

Méthodes statistiques en sciences humaines

David C. **HOWELL**

Traduction et révision scientifique de Vincent Yzerbyt,
Yves Bestgen, Nathalie Lefèvre et Aurélie Bertrand

**3^e édition
française**

MANUEL DE RÉFÉRENCE

- Tout le cours et résumés
- Nombreux exemples
- + de 270 graphiques, schémas et tableaux
- + de 80 exercices



Toutes les synthèses
en audio

+ EN LIGNE

— Pour les étudiants
+ de 450 exercices, + de 50 défis,
180 flash-cards

— Pour les professeurs
Supports de cours



Méthodes statistiques en sciences humaines

COLLECTION OUVERTURES PSYCHOLOGIQUES

Des manuels de qualité (originaux en langue française et traductions des plus grands ouvrages anglo-saxons), régulièrement mis à jour avec les données le plus récentes, qui privilégient une organisation pédagogique progressive et offrent à l'étudiant de nombreux outils d'apprentissage.

DAVID C. HOWELL

Méthodes statistiques en sciences humaines

TRADUCTION ET RÉVISION SCIENTIFIQUE
DE VINCENT YZERBYT, YVES BESTGEN, NATHALIE LEFÈVRE
ET AURÉLIE BERTRAND

3^e édition française

Ouvrage original :
Statistical Methods for Psychology by David C. Howell (8th edition)
© 2013 Wadsworth, Cengage Learning

Pour toute information sur notre fonds et les nouveautés dans votre domaine de spécialisation, consultez notre site web : www.deboecksuperieur.com

© De Boeck Supérieur s.a., 2024
Rue du Bosquet, 7 – B-1348 Louvain-la-Neuve

Tous droits réservés pour tous pays pour la traduction et l'adaptation en langue française.

Il est interdit, sauf accord préalable et écrit de l'éditeur, de reproduire (notamment par photocopie) partiellement ou totalement le présent ouvrage, de le stocker dans une banque de données ou de le communiquer au public, sous quelque forme et de quelque manière que ce soit.

Dépôt légal: 3^e édition

Bibliothèque nationale, Paris: juillet 2024

Bibliothèque royale de Belgique, Bruxelles: 2024/13647/034

ISSN 2030-420X

ISBN 978-2-8073-0552-6

Sommaire

Remerciements	VII
Préface de l'auteur	IX
À propos de l'auteur	XIV
Compléments numériques	XV
 CHAPITRE 1 Concepts de base	 1
CHAPITRE 2 Description et exploration des données	17
CHAPITRE 3 La distribution normale	67
CHAPITRE 4 Distributions d'échantillonnage et tests d'hypothèse	87
CHAPITRE 5 Concepts de base des probabilités	117
CHAPITRE 6 Données catégorielles et khi-carré	149
CHAPITRE 7 Le test d'hypothèse appliqué aux moyennes	189
CHAPITRE 8 La puissance	241
CHAPITRE 9 Corrélation et régression	267
CHAPITRE 10 Techniques alternatives de corrélation	321
CHAPITRE 11 L'analyse de variance simple	347
CHAPITRE 12 Comparaisons multiples entre des moyennes de traitement . .	395
CHAPITRE 13 L'analyse factorielle de la variance	443
CHAPITRE 14 Les plans à mesures répétées	495
CHAPITRE 15 La régression multiple	545
CHAPITRE 16 Les analyses de variance et de covariance comme modèles linéaires généraux	617
CHAPITRE 17 Méta-analyse et études de cas uniques	667
CHAPITRE 18 Approches de rééchantillonnage et non paramétriques des données	701
 Annexes	 731
Bibliographie	765
Glossaire des symboles	773

Remerciements

Je voudrais remercier les personnes suivantes qui ont lu le manuscrit et m'ont aidé à améliorer ce livre: Angus MacDonald (University of Minnesota), William Smith (California State University-Fulleton), Carl Scott (University of St. Thomas-Houston), Jamison Fargo (Utah State University), Susan Cashin (University of Wisconsin-Milwaukee) et Karl Wuensch (East Carolina University), qui a fourni des remarques très utiles à propos de cette édition, mais aussi des précédentes. Pour les éditions précédentes, j'ai bénéficié des commentaires et des suggestions de Kenneth J. Berry (Colorado State University), Tim Bockes (Nazareth College), Richard Lehman (Franklin and Marshall College), Tim Robinson (Virginia Tech), Paul R. Shirley (University of California-Irvine), Mathew Spackman (Brigham Young University), Mary Uley (Lindenwood University) et Christy Witt (Louisiana State University). Leur influence est encore évidente dans la présente édition.

