des rangs pour le groupe 1 est $\frac{4\,(13)}{2} = 26$.

L'hypothèse voulant que les groupes soient tirés de la même population de revenus revient donc à dire que la somme des rangs dans un groupe est égale à la moitié de la somme de tous les rangs (39 dans ce cas). Dans le cas contraire (différent de la moitié), on pourra conclure que le salaire des hommes et des femmes n'est pas le même.

Naturellement, la taille de la différence entre la somme des rangs pour les deux groupes dépend du nombre d'observations. Par conséquent, il faut standardiser cette différence afin d'éliminer l'influence du nombre de rangs sur le résultat. Plus bas, nous verrons la Formule 13.4 qui permet de réaliser cette standardisation. Cette formule, qui est le test non paramétrique Wilcoxon-Mann-Whitney semble assez complexe, mais sa logique est facile à saisir et à assimiler.

Comme nous en avons maintenant l'habitude, il nous faudra comparer la différence standardisée (qui sera produite par le biais de la Formule 13.4) à une valeur critique. Si la valeur standardisée que nous avons calculée est supérieure à la valeur critique, nous conclurons au rejet de l'hypothèse nulle, alors que si elle lui est égale ou inférieure, l'hypothèse nulle ne pourra être rejetée.

La valeur critique que nous utilisons pour ce test est très pratique. Elle se base sur le tableau de la distribution de la densité sous la courbe normale (Z, Annexe A1)[2]. Si la valeur Z produite par le Wilcoxon-Mann-Whitney est supérieure à 1,96, nous conclurons que la différence entre les deux groupes est statistiquement significative (au seuil alpha inférieur à 5 %). Si la valeur Z est supérieure à 2,58, la différence entre les deux groupes sera statistiquement significative au seuil $\alpha < 0{,}01$.

La standardisation s'obtient avec la Formule 13.4. Cette formulation exprime le test Wilcoxon-Mann-Whitney.

$$z = \frac{\left|SR_1 - N_1(N+1)/2\right| - 0{,}5}{\sqrt{N_1 N_2 (N+1)/12}} \qquad \text{Formule 13.4}$$

où :

- SR_1 est la somme des rangs dans le groupe 1 ;
- N_1 est la taille du groupe 1 ;

2. L'utilisation du tableau de la densité de la courbe normale est appropriée seulement si le nombre d'observations dans un groupe ou dans les deux groupes conjugués est égal ou plus grand que N = 10. Si N < 10, il faudra consulter un tableau spécialisé.

- N_2, la taille du groupe 2 ; et
- N est la taille totale des deux groupes (c.à.d. $N = N_1 + N_2$). Notez qu'il faut retirer 0,5 car les rangs peuvent être vus comme un nombre arrondi.

Le numérateur de la Formule 13.4 est essentiellement composé de deux quantités : SR_1, qui représente la somme des rangs attribués à l'un des groupes[3] et $N_1(N + 1) / 2$, qui représente la somme des rangs à laquelle nous pourrions nous attendre si l'hypothèse nulle était à retenir (c'est-à-dire que la somme des rangs attribués au groupe est exactement égale à la moitié de la somme de tous les rangs ou au prorata des rangs, lorsque le nombre d'observations dans les deux groupes est inégal). Nous voyons maintenant la logique du test : lorsque la somme des rangs obtenus dans un groupe (SR_1) est effectivement très proche de la moitié de la somme de tous les rangs, la soustraction donnera une différence très proche de zéro. Indépendamment du dénominateur, la quantité Z sera alors, elle aussi, proche de zéro. Une telle quantité ($z \approx 0$) étant inférieure à la valeur critique normalement acceptée pour la signification statistique ($z = 1{,}96$), nous ne pouvons pas rejeter l'hypothèse nulle et, dans notre exemple, il faudra alors conclure que le salaire des hommes et des femmes est équivalent. Dans cette formule, le dénominateur sert à établir la standardisation comme telle et c'est le rapport entre le numérateur et le dénominateur qui produit la valeur Z qui elle définit la taille standardisée de la différence entre les rangs obtenus par les deux groupes.

3. L'analyse ne porte que sur un seul groupe, peu importe lequel, car la somme des rangs de l'autre groupe est invariablement connue si l'on connaît la somme des rangs du groupe que l'on analyse. Par exemple, si la somme des rangs pour les hommes est égale à 21, la somme des rangs pour les femmes est obligatoirement 78 – 21 = 57. Si la somme des rangs pour un groupe est 39, l'autre obtiendra une somme de 39 aussi.

Un exemple du Wilcoxon-Mann-Whitney

Imaginons un monde idéal où il n'y aurait pas de différence entre les revenus des hommes et des femmes. On identifie le revenu de six femmes et de six hommes pris aléatoirement dans la population. Les données sont inscrites dans les colonnes 1 et 2 du Tableau 13.6.

Tableau 13.6
Deux exemples du test non paramétrique Wilcoxon-Mann-Whitney : les femmes et des hommes ont-ils le même salaire ?

Exemple 1		*Exemple 2*	
Col 1	*Col 2*	*Col 3*	*Col 4*
Femmes (K$)	*Hommes (K$)*	*Femmes (K$)*	*Hommes (K$)*
33 (2)	22 (1)	(1)	(7)
38 (3)	41 (4)	(2)	(8)
58 (6)	43 (5)	(3)	(9)
71 (7)	78 (8)	(4)	(10)
91 (10)	81 (9)	(5)	(11)
128 (11)	178 (12)	(6)	(12)

Le salaire est en milliers de dollars (K$) et le rang des douze observations est indiqué entre parenthèses.

Les revenus inscrits dans les colonnes 1 et 2 du Tableau 13.6 sont très asymétriques (c'est-à-dire ne sont pas distribués normalement), ce qui invite l'utilisation du Wilcoxon-Mann-Whitney. On calcule le rang des revenus (peu importe le sexe) que nous retrouvons entre parenthèses au Tableau 13.6. La personne qui gagne 22 000 $ par année (dans ce cas, il s'agit d'un homme) reçoit le rang 1 (le revenu le plus faible de tous) ; la personne (une femme dans ce cas) qui gagne 33 000 $ par année reçoit le rang 2, etc. L'hypothèse nulle prédit : H_0 : la somme des rangs des hommes est égale à $\frac{6(12+1)}{2} = 39$.
Et la règle est :

rejet de H_0 si $SR_{femmes} > SR_{Critique.}$

Mettons en œuvre le test statistique en faisant appel à la Formule 13.4.

$$Z = \frac{\left|SR_1 - N_1(N+1)/2\right| - 0{,}5}{\sqrt{N_1 N_2 (N+1)/12}}$$

$$Z = \frac{\left|39 - 6(13)/2\right| - 0{,}5}{\sqrt{6(6)(13)/12}} = -0{,}08$$

Nous pouvons maintenant tester l'hypothèse. Nous choisissons le seuil de signification $\alpha < 0{,}05$, qui, à partir du tableau de la densité sous la courbe normale, indique que la valeur critique est 1,96. Puisque le Z que nous avons calculé est –0,08 et que cette valeur est inférieure à 1,96, nous ne pouvons pas rejeter l'hypothèse nulle. Les hommes et les femmes ont donc des salaires équivalents. Cela ne devrait pas nous surprendre puisque la somme des rangs pour les hommes est égale, dans ce cas, à la moitié de la somme des rangs pour l'ensemble des données.

Refaisons le même exercice mais cette fois référons-nous aux données qui se trouvent dans les colonnes 3 et 4 du Tableau 13.6. Le tableau ne présente que les rangs et on constatera que ceux des femmes sont tous inférieurs aux rangs obtenus par les hommes (toutes les femmes sont moins payées que tous les hommes). Clairement, nous devons nous attendre à une différence statistiquement significative. Vérifions si cela est vrai.

D'abord, calculons la quantité SR_1 qui est la somme des rangs associés aux femmes : 1 + 2 + 3 + 4 + 5 + 6 = 21

$$Z = \frac{\left|SR_1 - N_1(N+1)/2\right| - 0{,}5}{\sqrt{N_1 N_2 (N+1)/12}}$$

$$Z = \frac{\left|21 - 6(13)/2\right| - 0{,}5}{\sqrt{6(6)(13)/12}} = -2{,}96$$

Le Z vaut 2,96. Cette valeur étant plus élevée que la valeur critique Z = 1,96, nous concluons à la signification statistique[4]. Cette conclusion est juste compte tenu que nous acceptons un risque d'erreur alpha de moins de 5 %. Donc, il existe bien une différence entre le salaire des femmes et des hommes.

4. Le signe est sans conséquence car il est déterminé par l'ordre d'entrée des données. On interprète le Z obtenu en fonction de sa taille et non pas de son signe.

Quiz rapide 13.5
Revenez à l'analyse portant sur les colonnes 3 et 4 du Tableau 13.6. L'un des auteurs du livre est en accord avec la conclusion voulant que les femmes et les hommes n'aient pas le même salaire, mais il croit que la conclusion est fausse quand le risque alpha est inférieur à 1 % au lieu du 5 % initialement présumé. A-t-il raison ?

Un exemple plus complexe du Wilcoxon-Mann-Whitney

Reprenons l'exemple ci-dessus avec des données plus réalistes. Ici, nous sommes confrontés à des groupes inégaux et, malheureusement, notre banque de données inclut des rangs ex æquo (tout va mal cette fois !). Les données sont :

- revenus des hommes en K$ (N_1) : 22, 33, 44, 57, 59, 71, 102, 111, 128, 178
- revenus des femmes en K$ (N_2) : 33, 36, 42, 56, 69, 70

où $N_1 = 10$, $N_2 = 6$ et $N = 16$. Après une conversion en rangs, on obtient :

- rangs des revenus des hommes (N = 10) : 1, 2.5, 6, 8, 9, 12, 13, 14, 15, 16 ($\Sigma = 96{,}5$)
- rangs des revenus des femmes (N = 6) : 2.5, 4, 5, 7, 10, 11 ($\Sigma = 39{,}5$).