Pour cette édition, j'aimerais remercier Deborah A. Carroll (Southern Connecticut University), Paul Chang (Edith Cowen University), Ann Huffman (Northern Arizona University), Samuel Moulton (Harvard University), Therese Pigott (Loyola University Chicago), Lucy Troup (Colorado State University) et Meng-Jia Wu (Loyola University Chicago). Les collaborateurs de la maison d'édition ont été d'une aide exceptionnelle tout au long du processus et je tiens à remercier Jessica Egbert (manager marketing), Vernon Boes et Pamela Galbreath (directeurs de la création), Mary Noel (éditeur des médias), Kristin Ruscetta (manager de projet de production), Lauren Moody (assistante éditoriale) et Timothy Matray (éditeur acquisition).

David C. HOWELL
Université du Vermont

Préface de l'auteur

Tout comme les éditions précédentes, la huitième édition des *Méthodes statistiques en sciences humaines* passe en revue les techniques statistiques communément utilisées dans les sciences sociales et comportementales, en particulier, en psychologie et dans les sciences de l'éducation. Bien que ce livre soit destiné aux étudiants de premier cycle avancé et aux étudiants ayant obtenu ce diplôme, il ne nécessite pas d'avoir suivi préalablement un autre cours de statistique ou un cours de mathématique qui dépasse le niveau d'algèbre du lycée. Le début du livre fournira une révision utile aux étudiants qui ont suivi un cours d'introduction. Ce manuel convient pour un cours s'étalant soit sur une année, soit sur un semestre. Comme je me suis rendu compte que les étudiants et les professeurs se référaient souvent à leur tout premier manuel lorsqu'ils étaient confrontés à un problème statistique, j'ai inclus certaines matières spécifiques dans ce livre afin qu'il constitue une référence utile pour toute consultation ultérieure. Les professeurs qui souhaitent omettre ces matières n'auront aucune difficulté à le faire. Toutefois, j'ai essayé de n'inclure que ce qui me semblait susceptible d'être le plus utile afin d'éviter d'intimider inutilement le lecteur.

D'une certaine façon, cette édition représente une rupture par rapport aux éditions précédentes. Au fil des ans, chaque édition a intégré une quantité raisonnable de nouveaux éléments et la discussion a évolué pour refléter les changements dans la théorie et la pratique. Cette édition ne fait pas exception, sauf que certains changements sont plus importants que d'habitude. J'estime que les étudiants doivent comprendre certains sujets supplémentaires, qui ne peuvent être traités sans modifier une partie de l'ancien matériel.

La première édition de ce livre a été rédigée en 1973-1974, alors que j'étais en congé sabbatique à l'Université de Durham, en Angleterre. À cette époque, la plupart des cours de statistiques en sciences humaines étaient destinés aux psychologues expérimentaux et un grand pourcentage de ce travail concernait l'analyse de la variance. L'ouvrage de B.J. Winner, *Statistical Principles in Experimental Design* (1962, 1971) a dominé une large partie du domaine statistique et a influencé la manière dont les expériences étaient conçues, menées et analysées. Mon livre a suivi une grande partie de cette tendance, et ce à juste titre. Depuis lors, l'analyse statistique des données a connu des changements considérables. Si l'analyse de la variance reste une technique dominante, suivie de près par la régression multiple, de nombreux autres sujets ont fait leur apparition sur la liste. Par exemple, les mesures de la taille de l'effet sont beaucoup plus importantes qu'elles ne l'étaient à l'époque, la méta-analyse est passée de pratiquement rien à un moyen important de résumer un domaine, et le traitement des données manquantes a changé de manière significative grâce à l'introduction des modèles mixtes linéaires et des méthodes d'imputation des données. Nous ne pouvons plus former des étudiants préparés à mener des recherches de qualité sans aborder ces questions.

Mais pour inclure ces sujets et d'autres sujets similaires, il fallait supprimer quelque chose si l'on voulait que le texte reste d'une taille raisonnable. Dans le passé, les procédures de comparaison multiple, par exemple, ont fait l'objet d'une attention considérable avec de nombreuses approches concurrentes. Mais si vous regardez ce que les gens font réellement dans la littérature, la plupart des approches concurrentes intéressent davantage les théoriciens que les praticiens. Je pense qu'il est beaucoup plus utile d'écrire sur le traitement des données manquantes que d'énumérer toutes les alternatives pour effectuer des comparaisons entre les méthodes.