Le revenu le plus faible est 22 K$ et nous attribuons le rang 1 à cette personne (un homme). Le revenu suivant, 33 000 $, est celui de deux personnes, un homme et une femme. Ces revenus étant ex æquo, nous attribuons le même rang (2,5) à cet homme et à cette femme. Nous poursuivons cette attribution jusqu'au dernier rang (16). La somme des rangs des hommes est égale à 96,5, celle des rangs des femmes à 39,5 et, au total, la somme de tous les rangs est 39,5 + 96,5 = 136. On peut vérifier que le total donne bien N(N + 1) / 2, soit 136. Comme le groupe des hommes est plus nombreux ($N_1 = 10$) que celui des femmes ($N_2 = 6$), il est entendu que la somme des rangs est plus grande pour le groupe des hommes (96,5) que pour celui des femmes (39,5). Dans ce cas, nous choisissons de travailler avec le groupe des hommes. L'hypothèse est :

H_1 : les hommes et les femmes n'ont pas le même revenu

H_0 : les hommes et les femmes ont les mêmes revenus.

La somme des rangs pour les hommes devrait être $\frac{10(16 + 1)}{2} = 85$.

La règle décisionnelle, pour un seuil de 5 %, est

rejet de H_0 si $Z > 1{,}96$.

Calculons Z :

$$z = \frac{\left|SR_1 - N_1(N+1)/2\right| - 0{,}5}{\sqrt{N_1 N_2 (N+1)/12}}$$

$$= \frac{\left|96{,}5 - 85\right| - 0{,}5}{\sqrt{10 \times 6(16+1)/12}}$$

$$= \frac{11{,}5 - 0{,}5}{\sqrt{1020/12}}$$

$$= \frac{11}{\sqrt{85}} = \frac{11}{9{,}21} = 1{,}19$$

Comme cette valeur n'est pas supérieure à la valeur critique (1,96), ces données ne permettent pas de rejeter l'hypothèse nulle. Il serait faux de conclure que les hommes et les femmes ont des salaires différents.

Quiz rapide 13.6
Plutôt que de considérer le rang des hommes pour réaliser le test, **dans la section « un exemple plus complexe du Wilcoxon-Mann-Whitney »**, pourriez-vous faire le test en considérant le rang des femmes ? Si oui, obtiendriez-vous le même résultat ? Sinon, pourquoi ?

UN TEST SUR K ÉCHANTILLONS INDÉPENDANTS

S'il y a plus de deux groupes indépendants à comparer, il faut appliquer un test tel que l'analyse de la variance à un facteur. Or, l'ANOVA, tout comme le test t, nécessite que les mesures soient obtenues sur une échelle de type II et, de plus, que la distribution de la population soit à peu près normale. Si l'un ou l'autre de ces prérequis n'est pas satisfait, il faut utiliser un test alternatif. Le test non paramétrique de Kruskal-Wallis sur plusieurs (k) échantillons indépendants est une alternative à l'ANOVA. On se souviendra que l'ANOVA est une généralisation du test t. De la même manière, le Kruskal-Wallis est une généralisation du Wilcoxon-Mann-Whitney

Le Kruskal-Wallis débute, tout comme le Wilcoxon-Mann-Whitney, par le remplacement des valeurs originales par des rangs. Nous établissons d'abord le rang de toutes les observations (sans tenir compte du groupe auquel chaque observation appartient). Puis nous calculons, pour chaque groupe, la somme des rangs. Par la suite, il faut mettre ces sommes au carré

et calculer une statistique g. Enfin, la valeur g que nous avons calculée est maintenant comparée à une valeur critique ($g_{critique}$). Cette valeur critique est pratique puisqu'il s'agit de la valeur critique du χ^2. Si g est plus grand que la valeur critique du χ^2, on rejette l'hypothèse nulle voulant que tous les groupes proviennent de la même population. Sinon, nous n'avons pas de base pour la rejeter.

La statistique g s'obtient avec la Formule 13.5 :

$$g = \frac{\sum_{i=1}^{k} \frac{SR_i^2}{N_i}}{N(N+1)/12} - 3(N+1) \qquad \text{Formule 13.5}$$

où :

SR_i est la somme des rangs du i^e groupe ;

N_i est la taille du i^e groupe ;

N est le nombre total d'observations dans tous les groupes (c.à.d. $N = N_1 + N_2 + \ldots + N_k$) ; et

k est le nombre de groupes.

Dans le Wilcoxon-Mann-Whitney, nous retirons une constante (dans ce cas 0,5). Pour le calcul du g de Kruskal-Wallis, nous retirons aussi une quantité qui, dans ce cas, s'obtient par l'expression 3 (N + 1).

La règle de décision est

$$\text{rejet de } H_0 \text{ si } g > g_{critique}.$$

Si tous les groupes ont plus de cinq observations, la valeur critique à laquelle notre statistique g sera comparée s'obtiendra directement dans le tableau des valeurs critiques de la distribution qui se trouve dans l'Annexe A4[5]. Comme nous en avons maintenant l'habitude, pour utiliser les tableaux des valeurs critiques, il faut établir le nombre de degrés de liberté et choisir le seuil de signification alpha. Le nombre de degrés de liberté est établi en appliquant la logique habituelle : la somme des rangs pour un des groupes est parfaitement déterminée par la somme des rangs obtenus dans les autres groupes. Par conséquent, les degrés de liberté s'obtiennent par le nombre de groupes moins un (K − 1). Par exemple, si nous analysons la différence entre trois groupes [dl = (K − 1) = (3 − 1) =2] et que nous

5. Si un des groupes est très petit (cinq observations ou moins pour ce test), il faut chercher la valeur critique dans une table spécialisée.

testons la signification avec un seuil alpha de 5 %, nous trouvons la valeur critique $_{\text{critique}}$= 9,488 à l'intersection de la rangée ν (nu) = 2 et α = 0,05 au Tableau A4. Si la statistique g que nous avons obtenue avec la Formule 13.5 est supérieure à 9,488, nous concluons au rejet de H_0.

Un exemple du test non paramétrique Kruskal-Wallis

Dans une entreprise, les employés doivent se servir d'une certaine machine pour exécuter le travail. Traditionnellement, la formation des employés à l'utilisation de la machine est assurée par le superviseur immédiat. Disons que c'est la formation « interne ». Par contraste, certains gestionnaires pensent qu'il serait préférable que cette formation soit assurée par une entreprise externe, spécialisée en formation. Appelons ça la formation « externe ». Enfin, d'autres gestionnaires pensent que la formation est un gaspillage d'argent et qu'il n'est pas véritablement requis de fournir un programme de formation, les employés pouvant maîtriser la machine sans cours formel. Appelons ça « pas de formation ». Face à ces trois choix, les gestionnaires décident de faire une expérience afin de déterminer qui a raison.

Trois groupes de nouveaux employés sont choisis : chaque groupe est composé de six employés (k = 3, $N_1 = N_2 = N_3 = 6$ et N = 18). Le premier groupe ne reçoit pas de formation (N_1), le deuxième (N_2) reçoit la formation interne, et le dernier (N_3) reçoit la formation externe. À la fin de la formation, on mesure pour chaque employé sa performance que nous mettons en rang. Dans ces conditions, l'analyse statistique des résultats se base sur la statistique g du test de Kruskal-Wallis. Le rendement de chaque employé est inscrit au Tableau 13.7, qui présente aussi (entre parenthèses) le rang de chacun.

D'abord, nous calculons la somme des rangs pour chaque groupe (SR_i) que nous mettons au carré (SR_i^2). Nous divisons chacune de ces quantités par le nombre d'observations dans chaque groupe (SR_i^2/N_i). Dans ce cas, toutes les quantités sont divisées par 6, le nombre d'employés dans chaque groupe. Pour obtenir le numérateur de la Formule 13.5, nous faisons la somme des trois quantités ($SR_1^2/N_1 + SR_2^2/N_2 + SR_3^2/N_3$). La somme des rangs est pour le groupe 1 : $SR_1 = 25$, pour le groupe 2 : $SR_2 = 65$, et pour le groupe 3 : $SR_3 = 81$. Nous pouvons maintenant calculer le numérateur de la Formule 13.5, la sommation $\sum_{i=1}^{3} \frac{SR_i^2}{N_i}$.

Tableau 13.7
Le succès relatif de trois techniques de formation : le Kruskal-Wallis

Groupe 1 (sans formation)	*Groupe 2 (formation interne)*	*Groupe 3 (formation externe)*
21 (1)	48 (7)	34 (4)
28 (2)	64 (9)	73 (12)
32 (3)	70 (10)	76 (14)
44 (5,5)	71 (11)	80 (17)
44 (5,5)	75 (13)	80 (17)
54 (8)	77 (15)	80 (17)

$$\sum_{i=1}^{3} \frac{SR_i^2}{N_i} = \frac{SR_1^2}{N_1} + \frac{SR_2^2}{N_2} + \frac{SR_3^2}{N_3}$$
$$= \frac{25^2}{6} + \frac{65^2}{6} + \frac{81^2}{6}$$
$$= \frac{625}{6} + \frac{4225}{6} + \frac{6561}{6}$$
$$= 104{,}16 + 704{,}16 + 1093{,}5$$
$$= 1901{,}8$$

Le numérateur étant obtenu, il est enfin possible de calculer la statistique g de Kruskal-Wallis avec la Formule 13.5.

$$= \frac{\sum_{i=1}^{k} \frac{SR_i^2}{N_i}}{N(N+1)/12} - 3(N+1)$$

$$= \frac{1901{,}8}{\left(\frac{18\,x\,19}{12}\right)} - 3\,x19$$

$$= 9{,}73$$

Le g de Kruskal-Wallis est 9,7. Nous avons trois groupes et donc nous nous retrouvons avec K – 1 degrés de liberté, dl = (3 – 1) = 2. Nous vérifions notre hypothèse en nous référant au tableau de la distribution des valeurs critiques du χ^2 qui est (Annexe A4), dans ce cas, 9,488. Puisque la statis-

tique g obtenue (9,7) est supérieure à la valeur critique (9,488), nous rejetons l'hypothèse nulle et concluons que les trois programmes de formation ne produisent pas des résultats équivalents.

LE TEST DE WILCOXON SUR DONNÉES APPAREILLÉES

Nous terminons ce chapitre en présentant l'analyse non paramétrique requise lorsque nous travaillons avec des données appareillées. Celles-ci sont souvent utilisées en pratique lorsque l'on veut vérifier le changement : par exemple, si à la suite d'un traitement médical ou psychologique, l'état de santé des patients s'est amélioré. Nous voulons ainsi déterminer si la différence entre une mesure prise avant et après sur le même groupe de personnes est statistiquement significative. Ainsi, nous allons comparer deux échantillons, mais puisqu'ils sont tous deux composés des mêmes individus, ils ne sont pas indépendants.

Nous avons déjà étudié ce problème dans le chapitre portant sur le test t pour données appareillées (chapitre 10). Mais, on se souviendra que le test t exige que la variable à l'étude soit une variable de type II et que les données soient extraites d'une population normale. Lorsque ces exigences ne peuvent être respectées, le test non paramétrique de Wilcoxon est la forme d'analyse statistique appropriée. Le test de Wilcoxon nécessite uniquement que l'on puisse calculer l'écart qui existe pour chaque observation entre sa paire de données.

Pour réaliser ce test, il faut premièrement calculer l'écart entre les mesures. Certains écarts seront positifs, d'autres négatifs. Par exemple, à la suite d'une intervention psychologique, la santé mentale de certaines personnes pourrait s'améliorer (écart positif), elle pourrait se détériorer (écart négatif), ou elle pourrait demeurer inchangée (écart nul). Certains écarts pourraient être numériquement importants alors que d'autres pourraient être faibles. En ignorant le signe des écarts, on peut calculer leurs rangs en fonction de la taille de l'écart. L'écart le plus petit obtiendrait le rang 1, et l'écart le plus grand obtiendrait le rang N. Si les deux mesures proviennent de la même population (H_0), le total des rangs d'écarts positifs devrait être semblable au total des rangs d'écarts négatifs, et valoir la moitié de la somme de tous les rangs.

Le jeu d'hypothèses suit l'approche traditionnelle.

H_1 : la somme des rangs ayant un écart positif n'est pas égale à la moitié des rangs.

H_0 : la somme des rangs des gens ayant un écart positif vaut la moitié de la somme totale des rangs, soit $\frac{N(N+1)}{2}$ et N est le nombre de paires d'observations.

Ainsi, dans notre exemple, si l'intervention psychologique n'a pas d'effet, les améliorations devraient être égales aux détériorations dans la santé mentale des patients.

La règle de décision est :

$$\text{rejet de } H_0 \text{ si } SR_+ > SR_{\text{Critique}}$$

où SR_+ est la somme des rangs pour ceux ayant un écart positif.

Lorsque nous avons plus que 15 observations pairées (N > 15), nous pouvons standardiser SR_+ avec la Formule 13.6 qui définit le Wilcoxon :

$$z = \frac{\left|SR_+ - N(N+1)\,/\,4\right|}{\sqrt{N(N+1)(2N+1)\,/\,24}} \qquad \text{Formule 13.6}$$

Vous remarquerez à la Formule 13.6 que la standardisation se fait par l'entremise d'une valeur Z. Par conséquent, la valeur critique à laquelle notre résultat sera comparé proviendra du tableau de densité sous la courbe normale (Annexe A1)[6].

La règle de décision devient :

$$\text{rejet de } H_0 \text{ si } z > z_{\text{critique}}$$

où la valeur critique est tirée d'une table normale standardisée telle que celle de l'Annexe A1.

Un exemple du test Wilcoxon pour des données appareillées

À la suite d'une crise financière, un grand nombre d'investisseurs à la Bourse tombent en dépression. Pour réduire leurs symptômes et peut-être relancer l'économie, une banque centrale injecte un giga milliard de dollars

6. Lorsque N ≤ 15, il faut consulter une table spécialisée pour trouver la valeur critique (qui dépend de ce N et du seuil de décision alpha).

dans les marchés boursiers. Cette banque, qui veut vérifier si le moral de ces investisseurs s'est amélioré, a demandé à un échantillon de 16 investisseurs combien ils croyaient gagner dans l'année à venir avant l'annonce de l'injection. Puis elle repose cette même question après l'annonce. Le Tableau 13.8 montre les résultats découlant de cette collecte de données pour chacun de ces investisseurs. La troisième colonne du Tableau 13.8 calcule la différence entre la mesure pré-injection et la mesure post-injection de fonds. Un signe négatif dans cette colonne indique que l'investisseur croit qu'il aura des pertes alors qu'un signe positif indique qu'il estime qu'il fera des profits.

Tableau 13.8
Profit anticipé par des investisseurs avant et après une injection massive de fonds dans les marchés boursiers
Le test de Wilcoxon pour des données appareillées

Profit anticipé avant l'injection $	*Profit anticipé après l'injection $*	*Écart (le signe est entre parenthèses)*	*Rang absolu de la différence*
–1 000 000	1 000 000	2 000 000 (+)	12
–500 000 000	500 000 000	1 000 000 000 (+)	16
1	10 000 000	9 999 999 (+)	13
200 000	180 000	20 000 (–)	5
–100 000 000	100 000 000	200 000 000 (+)	15
1 000 000	999 999	1 (–)	1
0	–100	100 (–)	2
100	–500	600 (–)	3
600 000	500 000	100 000 (–)	7
0	10 000 000	10 000 000 (+)	14
100	1 000 000	999 900 (+)	10
1 000	2 000 000	1 999 999 (+)	11
500	10 000	9 500 (+)	4
1 000	100 000	99 000 (+)	6
5 000	1 000 000	995 000 (+)	9
10 000	1 000 000	990 000 (+)	8

Ces mesures, qui sont des évaluations subjectives, ne suivent pas du tout une distribution normale et donc l'approche non paramétrique de Wilcoxon est indiquée.

L' écart négatif indique que l'investisseur croit qu'il fera moins d'argent après l'injection d'argent. Nous allons calculer le signe de l'écart, aussi bien que sa taille (c'est-à-dire l'écart sans le signe qui se trouve dans la troisième colonne). Enfin, nous établirons le rang de la différence entre ces derniers (la quatrième colonne du Tableau 13.8). Nous prenons en considération ceux qui s'attendent à une amélioration (c'est-à-dire ceux dont l'écart est associé à un signe + au Tableau 13.8).

Les rangs pour ceux qui s'attendent à une amélioration (c'est-à-dire ceux qui ont le signe +) sont :

12, 16, 13, 15, 14, 10, 11, 4, 6, 9, 8 ($\Sigma = 118$)

et la somme de ces rangs, SR_+, vaut 118. Par comparaison, la somme de tous les rangs de 1 à 16 vaut N (N + 1)/2, soit 136 ; 118 est loin de la moitié.

Puisqu'il y a plus de 15 données, on procède à la standardisation (Formule 13.6) :

$$
\begin{aligned}
z &= \frac{\left|SR_+ - N(N+1)\,/\,4\right|}{\sqrt{N(N+1)(2N+1)\,/\,24}} \\
&= \frac{\left|118 - 16(16+1)\,/\,4\right|}{\sqrt{16(16+1)(2\times 16+1)\,/\,24}} \\
&= \frac{\left|118 - 16\times 17\,/\,4\right|}{\sqrt{16\times 17\times 33\,/\,24}} \\
&= \frac{\left|118 - 68\right|}{\sqrt{374}} \\
&= \frac{50}{19{,}3} \\
&= 2{,}59
\end{aligned}
$$

Cette valeur étant supérieure à la valeur critique habituelle (Z = 1,96), nous rejetons l'hypothèse nulle, en acceptant un risque d'erreur de type I qui est inférieur à 5 % (voir l'Annexe A1). De toute évidence, renflouer les marchés boursiers à coups de giga dollars améliore l'état d'âme des investisseurs !

Quiz rapide 13.7
Plutôt que de travailler sur les rangs ayant des signes +, refaites le test en considérant cette fois-ci les rangs ayant des signes négatifs. Obtenez-vous un résultat similaire ? Est-ce logique d'obtenir ce résultat ?

SOMMAIRE DU CHAPITRE

Les analyses paramétriques exigent que les données soient de type II (intervalle ou de rapport) et que les variables en jeu soient extraites d'une population distribuée normalement. Lorsque nous ne pouvons respecter l'une ou l'autre de ces conditions — car nous travaillons avec une variable qui est non normale dans la population ou qui est mesurée sur une échelle de type I (nominale ou ordinale) —, nous devons alors faire appel aux procédures non paramétriques. Ces analyses non paramétriques sont le pendant des analyses paramétriques telles que la corrélation, le test t et l'analyse de variance. Elles servent à établir la corrélation entre deux variables ou la différence entre deux ou plusieurs groupes, qu'ils proviennent ou non d'échantillons indépendants. Comme pour l'ensemble des statistiques décrites dans ce livre, l'inférence statistique consiste en une comparaison du résultat obtenu avec une valeur critique tabulée, généralement un χ^2 ou le tableau de la densité sous la courbe normale (Z).

À l'exception du χ^2 — qui teste des variables nominales —, les analyses non paramétriques décrites dans le chapitre exigent que les données à être traitées soient ordinales (des rangs).

Le test du chi deux (χ^2) sert à déterminer si la taille des effectifs est différente pour les diverses catégories d'une variable nominale. Ce test établit la différence entre les fréquences observées dans les catégories et la taille des effectifs qui devraient exister s'il n'y avait pas de différence entre les effectifs de ces catégories.

Pour la corrélation de Spearman, le test vérifie le degré avec lequel le rang obtenu par chaque observation sur une variable est maintenu sur une deuxième.

Quant aux analyses qui vérifient la différence entre les groupes, la logique des diverses techniques est très similaire : nous mettons en rang toutes

les observations (sans prendre en considération le groupe dont elles proviennent). Puis nous faisons la sommation des rangs qu'occupent les membres d'un des groupes. Lorsque cette sommation est assurément différente que celle attendue, si les rangs étaient également répartis entre les groupes, nous pouvons conclure au rejet de l'hypothèse nulle.

EXERCICES DE COMPRÉHENSION

1. Si les données sont déjà des rangs, mais par rapport aux autres personnes du groupe, est-ce possible de réaliser un test comme le Wicoxon-Mann-Whitney ?
2. Dans quelle situation peut-on faire un test de Kruskal-Wallis ?
   a) Quand les données ne sont pas normalement distribuées.
   b) Quand les données ne sont pas de type II.
   c) Quand les données ne sont pas continues mais distribuées normalement.
   d) a ou b
   e) a ou c
3. Si l'on veut obtenir des données sur le QI avant et après une cuite, quel test peut-on utiliser ?
   a) Un test de Kruskal-Wallis.
   b) Un test de Wilcoxon-Mann-Whitney.
   c) Un test de chi deux.
   d) Une corrélation de Spearman.
   e) Aucune de ces réponses.
4. Si vous avez quatre groupes de huit personnes chacun, quelle est la somme des rangs à laquelle on s'attend dans le premier groupe dans le cas d'une hypothèse nulle ?
   a) 124
   b) 66
   c) 132
   d) 100

Les questions 5, 6 et 7 se rapportent à ce problème

Le directeur d'une clinique externe doit conjuguer avec quatre formes de maladie mentale (A, B, C et D), chacune exigeant d'être traitée par un psychiatre de spécialité différente. La clinique a traité le nombre suivant de patients pour chaque maladie le mois précédent : maladie A = 15 patients ; maladie B = 25 patients ; maladie C = 11 patients ; maladie D = 29 patients. Le directeur décide d'augmenter le nombre de médecins spécialistes afin d'assurer un rapport spécialiste/type de maladie égal pour les quatre types de patient.