L'objectif que j'ai poursuivi en rédigeant cet ouvrage consistait à expliquer la matière à un niveau intuitif. Cela ne signifie pas que cette matière soit *édulcorée*, mais uniquement que l'accent est mis sur une *compréhension conceptuelle*. Ainsi, il se peut qu'un étudiant capable de dériver la distribution d'échantillonnage de t ne comprenne pas comment cette distribution doit être utilisée. Dans ce cas-ci, je me suis efforcé d'axer le raisonnement sur la signification d'une distribution d'échantillonnage et de montrer le rôle joué par cette distribution dans la théorie générale du test d'hypothèse. À mon avis, cette approche conviendra mieux qu'une approche plus technique pour faire comprendre aux étudiants le fonctionnement d'un test spécifique et les relations entre les tests.

Contrairement à l'opinion populaire, les méthodes statistiques sont constamment en évolution. Cela s'explique en partie par le fait que les sciences humaines s'étendent à de nombreux nouveaux domaines et, en partie, par le fait que nous trouvons de meilleures façons de poser des questions à propos de nos données. Aucun livre ne peut couvrir l'ensemble des sujets à traiter, mais il est essentiel de préparer les étudiants et les professionnels à aborder ces sujets en cas de besoin. Par exemple, les modèles multiniveaux/hiérarchiques sont de plus en plus courants dans la littérature de recherche. La compréhension de ces modèles nécessite des textes spécialisés, mais pour commencer à trier cette littérature, il est essentiel de comprendre les notions de facteurs fixes par rapport aux facteurs aléatoires et les plans multiniveaux. Ce livre ne peut pas se charger du premier aspect, en dérivant les modèles nécessaires, mais il peut aborder, et le fait, le second aspect, en établissant une base pour les plans fixes et aléatoires et les plans imbriqués. J'ai essayé d'établir des bases similaires pour d'autres sujets; par exemple, des dispositifs graphiques plus modernes et des statistiques de rééchantillonnage, là où je pouvais le faire sans entraîner le lecteur dans des sujets trop complexes. D'une certaine façon, ma responsabilité est d'essayer d'anticiper où nous allons et de donner au lecteur une base pour aller dans cette direction.

Modifications principales dans cette huitième édition

Cette 8^e édition (3^e édition française) introduit ou développe plusieurs thématiques qui visent à renforcer l'attrait de l'ouvrage pour l'étudiant et à mieux adapter son contenu aux tendances actuelles en méthodologie et en analyse des données.

- J'ai continué à attacher beaucoup d'importance aux suggestions du comité de l'Association américaine de Psychologie, comité qui a longuement examiné la question des tests d'hypothèse, et j'ai inclus davantage d'informations à propos de la taille d'effet et de l'importance d'un effet. Ces questions sont abordées d'une manière plus approfondie dans la présente édition par rapport aux précédentes afin de donner au lecteur une présentation très détaillée de cette problématique.
- Dans la septième édition, j'avais inclus du nouveau matériel sur des représentations graphiques, y compris des graphiques de probabilité, des graphiques de densité estimée par noyau et des graphiques résiduels. Chacun de ces éléments nous aide à mieux comprendre nos données et à évaluer le caractère raisonnable de nos hypothèses. Plus généralement, j'ai étendu l'utilisation des représentations graphiques dans cette édition.
- J'ai développé davantage le concept de rééchantillonnage pour illustrer ce que les approches plus traditionnelles tentent de faire sur la base d'hypothèses sous-jacentes. Il y a eu un mouvement accéléré dans la direction du rééchantillonnage, et de nombreuses solutions informatiques offrent aux utilisateurs des options telles que « *bootstrap* », « *simuler* » et « *rééchantillonnage* », qui reposent toutes sur cette approche.
- La couverture de l'analyse de Cochran-Mantel-Haenszel des tables de contingence est liée à l'exemple classique du paradoxe de Simpson appliqué aux données d'admission des diplômés de Berkeley. Ceci est en lien avec la motivation sous-jacente d'amener les étudiants à réfléchir profondément à la signification de leurs données.
- J'ai en outre modifié le chapitre 12 sur les techniques de comparaison multiple pour réduire le large éventail de tests discutés dans les précédentes éditions et pour couvrir le taux de fausse découverte de Benjamini et Hochberg. Au fur et à mesure que nous nous éloignons des taux d'erreur connus pour passer au taux de fausses découvertes, nous augmentons la puissance de nos analyses à un coût relativement faible en termes d'erreurs de type I.
- Une section du chapitre sur l'analyse de la variance pour les mesures répétées étend la discussion des modèles mixtes. Cette approche permet un bien meilleur traitement des données manquantes et relâche les hypothèses déraisonnables sur la symétrie composée. Cela sert d'introduction aux modèles mixtes sans essayer d'aborder tout un nouveau domaine en une fois.
- Plusieurs éditions précédentes contenaient un chapitre sur les modèles log-linéaires. J'ai remplacé ce chapitre par du matériel sur la méta-analyse et les plans à sujet unique. Les méta-analyses ont gagné en importance dans les sciences humaines et sous-tendent l'accent mis sur la « médecine basée

sur les preuves». Étant donné qu'un texte ne peut couvrir qu'une quantité limitée de matériel, je pense qu'une introduction à ces deux sujets est plus utile qu'un chapitre sur les modèles log-linéaires.