5. Quelle est la forme statistique requise dans cette situation ?
   a) Analyse de variance.
   b) Chi deux.
   c) Kruskal-Wallis.
   d) Wilcoxon.
   e) Spearman.

6. Dans le cas d'une hypothèse nulle, combien de patients chaque maladie affligera-t-elle ?
   a) 20
   b) entre 11 et 29
   c) plus de 11 mais moins de 29
   d) 15, 25, 11 et 29 pour les maladies A à D respectivement
   e) Impossible à dire avec les informations disponibles.

7. En vous basant sur votre analyse de ces données, le directeur devrait-il rejeter l'hypothèse nulle ?

8. À la suite de votre analyse, apportez la conclusion qui s'impose. Le directeur devrait engager _____.
   a) moins de spécialistes pour la maladie C
   b) deux fois plus de médecins pour la maladie D que pour la maladie A
   c) e même nombre de médecins pour les maladies A et C et plus de médecins pour les maladies B et D
   d) un nombre égal de médecins spécialisés pour chaque type de maladie

9. Maintenant, c'est au tour du directeur d'un hôpital d'avoir un problème. Il vérifie le nombre de jours que les patients accablés des maladies A, B, C et D sont hospitalisés. Il souhaite déterminer s'il doit accroître le nombre de lits disponibles pour chaque type de maladie. Ses dossiers indiquent le nombre de jours d'hospitalisation pour chaque patient, et pour chaque type de maladie, durant la dernière année. Le nombre de jours d'hospitalisation varie entre 0 et 30. Quel type d'analyse statistique est approprié dans ce cas ?

a) Analyse de variance.
b) Chi deux.
c) Kruskal-Wallis.
d) Wilcoxon.
e) Spearman.

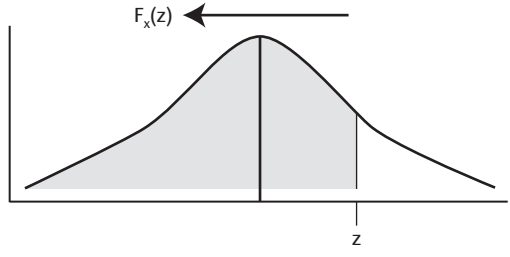
Réponses

1. Non, ce n'est pas possible, car il faut recoder les données pour connaître le rang de chaque score si toutes les données sont dans un seul groupe. Comme on ignore si la personne qui occupe le rang 2 dans le premier groupe est meilleure ou pire que la personne qui occupe le rang 2 dans le second groupe, il n'est pas possible de recoder ces scores.
2. d (Notez que l'alternative c est impossible : une variable non continue ne peut pas être distribuée normalement.)
3. e
4. c
5. b
6. a (Notez que ces données proviennent d'une échelle nominale, la fréquence.)
7. Le $\chi^2_{observé} = 10{,}6$. (Au Tableau A4, la valeur critique du χ^2 pour dl = 3 et $\alpha = 0{,}05$ est 12,59. Le $\chi^2_{observé}$ est inférieur à la valeur critique. Non, nous ne pouvons pas rejeter H_0.)
8. d (H_0 ne pouvant pas être rejeté, le directeur n'a pas de base valide pour déterminer qu'une maladie survient plus fréquemment qu'une autre.)
9. a (Notez que le nombre de jours est une variable de rapport — il est possible d'avoir 0 jour d'hospitalisation. Dans ce cas, il est possible de présumer que la distribution est normale, ce qui indique qu'une analyse paramétrique est de rigueur. Nous avons quatre groupes, ce qui montre que l'ANOVA est la bonne technique à appliquer.)

ANNEXE

Page laissée blanche

Tableau A.1
Tableau de la densité sous la courbe normale

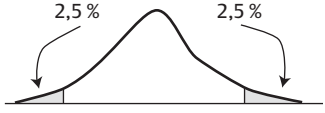


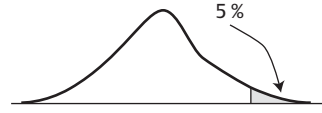
z	$F_X(z)$	z	$F_X(z)$	z	$F_X(z)$	z	$F_X(z)$
0,00	0,5000	**1,00**	0,8413	**2,00**	0,9772	**3,00**	0,9987
0,02	0,5080	**1,02**	0,8461	**2,02**	0,9783	**3,02**	0,9987
0,04	0,5160	**1,04**	0,8508	**2,04**	0,9793	**3,04**	0,9988
0,06	0,5239	**1,06**	0,8554	**2,06**	0,9803	**3,06**	0,9989
0,08	0,5319	**1,08**	0,8599	**2,08**	0,9812	**3,08**	0,9990
0,10	0,5398	**1,10**	0,8643	**2,10**	0,9821	**3,10**	0,9990
0,12	0,5478	**1,12**	0,8686	**2,12**	0,9830	**3,12**	0,9991
0,14	0,5557	**1,14**	0,8729	**2,14**	0,9838	**3,14**	0,9992
0,16	0,5636	**1,16**	0,8770	**2,16**	0,9846	**3,16**	0,9992
0,18	0,5714	**1,18**	0,8810	**2,18**	0,9854	**3,18**	0,9993
0,20	0,5793	**1,20**	0,8849	**2,20**	0,9861	**3,20**	0,9993
0,22	0,5871	**1,22**	0,8888	**2,22**	0,9868	**3,22**	0,9994
0,24	0,5948	**1,24**	0,8925	**2,24**	0,9875	**3,24**	0,9994
0,26	0,6026	**1,26**	0,8962	**2,26**	0,9881	**3,26**	0,9994
0,28	0,6103	**1,28**	0,8997	**2,28**	0,9887	**3,28**	0,9995
0,30	0,6179	**1,30**	0,9032	**2,30**	0,9893	**3,30**	0,9995
0,32	0,6255	**1,32**	0,9066	**2,32**	0,9898	**3,32**	0,9995
0,34	0,6331	**1,34**	0,9099	**2,34**	0,9904	**3,34**	0,9996
0,36	0,6406	**1,36**	0,9131	**2,36**	0,9909	**3,36**	0,9996
0,38	0,6480	**1,38**	0,9162	**2,38**	0,9913	**3,38**	0,9996
0,40	0,6554	**1,40**	0,9192	**2,40**	0,9918	**3,40**	0,9997
0,42	0,6628	**1,42**	0,9222	**2,42**	0,9922	**3,42**	0,9997
0,44	0,6700	**1,44**	0,9251	**2,44**	0,9927	**3,44**	0,9997

Tableau A.1
Tableau de la densité sous la courbe normale (suite)

0,46	0,6772	**1,46**	0,9279	**2,46**	0,9931	**3,46**	0,9997
0,48	0,6844	**1,48**	0,9306	**2,48**	0,9934	**3,48**	0,9997
0,50	0,6915	**1,50**	0,9332	**2,50**	0,9938	**3,50**	0,9998
0,52	0,6985	**1,52**	0,9357	**2,52**	0,9941	**3,52**	0,9998
0,54	0,7054	**1,54**	0,9382	**2,54**	0,9945	**3,54**	0,9998
0,56	0,7123	**1,56**	0,9406	**2,56**	0,9948	**3,56**	0,9998
0,58	0,7190	**1,58**	0,9429	**2,58**	0,9951	**3,58**	0,9998
0,60	0,7257	**1,60**	0,9452	**2,60**	0,9953	**3,60**	0,9998
0,62	0,7324	**1,62**	0,9474	**2,62**	0,9956	**3,62**	0,9999
0,64	0,7389	**1,64**	0,9495	**2,64**	0,9959	**3,64**	0,9999
0,66	0,7454	**1,66**	0,9515	**2,66**	0,9961	**3,66**	0,9999
0,68	0,7517	**1,68**	0,9535	**2,68**	0,9963	**3,68**	0,9999
0,70	0,7580	**1,70**	0,9554	**2,70**	0,9965	**3,70**	0,9999
0,72	0,7642	**1,72**	0,9573	**2,72**	0,9967	**3,72**	0,9999
0,74	0,7704	**1,74**	0,9591	**2,74**	0,9969	**3,74**	0,9999
0,76	0,7764	**1,76**	0,9608	**2,76**	0,9971	**3,76**	0,9999
0,78	0,7823	**1,78**	0,9625	**2,78**	0,9973	**3,78**	0,9999
0,80	0,7881	**1,80**	0,9641	**2,80**	0,9974	**3,80**	0,9999
0,82	0,7939	**1,82**	0,9656	**2,82**	0,9976	**3,82**	0,9999
0,84	0,7995	**1,84**	0,9671	**2,84**	0,9977	**3,84**	0,9999
0,86	0,8051	**1,86**	0,9686	**2,86**	0,9979	**3,86**	0,9999
0,88	0,8106	**1,88**	0,9699	**2,88**	0,9980	**3,88**	0,9999
0,90	0,8159	**1,90**	0,9713	**2,90**	0,9981	**3,90**	1,0000
0,92	0,8212	**1,92**	0,9726	**2,92**	0,9982	**3,92**	1,0000
0,94	0,8264	**1,94**	0,9738	**2,94**	0,9984	**3,94**	1,0000
0,96	0,8315	**1,96**	0,9750	**2,96**	0,9985	**3,96**	1,0000
0,98	0,8365	**1,98**	0,9761	**2,98**	0,9986	**3,98**	1,0000

Tableau A.2
Distribution des valeurs critiques de la statistique t





Hypothèse bicaudale (non directionnelle)				*Hypothèse unicaudale (directionnelle)*			
dl	**Seuil alpha (p <)**			***dl***	**Seuil alpha (p <)**		
	0,05	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**
1	12,706	63,656	636,578	**1**	6,314	31,821	318,289
2	4,303	9,925	31,600	**2**	2,920	6,965	22,328
3	3,182	5,841	12,924	**3**	2,353	4,541	10,214
4	2,776	4,604	8,610	**4**	2,132	3,747	7,173
5	2,571	4,032	6,869	**5**	2,015	3,365	5,894
6	2,447	3,707	5,959	**6**	1,943	3,143	5,208
7	2,365	3,499	5,408	**7**	1,895	2,998	4,785
8	2,306	3,355	5,041	**8**	1,860	2,896	4,501
9	2,262	3,250	4,781	**9**	1,833	2,821	4,297
10	2,228	3,169	4,587	**10**	1,812	2,764	4,144
11	2,201	3,106	4,437	**11**	1,796	2,718	4,025
12	2,179	3,055	4,318	**12**	1,782	2,681	3,930
13	2,160	3,012	4,221	**13**	1,771	2,650	3,852
14	2,145	2,977	4,140	**14**	1,761	2,624	3,787
15	2,131	2,947	4,073	**15**	1,753	2,602	3,733
16	2,120	2,921	4,015	**16**	1,746	2,583	3,686
17	2,110	2,898	3,965	**17**	1,740	2,567	3,646
18	2,101	2,878	3,922	**18**	1,734	2,552	3,610
19	2,093	2,861	3,883	**19**	1,729	2,539	3,579
20	2,086	2,845	3,850	**20**	1,725	2,528	3,552
21	2,080	2,831	3,819	**21**	1,721	2,518	3,527
22	2,074	2,819	3,792	**22**	1,717	2,508	3,505
23	2,069	2,807	3,768	**23**	1,714	2,500	3,485
24	2,064	2,797	3,745	**24**	1,711	2,492	3,467

Tableau A.2
Distribution des valeurs critiques de la statistique t (suite)

Hypothèse bi caudale (non directionnelle)				*Hypothèse uni caudale (directionnelle)*			
dl	**Seuil alpha (p <)**			***dl***	**Seuil alpha (p <)**		
	0,05	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**
25	2,060	2,787	3,725	**25**	1,708	2,485	3,450
26	2,056	2,779	3,707	**26**	1,706	2,479	3,435
27	2,052	2,771	3,689	**27**	1,703	2,473	3,421
28	2,048	2,763	3,674	**28**	1,701	2,467	3,408
29	2,045	2,756	3,660	**29**	1,699	2,462	3,396
30	2,042	2,750	3,646	**30**	1,697	2,457	3,385
32	2,037	2,738	3,622	**32**	1,694	2,449	3,365
34	2,032	2,728	3,601	**34**	1,691	2,441	3,348
36	2,028	2,719	3,582	**36**	1,688	2,434	3,333
38	2,024	2,712	3,566	**38**	1,686	2,429	3,319
40	2,021	2,704	3,551	**40**	1,684	2,423	3,307
45	2,014	2,690	3,520	**45**	1,679	2,412	3,281
50	2,009	2,678	3,496	**50**	1,676	2,403	3,261
55	2,004	2,668	3,476	**55**	1,673	2,396	3,245
60	2,000	2,660	3,460	**60**	1,671	2,390	3,232
65	1,997	2,654	3,447	**65**	1,669	2,385	3,220
70	1,994	2,648	3,435	**70**	1,667	2,381	3,211
75	1,992	2,643	3,425	**75**	1,665	2,377	3,202
80	1,990	2,639	3,416	**80**	1,664	2,374	3,195
85	1,988	2,635	3,409	**85**	1,663	2,371	3,189
90	1,987	2,632	3,402	**90**	1,662	2,368	3,183
95	1,985	2,629	3,396	**95**	1,661	2,366	3,178
100	1,984	2,626	3,390	**100**	1,660	2,364	3,174

Tableau A.3
Distribution des valeurs critiques de F ($\alpha = 0,05$)

v_2 (dl_{intra})	v_1 *Les degrés de liberté intergroupes*										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
1	161,446	199,499	215,707	224,583	230,160	233,988	236,767	238,884	240,543	241,882	243,905
2	18,513	19,000	19,164	19,247	19,296	19,329	19,353	19,371	19,385	19,396	19,412
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,785	8,745
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,912
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,678
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,000
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,575
8	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,388	3,347	3,284
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,073
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,913
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,788
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,687
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,475
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,278
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282	2,236	2,165
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,092

Tableau A.3
Distribution des valeurs critiques de F ($\alpha = 0,05$) (suite)

	1	2	3	4	5	6	7	8	9	10	12
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077	2,003
50	4,034	3,183	2,790	2,557	2,400	2,286	2,199	2,130	2,073	2,026	1,952
100	3,936	3,087	2,696	2,463	2,305	2,191	2,103	2,032	1,975	1,927	1,850
200	3,888	3,041	2,650	2,417	2,259	2,144	2,056	1,985	1,927	1,878	1,801
250	3,879	3,032	2,641	2,408	2,250	2,135	2,046	1,976	1,917	1,869	1,791
300	3,873	3,026	2,635	2,402	2,244	2,129	2,040	1,969	1,911	1,862	1,785
350	3,868	3,022	2,630	2,397	2,240	2,125	2,036	1,965	1,907	1,858	1,780
400	3,865	3,018	2,627	2,394	2,237	2,121	2,032	1,962	1,903	1,854	1,776
500	3,860	3,014	2,623	2,390	2,232	2,117	2,028	1,957	1,899	1,850	1,772
600	3,857	3,011	2,620	2,387	2,229	2,114	2,025	1,954	1,895	1,846	1,768
800	3,853	3,007	2,616	2,383	2,225	2,110	2,021	1,950	1,892	1,843	1,764
1000	3,851	3,005	2,614	2,381	2,223	2,108	2,019	1,948	1,889	1,840	1,762

Tableau A.3
Distribution des valeurs critiques de F ($\alpha = 0{,}01$)

v_2 (dl_{intra})	v_1 *Les degrés de liberté intergroupes*										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
1	4052,185	4999,340	5403,534	5624,257	5763,955	5858,950	5928,334	5980,954	6022,397	6055,925	6106,682
2	98,502	99,000	99,164	99,251	99,302	99,331	99,357	99,375	99,390	99,397	99,419
3	34,116	30,816	29,457	28,710	28,237	27,911	27,671	27,489	27,345	27,228	27,052
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659	14,546	14,374
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158	10,051	9,888
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,718
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,469
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,667
9	10,562	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257	5,111
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,706
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539	4,397
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296	4,155
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,666
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368	3,231
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,217	3,129	2,993
30	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	3,067	2,979	2,843

Tableau A.3
Distribution des valeurs critiques de F ($\alpha = 0{,}01$) (suite)

	1	2	3	4	5	6	7	8	9	10	12
40	7,314	5,178	4,313	3,828	3,514	3,291	3,124	2,993	2,888	2,801	2,665
50	7,171	5,057	4,199	3,720	3,408	3,186	3,020	2,890	2,785	2,698	2,563
100	6,895	4,824	3,984	3,513	3,206	2,988	2,823	2,694	2,590	2,503	2,368
200	6,763	4,713	3,881	3,414	3,110	2,893	2,730	2,601	2,497	2,411	2,275
250	6,737	4,691	3,861	3,395	3,091	2,875	2,711	2,583	2,479	2,392	2,256
300	6,720	4,677	3,848	3,382	3,079	2,862	2,699	2,571	2,467	2,380	2,244
350	6,708	4,666	3,838	3,373	3,070	2,854	2,691	2,562	2,458	2,372	2,236
400	6,699	4,659	3,831	3,366	3,063	2,847	2,684	2,556	2,452	2,365	2,229
500	6,686	4,648	3,821	3,357	3,054	2,838	2,675	2,547	2,443	2,356	2,220
600	6,677	4,641	3,814	3,351	3,048	2,832	2,669	2,541	2,437	2,351	2,214
800	6,667	4,632	3,806	3,343	3,040	2,825	2,662	2,533	2,429	2,343	2,207
1000	6,660	4,626	3,801	3,338	3,036	2,820	2,657	2,529	2,425	2,339	2,203

Tableau A.3
Distribution des valeurs critiques de F (α = 0,001)

v_2 (dl_{intra})	v_1 *Les degrés de liberté intergroupes*										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
1	405311,584	499725,342	540256,500	562667,847	576496,124	586032,867	593185,425	597953,796	602245,331	605583,191	610351,563
2	998,378	998,843	999,309	999,309	999,309	999,309	999,309	999,309	999,309	999,309	999,309
3	167,056	148,488	141,095	137,079	134,576	132,830	131,608	130,618	129,861	129,221	128,319
4	74,127	61,249	56,170	53,435	51,718	50,524	49,651	48,996	48,472	48,050	47,410
5	47,177	37,122	33,200	31,083	29,751	28,835	28,165	27,649	27,241	26,914	26,419
6	35,507	27,001	23,705	21,922	20,802	20,031	19,463	19,030	18,688	18,412	17,990
7	29,246	21,690	18,772	17,197	16,207	15,520	15,018	14,634	14,330	14,083	13,708
8	25,415	18,494	15,829	14,392	13,484	12,858	12,398	12,045	11,767	11,540	11,194
9	22,857	16,387	13,901	12,560	11,714	11,129	10,697	10,368	10,106	9,894	9,570
10	21,038	14,905	12,553	11,283	10,481	9,926	9,517	9,204	8,956	8,754	8,446
11	19,687	13,812	11,561	10,346	9,579	9,047	8,655	8,355	8,116	7,923	7,625
12	18,645	12,973	10,805	9,633	8,892	8,378	8,001	7,711	7,480	7,292	7,005
15	16,587	11,340	9,335	8,253	7,567	7,091	6,741	6,471	6,256	6,081	5,812
20	14,819	9,953	8,098	7,096	6,461	6,019	5,692	5,440	5,239	5,075	4,823
25	13,877	9,222	7,451	6,493	5,885	5,462	5,148	4,906	4,713	4,555	4,311
30	13,293	8,773	7,054	6,125	5,534	5,122	4,817	4,582	4,393	4,239	4,001
40	12,609	8,251	6,595	5,698	5,128	4,731	4,436	4,207	4,024	3,874	3,643

Tableau A.3
Distribution des valeurs critiques de F (α = 0,001) (suite)

	v_1 *Les degrés de liberté intergroupes*										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
50	12,222	7,956	6,336	5,459	4,901	4,512	4,222	3,998	3,819	3,671	3,443
100	11,496	7,408	5,857	5,017	4,482	4,107	3,829	3,612	3,439	3,296	3,074
200	11,154	7,152	5,634	4,812	4,287	3,920	3,647	3,434	3,263	3,123	2,904
250	11,089	7,102	5,591	4,772	4,250	3,884	3,612	3,400	3,230	3,089	2,871
300	11,044	7,070	5,562	4,746	4,225	3,860	3,588	3,377	3,207	3,067	2,849
350	11,012	7,046	5,542	4,727	4,207	3,843	3,572	3,361	3,191	3,051	2,834
400	10,989	7,029	5,527	4,713	4,194	3,830	3,560	3,349	3,179	3,040	2,822
500	10,957	7,004	5,506	4,693	4,175	3,813	3,542	3,332	3,163	3,023	2,806
600	10,935	6,988	5,492	4,681	4,163	3,801	3,531	3,321	3,152	3,012	2,795
800	10,908	6,968	5,474	4,665	4,148	3,786	3,517	3,307	3,138	2,999	2,782
1000	10,892	6,956	5,464	4,655	4,139	3,778	3,508	3,299	3,130	2,991	2,774

Tableau A.4
Distribution des valeurs critiques de χ^2

	seuil α		
v	< 10 %	< 5 %	< 1 %
1	2,706	5,991	15,086
2	4,605	9,488	21,666
3	6,251	12,592	26,217
4	7,779	14,067	29,141
5	9,236	16,919	32,000
6	10,645	18,307	34,805
7	12,017	21,026	38,932
8	13,362	22,362	40,289
9	14,684	23,685	41,638
10	15,987	24,996	42,980
11	17,275	27,587	46,963
12	18,549	28,869	48,278
15	22,307	33,924	54,776
20	28,412	41,337	64,950
25	34,382	48,602	73,683
30	40,256	55,758	82,292
40	51,805	68,669	98,028
50	63,167	82,529	114,695
100	118,498	144,354	186,393

Page laissée blanche

RÉPONSES AUX QUIZ RAPIDES

Chapitre 1

1.1 Variables : Nom, prénom, position, salaire ; constantes : équipe et ville.

1.2 Il faudra créer une nouvelle catégorie « 4 », telle que « Autre » ou « Yeux de couleurs différentes » ou même « Œil bleu/œil vert ».

1.3 Julie = 1, Marie = 2, Paul = 3.

1.4 Oui, parce qu'il s'agit d'une échelle de rapport avec un vrai point « zéro ».

1.5 a) *Échelle nominale :* Marie, Paul et Julie ont obtenu trois valeurs différentes.

b) *Échelle ordinale :* Marie est arrivée première, Paul deuxième et Julie troisième (ou dernière).

c) *Échelle à intervalles :* Marie a beaucoup mieux réussi que Paul, qui n'a que légèrement mieux réussi que Julie.

d) *Échelle de rapport :* la performance de Marie est de (90 – 71)/71 = 26,7 % meilleure que celle de Paul, et la performance de Paul est de (71 – 70)/70 = 1,4 % supérieure à celle de Julie.

Chapitre 2

2.1 13 joueurs gagnent 2 000 000 $.

2.2 L'ajout d'un autre joueur pourrait changer la taille de l'écart entre le salaire le plus bas et le salaire le plus élevé, ce qui aurait comme effet de changer la taille des catégories. La reconstruction du tableau est donc une procédure sage.

2.3 96,2 %, soit 25 joueurs sur 26.

2.4 Les distributions de salaires tendent, en général, à suivre ces formes de distribution (très asymétrique positive). Pour les joueurs de basket-ball, cela sera donc probablement vrai.
Quant au nombre de poissons, la distribution sera moins asymétrique.

2.5 Très difficiles.

2.6 Bimodale.

2.7 Probablement asymétrique positive.

Chapitre 3

3.1 Les modes sont : 49, 65, 71 et 76 ; ce qui fait que cette distribution est multimodale.

3.2 a) Les six dernières observations sont 76, 76, 77, 78, 87, 90. La médiane pour les six dernières est 77,5 et pour les cinq dernières, elle est 78.

b) Oui. Par exemple, si nous ajoutons le nombre 100 aux 6 dernières observations, la médiane devient 78.

3.3 L'erreur moyenne sera invariablement zéro. Il y aura invariablement une erreur plus grande que zéro, dans le mode et la médiane, à moins que la distribution ne soit parfaitement symétrique.

3.4 Asymétrique positive. Dans ces cas, la moyenne est toujours plus élevée que la médiane.

3.5 Montréal : 4 450 000 $; Atlanta : 2 875 000 $. Elle est de 17 dans le Tableau 3.2.

3.6 Lorsque toutes les valeurs de la distribution sont identiques.

3.7 La valeur 60 se trouvant exactement à la moyenne, elle ajoutera la quantité zéro au numérateur de la formule de la variance. La variance n'augmentera pas (elle se réduira un peu). Mais l'ajout de la note 20 accroîtra la quantité au numérateur ; par conséquent, la variance augmentera.

3.8 1 et 4 respectivement.

3.9 Positif.

3.10 L'asymétrie sera zéro, alors que le degré d'aplatissement aura un signe négatif.

3.11 Examen A : CV = 0,20. (*Nota bene :* l'énoncé propose une variance de 100. Il faut d'abord calculer son écart type $\sqrt{100} = 10$. Maintenant, nous pouvons calculer le coefficient de variabilité.) Examen B : CV = 0,25. L'examen B démontre plus de variance relativement à la moyenne.

Chapitre 4

4.1 Le rang de M. X est 679. Si au moins une personne d'affaires dans la distribution touche un salaire plus faible que 11 millions de dollars, le rang attribué à M. X sera plus élevé. Mais si les gens d'affaires dans la distribution touchent tous un salaire supérieur à 11 millions, le rang attribué à M. X restera le même (679).

4.2 a) (16,5-0,5)/32 = 0,50.
b) Le même.

4.3 Les notes de 52, 65, 74 b) quartile 1 = 52 ou moins ; quartile 2 = plus de 52 et au moins 65 ; quartile 3 = plus de 65 et au moins 74 ; quartile 4 = plus de 74.

4.4 La moyenne est égale à zéro dans les deux cas.

4.5 a) Les distributions X et Y ne démontrent pas le même niveau de variabilité lorsque nous analysons les données brutes.
b) Après la conversion en valeurs étalons Z, les moyennes et les variances sont toujours 0 et 1 respectivement. Ayant tous deux la même moyenne (0) et la même variance (1), X et Y produiront un coefficient de variabilité identique.

4.6 La température du mois d'avril est Z = –0,46, que la valeur initiale soit en Celsius ou en Fahrenheit. La valeur T correspondante sera alors T = (10 X –0,46) + 50 = 45,4.

4.7 Le salaire d'Ingrid serait de 45 000 euros.

Chapitre 5

5.1 La distribution en pointillé n'est pas symétrique.

5.2 L'aire totale sous cette courbe sera de 100 %, mais la densité sera plus élevée au centre. En conséquence, entre –1 et +1, il y aura plus que 68,26 % des observations.

5.3 Cette observation (140) se situe à +2 écarts types de la moyenne [Z = (140 – 100)/20 = +2]. La Figure 5.4 indique que 97,72 % des observations sont égales ou inférieures à cette observation (50 % + 34,13 % + 13,59 % = 97,72 %). Par conséquent, seulement 2,28 % des observations (100 % – 97,72 % = 2,28 %) seront plus loin de la moyenne. La probabilité sera alors p = 0,0228.

5.4 Cinquante observations sont supérieures ou égales à 10 (la moyenne). Entre 10 et 14, nous avons deux écarts types, ce qui équivaut à 34,13 % + 13,59 % = 47,72 %. Par conséquent, environ 48 observations se trouveront entre la moyenne (10) et 14. Enfin, si 34,13 % des observations se trouvent entre la moyenne et –1 écart type, il s'ensuit que seulement 15,87 % des observations auront une valeur moins grande que 8.

5.5 La proportion au-dessous de la moyenne est 0,50. La proportion entre la moyenne +1 écart type est 0,3413 et la proportion entre +1 et +2 écarts types est 0,1359. Enfin, la proportion se trouvant entre +2 et +3 écarts types est p = 0,0214. En additionnant ces proportions, on obtient 0,50 + 0,3413 + 0,1359 + 0,0214 = 99,86. Le rang percentile pour cette observation est 99,86 % et, puisque nous ne désirons pas conclure à un percentile de 100 (voir le chapitre 4), nous attribuons le rang percentile 99 à cette observation.

5.6 a) En se référant au Tableau 5.4, on sait que 0,013 % des observations seront plus petites que –3 écarts types. Le rang percentile sera alors plus petit que 1, mais, en arrondissant, nous lui attribuons le rang percentile de 1. En fait, plus de 99 % des observations seront supérieures à cette observation.

b) Nous savons que seulement 0,013 % des observations se situent à plus de +3 écarts types de la moyenne. Puisque la distribution normale est symétrique, seulement 0,013 % des observations se situent encore plus loin que –3 écarts types de la moyenne.

Au total, alors, 0,013 % + 0,013 % = 0,026 % des observations seront plus grandes et plus petites que ±3 écarts types et 99,974 % seront comprises entre ces limites.

Chapitre 6

6.1 {87 ; 17,4}.

6.2 En présumant que les boules sont faites avec le même matériau, une boule plus grande devrait contenir plus de ce matériau, impliquant que la relation entre la taille et le poids des boules sera positive. En présumant que nous avons un montant limité d'argent dans nos poches, plus nous dépensons pour le disque, moins il nous restera d'argent. La relation sera négative.

6.3 La corrélation sera zéro.

6.4 La relation entre le nombre d'heures de travail dans une journée et le nombre de minutes passées au travail sera nécessairement parfaite et positive. La relation entre le nombre de dossiers résolus et les heures de travail sera probablement positive, mais il est peu probable qu'elle soit parfaite. Ce n'est pas parce que nous passons plus de temps sur une tâche que nous sommes nécessairement plus productifs.

6.5 La variance de la variable « nombre de nez » est égale à zéro. La corrélation sera zéro parce que le nombre de nez ne permet pas de savoir si une personne fume peu ou beaucoup.

6.6 La corrélation sera probablement plus forte pour le groupe A que pour le groupe B. Cela est vrai, car les enfants âgés d'un an auront des tailles très différentes de ceux âgés de huit ans (groupe A). Dans le groupe B, les enfants de six ans auront des tailles plutôt semblables à celles des enfants de sept ans. Techniquement, la variance pour la variable « taille » sera plus forte pour le groupe A que pour le groupe B. De plus, pour le groupe A, la différence d'âge des enfants est plus grande (l'âge peut prendre des valeurs allant de 1 à 8). Pour le groupe B, il n'existe que deux valeurs possibles pour l'âge (6 ou 7). La variable « âge » aura plus de variance pour le groupe A que pour le groupe B.

6.7 La corrélation restera quasiment inchangée, que la corrélation initiale soit forte ou faible.

6.8 La corrélation entre la mesure d'aptitude (X) et la performance au travail (Y) sera de zéro : la performance au test ne réduit pas l'incertitude au sujet de la performance au travail.

Chapitre 7

7.1 Variable indépendante (X) = nombre de voitures et variable dépendante (Y) = pollution ; variable indépendante (X) = pollution et variable dépendante (Y) = nombre de voitures.

7.2 Voir le Tableau 7.1.

7.3 Le graphique est imprécis parce qu'un graphique n'est qu'une représentation visuelle des deux statistiques, l'ordonnée à l'origine et le coefficient de régression.

7.4 $Y = 32 + 1{,}8 \times 22 = 71{,}6$.

7.5 La valeur prédite est surestimée.

7.6 Étudiant A : forte recommandation d'abandonner le cours ; étudiant B : recommandation d'abandon ; étudiant C : recommandation de poursuivre le cours.

Chapitre 8

8.1 Une population. Un recensement inclut, par définition, tous les membres d'une population.

8.2 La procédure de sélection des échantillons viole les notions d'égalité des chances et d'indépendance. Égalité des chances : seuls les échantillons trafiqués ont été choisis. Les autres échantillons potentiels (non trafiqués) n'avaient aucune chance d'être choisis. Indépendance : la quantité d'or contenue dans les échantillons était prédéterminée, puisqu'elle ne pouvait pas varier librement.

8.3 Le sujet d'étude est « les familles avec des enfants ». La procédure de sélection des échantillons est appropriée, car chaque famille avec enfants a une chance égale et indépendante d'être choisie.

8.4 La mesure de Weschler a été administrée à plusieurs millions de personnes. Ces échantillons produisent une moyenne de 100 et un écart type de 16. Les informations descriptives sont donc des statistiques et non pas des paramètres. Par conséquent, nous écrivons $M_{QI} = 100$. Mais puisque nous avons des millions d'observations, nous pouvons compter sur le fait que les statistiques sont très proches des paramètres.

8.5 En utilisant C = 4, la moyenne devient (1 + 2 + 4)/3 = 2,33, ce qui prouve que l'observation C ne peut pas être 4, puisque la moyenne est de 2. Si C = 3, alors M = 2.

8.6 Cas 1 : $\sum(X - \mu)^2/N$, car nous nous intéressons exclusivement aux notes de la classe, ce qui fait que, dans ce cas, notre « échantillon » est la population. Cas 2 : $\sum(X - M)^2/(N - 1)$. Nous voulons estimer la variance de la population (tous les étudiants) à partir des notes de l'échantillon.

8.7 L'état de la chaussée leur causerait une inquiétude, ce qui les amènerait à ralentir. H : la vitesse des véhicules, après que les chauffeurs ont remarqué l'état de la route, sera plus faible qu'avant. H_0 : la vitesse des véhicules, avant-après, sera la même.
$H : \mu_{avant} \neq \mu_{après}$; $H_0 : \mu_{avant} = \mu_{après}$.

8.8 La théorie indique que les nordiques sont moins « émotifs » que les sudistes. Une hypothèse vérifiable pourrait être celle-ci : H : le nombre de crimes passionnels par 100 000 habitants enregistrés l'année dernière dans les pays scandinaves est plus petit que celui enregistré dans les pays qui bordent la Méditerranée ; et l'hypothèse nulle : H_0 : le nombre de crimes passionnels n'est pas différent.

8.9 Les hypothèses se posent de la manière suivante : H : le tabac cause le cancer ; H_0 : le tabac ne cause pas le cancer. La variable dépendante est le niveau de cancer moyen dans les deux groupes de rats. $H : \mu_{avec\ tabac} \neq \mu_{sans\ tabac}$ et l'hypothèse nulle H_0 devient : $H_0 : \mu_{avec\ tabac} = \mu_{sans\ tabac}$. La conclusion du chercheur n'est pas appropriée. Il doit se limiter à dire qu'il n'existe pas de preuve voulant que le tabac cause le cancer (il ne prouve pas que le tabac ne cause pas le cancer). S'il avait prolongé l'exposition au tabac pendant plus de trois semaines, ou augmenté le niveau de fumée dans les cages, les résultats obtenus auraient pu être différents.

Chapitre 9

9.1 C'est possible que les Ph.D. proviennent d'une population de salaire différente de celle des Canadiens en général. Mais puisque nous ne connaissons pas l'erreur type de la moyenne, cette conclusion est prématurée.

9.2 Les citoyens seront probablement la population ayant une plus grande variance dans les attitudes. En effet, dans la catégorie des citoyens se trouvent des médecins, mais aussi des gens qui n'ont jamais mis les pieds dans un hôpital. Il est impossible que la variance des citoyens soit inférieure à la variance des médecins, car les médecins étant des citoyens, l'attitude des citoyens variera au moins autant que celle des médecins.

9.3 Puisque nous avons postulé que nous pouvions avoir TOUS les sous-ensembles possibles de taille K, sans exception, il ne manque aucun échantillon. Nous avons la population des échantillons à notre disposition. En particulier, il ne manque pas les échantillons extrêmes, même s'ils sont rares, et donc, il n'est pas nécessaire de diviser par K – 1.

9.4 La valeur Z est de +4. La probabilité d'être au-delà de +4 est infime : 0,00003 !

9.5 Il nous faut au préalable connaître l'erreur type de la moyenne. L'écart type de l'échantillon est $\sqrt{100} = 10$. Nous appliquons la Formule 9.3 et trouvons que l'erreur type de la moyenne est $10/\sqrt{100} = 10/10 = 1$. La moyenne de la population se trouvera probablement entre 109 et 111. Lorsque nous réduisons l'échantillon à 25 personnes, l'erreur type de la moyenne est $10/\sqrt{25} = 10/5 = 2$. La moyenne de la population se trouve entre 108 et 112. Les deux estimations ne sont pas pareilles. Lorsque nous avons moins d'information au sujet d'une population, parce que notre échantillon est plus petit, l'estimation que nous faisons de la population est moins précise.

9.6 L'écart entre ce qui est obtenu et ce que l'on affirme étant de 2,58 (6,58 – 4) et l'erreur type de la moyenne étant 1, l'écart type standardisé est aussi de 2,58 (2,58/1). La probabilité d'obtenir une telle différence (voir la table de la distribution normale standardisée) est de 1 %.

9.7 Oui, vous le pouvez. Il s'agit d'une possible hypothèse nulle car elle représente une situation bien précise ($H_0 : \mu = 4$). L'hypothèse est que le quidam a tort ($H : \mu \neq 4$).

9.8 Avec un score de 140, on obtient un score Z de (140 – 100)/10, soit +4. La probabilité d'obtenir +4 est bien plus faible que le seuil de signification et donc cet échantillon ne provient probablement pas de la population. Le score critique en score Z est +2,58 (la probabilité de dépasser +2,58 est de 1 %). En dénormalisant, on obtient (2,58 × 10) + 100, soit 125,8. Dès que la moyenne d'un échantillon dépasse 125,8, on juge que celle-ci est est trop élevée pour conclure que l'échantillon provient de la population indiquée dans l'énoncé.

9.9 Si nous avons la population en main, l'erreur que nous commettons en calculant la moyenne ne peut être que nulle. L'erreur type de la moyenne d'une population calculée à partir de la population entière est zéro.

9.10 En tant que patron, vous ne voulez surtout pas une erreur de type II (conclure qu'il n'y a pas d'effet alors qu'il y en a réellement un). Une façon de réduire le risque d'erreur de type II est d'avoir un plus grand échantillon. Une autre façon est d'accroître le seuil de décision ($\alpha = 0{,}10$ au lieu de $\alpha = 0{,}05$). Cette dernière solution n'est pas souhaitable du point de vue du gouvernement car le risque d'une erreur de type I double (conclure que le médicament fonctionne alors qu'en réalité, il ne fait rien).

Chapitre 10

10.1 Les valeurs critiques sont 2,201 et 3,106 respectivement. Pour N=12, dl=11.

10.2 Dans le premier cas, il s'agit de deux échantillons jumelés. Dans le second cas, il s'agit d'échantillons indépendants, car ce ne sont pas les mêmes élèves. Par contre, s'il existe des élèves qui ont redoublé, il faut les retirer de l'échantillon car, pour eux, ce sont des données pairées !

10.3 Non, nous obtenons $\frac{49 \times 12 + 499 \times 20}{49 + 499} = 19{,}3$. Plus un groupe est grand par rapport à l'autre, plus la variance combinée tend vers la

variance de ce groupe. Contenant plus d'observations, nous avons plus confiance dans les statistiques produites par ce grand groupe.

10.4 Oui, dans tous les cas, (0,05,0,01 et 0,001) le t est significatif.

10.5 Oui.

10.6 Non, la conclusion ne change pas. La valeur du $t_{critique}$ est unicaudale et est de 2,42 au lieu de 2,70 (bicaudale). Le $t_{observé}$ étant supérieur au $t_{critique}$, nous maintenons notre conclusion que la différence entre les groupes est statistiquement significative.

10.7 Le chercheur a fait un test t. Puisqu'il y a 134 degrés de liberté, cela implique qu'il a mesuré 135 personnes. Le risque d'erreur dans sa conclusion, qu'il jugeait acceptable, est de 5 % (le seuil de signification). Même avec un seuil de signification de 1 %, la conclusion serait restée la même, car la valeur critique est dans ce cas de 3.36, et il a obtenu un t observé de 6,4.

Chapitre 11

11.1 Il y a (25 x 24)/2, soit 300 paires de tests possibles à vérifier. Concernant le toast, c'est la même formule : il y a (10 x 9)/2 = 45 paires de convives, et donc 45 tintements.

11.2 A. Non, il n'y a pas de cumul des erreurs. Avec la Formule 11.1, on trouve que c = 1 (c'est bien le cas, on ne fait qu'une seule comparaison, un seul test). Avec la Formule 11.2, on trouve que $p = \alpha$.
B. Pas plus de trois comparaisons. À quatre comparaisons, le cumul devient 0,06, ce qui excède le risque choisi (0,05).

11.3 La variable indépendante est le traitement administré. Elle a quatre niveaux : thérapie comportementale, psychanalyse, thérapie cognitive, thérapie chimique.

11.4 Il ne faut pas confondre la variable indépendante et la variable dépendante (la mesure). Ici, il s'agit du niveau de dépression après le traitement.

11.5 Dans les deux cas, le numérateur est une mesure de la différence entre les groupes (intergroupe). S'il n'y a que deux groupes, on soustrait les moyennes ; s'il y a plus de deux groupes, on calcule le carré moyen (comme on le verra plus loin). Quant au dénomina-

teur, il s'agit dans les deux cas d'une mesure de l'erreur probable. Pour le test t, on l'obtient avec l'erreur type de la moyenne ; pour le test F, avec le carré moyen intragroupe (voir plus loin). Un grand nombre de tests sont de cette forme : une mesure de l'effet pondérée (divisée par) une mesure de l'erreur probable.

11.6 Le résultat sera zéro : il n'y a aucune différence intergroupe. Dans ce cas, inutile de faire un test statistique : tous les groupes sont très probablement de la même population.

11.7 Les degrés de liberté intra- et intergroupe sont respectivement 90 (nombre total d'observations moins nombre de groupes) et 9 (nombre de groupes –1).

11.8 Il s'agit d'une situation où il n'y a aucune variabilité intragroupe. Le CM étant une mesure de variabilité, il sera de zéro. Quant au CM_{inter}, nous n'avons pas assez d'information pour le quantifier.

11.9 La valeur F pour 9 et 90 degrés de liberté [noté en court F(9, 90)] avec un seuil de 5 % vaut environ 1,975 (quand un tableau de valeurs critiques ne contient pas l'entrée exacte, prendre la ligne qui suit ou utiliser un logiciel) ; pour un seuil de 1 %, on trouve que F(9, 90) vaut 2,590.

11.10 Dans le cas où nous acceptons un risque de 5 %, nous disons qu'il existe une différence, car le $F_{critique}$ F(5, 94) vaut 2,31 et le $F_{observé}$ est plus grand. Nous écririons : « Il existe au moins une différence significative entre les groupes [$F(5, 94) = 4{,}3$, $p < 0{,}05$]. » Dans le cas où nous ne voulons courir presque aucun risque (1 sur 1 000), la valeur critique recule à 4,48. Dans ce cas, nous restons avec l'hypothèse nulle : « Il n'existe aucune différence significative entre les groupes [$F(5, 94) = 4{,}3$, $p > 0{,}001$]. »

11.11 Vous pouvez sans problème augmenter la taille des échantillons. Vous pouvez aussi augmenter le seuil α, mais ça accroît vos risques d'erreur de type I. Finalement, s'il est possible d'avoir moins de variabilité dans les groupes (par exemple en mesurant un seul type de quartier très précis), cela peut réduire le CM_{intra} et donc augmenter le $F_{observé}$.

11.12 Zéro, car la SC_{intra} vaut zéro, et zéro divisé par une SC_{total} quelconque donne toujours zéro.

Chapitre 12

12.1 Il n'y a que deux niveaux. Soit ils diffèrent (et l'ANOVA sur le facteur « type de maladie » est significatif), soit ils ne diffèrent pas, auquel cas il n'y a pas de raison de faire un test de comparaison multiple.

12.2 Disons que 100 est le niveau habituel de dépression (mesuré par un questionnaire) et qu'un score supérieur à 100 représente la direction de la guérison. Si nos deux groupes sans médicament ont 100, disons que le groupe avec médicament ET avec thérapie a 120 (peu importe le chiffre exact). Disons aussi que le groupe avec médicament mais sans thérapie a obtenu une moyenne de 90 (les dépressifs le sont plus qu'avant !). Si on calcule les moyennes marginales, on obtient : avec médicament : 105 ; sans médicament : 100. C'est l'information qu'a utilisée la compagnie pharmaceutique dans sa proclamation. Par contre, regardons aussi les autres moyennes marginales : avec thérapie : 110 ; sans thérapie : 95. La compagnie ne nous avait pas fait part de ceci : la thérapie a beaucoup d'effets positifs (en fait, plus que le médicament). Finalement, si vous faites un graphique de l'interaction, vous verrez que les droites ne sont pas parallèles. L'interaction est probablement significative dans ce cas-ci (il faudrait vérifier). Autrement dit, dans le cas de ces données fictives, le médicament peut être valable si et seulement si la personne accepte de suivre une thérapie en même temps.

Chapitre 13

13.1 dl=9.

13.2 dl= R=3 (années d'étude) ; C=3 (ce qu'ils feront pour le lunch) ; dl= (R-1) × (C-1) = (3-1) × (3-1)=4. La valeur critique pour 4 degrés de libertés au seuil alpha = 5 % est 14,067. Nous avons obtenu un chi deux de 13, une valeur inférieure à la valeur critique. La différence n'est pas statistiquement significative.

13.3 La corrélation est $\rho = -0{,}89$, une relation négative. Nous concluons que ceux qui avaient des revenus plus élevés pendant leur carrière ont tendance à avoir des revenus plus faibles lors de leur retraite.

N'auraient-ils pas, comme la cigale et la fourmi, économisé insuffisamment pendant leur vie active ?

13.4 a) Il faut faire une mise en rang (1 à 15). On peut anticiper que la bouteille d'oxygène occupera le rang 1 (plus important) et que les allumettes seront les objets les moins important (rang 15). Les autres objets occuperont les rangs entre ces deux extrêmes.

b) et c). Établissez une corrélation par rang de Spearman entre votre réponse et celle de la NASA (qui se trouve ci-dessous) et répétez cela pour la réponse de votre collègue.

d) Si le ρ entre votre réponse et celle de la NASA est plus élevé que celui obtenu par votre collègue, votre réponse est supérieure car votre mise en rang est plus proche de la mise en rang « idéale » produite par la NASA. Si votre réponse (ou celle de l'autre) correspond à une corrélation proche de zéro avec la réponse NASA ou, pire encore, si elle le coefficient ρ est négatif, il serait prudent pour vous (ou votre collègue) soit de poursuivre une carrière autre que celle d'astronaute soit de ne jamais avoir d'accident !

Objet	***(NASA)***	***Explication***
Allumettes	15	Inutiles (pas d'oxygène sur la lune).
Nourriture sèche	4	Produit d'alimentation efficace.
Corde	6	Utile pour grimper.
Toile de parachute	8	Protection (rayons du soleil).
Chaufferette solaire	13	Inutile : si l'atterrissage se fait sur la face cachée, la pile ne fonctionnera pas. Sinon elle ne sera pas requise.
Revolvers	11	Utiles pour l'autopropulsion.
Lait en poudre	12	Redondant avec la nourriture sèche.
Oxygène	1	Essentiel à la survie.
Carte lunaire	3	Instrument de navigation essentiel.
Radeau avec CO_2	9	Le CO_2 est utilisable pour l'autopropulsion. Le radeau peut servir à transporter les objets.

Compas	14	Le compas ne fonctionne pas sur la lune.
Eau	2	Essentielle à la survie.
Fusées de signalisations	10	Peuvent être utilisées pour signaler sa position.
Trousse de 1ers soins	7	Importante en cas d'accident ou de malaise.
Walkie-talkie	5	Pour la communication (courte portée) avec les sauveteurs.

13.5 Le risque de commettre une erreur de type I plus petit que 1 % implique, d'après le tableau de la densité de la courbe normale, que le Z observé doit être égal ou supérieur à une valeur Z = 2,58. Puisque nous avons obtenu un Z = 2,96, il serait juste de conclure que le salaire des hommes et des femmes est différent et ce faisant, le risque de commettre une erreur d'inférence de type I est inférieur à 1 %. Non, l'auteur n'a pas raison.

13.6 Oui, nous pouvons travailler avec le groupe des femmes au lieu du groupe des hommes. L'inférence finale sera identique car la taille de la différence entre les hommes et les femmes est exactement la même que celle entre les femmes et les hommes !

BIBLIOGRAPHIE

Bennet, J. O., W. L. Briggs et M. F. Triola (2003), *Statistical Reasoning for Everyday Life*, Boston, Addison Wexley (2e édition).

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, New York, Academic Press.

Darlington, R. B. (1990), *Regression and Linear Models*, New York, McGraw-Hill.

Ferguson, G. A. (1976), *Statistical Analysis in Psychology and Education*, New York, McGraw-Hill, McGraw-Hill Series in Psychology.

Gauvrit, N. (2005), *Stats pour psycho*, Bruxelles, De Boeck/Larcier.

Gravetter, F. J. et L. B. Wallnau (2005), *Essentials of Statistics for the Behavioural Sciences*, Belmont (CA), Wadsworth/Thompson (5e édition).

Howell, D. C. (2008), *Méthodes statistiques en sciences humaines* (traduction de la 6e édition par Rogier, M., V. Yzerbyt et Y. Bestgen), Bruxelles, De Boeck.

Kranzler, J. H. (2003), *Statistics for the Terrified*, Upper Saddle River (NJ), Prentice-Hall.

Pedhazur, E. J. et L. P. Schmelkin (1991), *Measurement, Design and Analysis: An Integrated Approach*, Hillsdale (NJ), Lawrence Erlbaum Associates.

Tabachnick, B. G. et L. S. Fidell (2007), *Using Multivariate Statistics*, Boston, Pearson (5e édition).

Witte, R. S. et J. S. Witte (2010), *Statistics*, Hoboken (NJ), John Wiley & Sons (9e édition).

Page laissée blanche

Autres livres publiés par Robert R. Haccoun

Saks, A. et R. R. Haccoun (2010), *Managing Performance through Training and Development* (5e éd.), Scarborough, Ontario, Nelson Publishing.

Haccoun, R. R. et D. Cousineau (2007), *Statistiques. Concepts et applications* (1re éd.), Presses de l'Université de Montréal.

Saks, A. et R. R. Haccoun (2007), *Managing Performance through Training and Development* (4e éd.), Scarborough, Ontario, Nelson Publishing.

Saks, A. et R. R. Haccoun (2004), *Managing Performance through Training and Development* (3e éd.), Scarborough, Ontario, Nelson Publishing.

Bordeleau, Y., L. Brunet, R. R. Haccoun, A.-J. Rigny et A. Savoie (1987), *Modelos de investigacion para el desarrollo de recursos humanos* (traduction en espagnol de Bordeleau et coll., 1982).

Bordeleau, Y., L. Brunet, R. R. Haccoun, A.-J. Rigny et A. Savoie (1982), *Comprendre l'organisation : Approches de recherches*, Montréal, Agence d'Arc.

Page laissée blanche

Ce livre a été imprimé au Québec en janvier 2010 sur du papier
entièrement recyclé sur les presses de Marquis imprimeur.