- Depuis des années, un environnement informatique gratuit nommé *R*, avec son package commercial *S-PLUS*, voit son influence augmenter en analyse statistique. Je n'ai pas l'intention de fournir des instructions en *R* en plus de tout le reste, mais je présente un grand nombre de programmes ou de sélections de code *R* disponibles sur le site Web du livre. Cela fournit un point de départ pour ceux qui sont intéressés par *R* et permet des démonstrations intéressantes des points soulevés dans le texte.
- Depuis plus de vingt ans, je rassemble sur Internet différentes matières destinées aux étudiants et aux professeurs. Ceux-ci y trouveront notamment des exemples supplémentaires (parfois complexes), des sujets non abordés dans ce livre, des données supplémentaires, d'autres sources Internet, différentes démonstrations qui conviendraient pour un cours ou une séance de travaux pratiques, etc. De nombreux endroits du livre renvoient spécifiquement à ce matériel si l'étudiant souhaite approfondir un sujet. Ce matériel est rédigé en anglais et est disponible sur le site Web du livre, à l'adresse suivante :



À l'instar des éditions précédentes, cette 8^e édition se caractérise par deux thèmes sous-jacents, plus ou moins indépendants de ce qui constitue la matière principale du livre (les tests d'hypothèse).

- Le premier thème souligne qu'il est essentiel d'observer les données avant de se précipiter sur un test d'hypothèse. Dans cette optique, j'examine en détail la représentation graphique des données, la recherche d'éventuelles valeurs aberrantes et la vérification des conditions d'application (tout cela à grand renfort de graphiques). J'essaie de procéder ainsi avec tout ensemble de données, même lorsque cet ensemble est destiné à illustrer une technique statistique complexe.
- Le second thème évoque l'importance du lien entre le test statistique à employer et les questions théoriques posées par l'expérience. Pour souligner ce lien, j'ai recours à des exemples authentiques en vue de faire comprendre aux étudiants l'objectif poursuivi par l'expérience et les prédictions liées à la théorie. À cette fin, je choisis parfois un exemple important sur lequel s'axe une section entière, voire un chapitre entier. C'est ainsi que des données intéressantes relatives à l'illusion de la lune, extraites d'une étude bien connue de Kaufman et Rock (1962), sont utilisées dans diverses formes du test *t*

(chapitre 7), et que la majeure partie du chapitre 12 s'articule autour d'une étude importante de Siegel (1975) sur la tolérance à la morphine. Chacun de ces exemples devrait directement intéresser les étudiants. L'accent mis sur les tailles d'effet dans la présente édition souligne également combien il est indispensable de bien réfléchir aux données recueillies et aux questions auxquelles on tente de répondre.

Même si personne ne qualifierait ce livre de controversé, j'ai néanmoins estimé important de me prononcer sur certains sujets de dissension. Après tout, l'intérêt de la statistique réside en partie dans les controverses qui la caractérisent. J'affirme par exemple que les échelles de mesure ne méritent pas toute l'importance que certains leur accordent, et je plaide pour une façon particulière de considérer les analyses de variance impliquant des échantillons de tailles inégales (à moins qu'il n'existe une raison impérieuse de procéder autrement). Je ne m'attends pas à ce que tous les professeurs partagent mes opinions, et en réalité, j'espère plutôt que ce ne sera pas le cas. Cela permet de présenter aux étudiants des positions divergentes et ainsi de les aider à comprendre les problèmes. Il me semble par ailleurs qu'il serait frustrant et malhonnête vis-à-vis des étudiants de leur présenter plusieurs procédures différentes de comparaisons multiples (ce que je fais) puis d'en rester là sans leur fournir aucune recommandation quant à la procédure la mieux adaptée à leur problème précis. Bien que j'aie réduit le nombre de procédures que j'aborde, je m'efforce d'expliquer à l'étudiant ce qu'il faut prendre en compte dans le choix d'une approche.

Il existe un manuel de solutions pour les étudiants, avec des solutions détaillées pour une partie des exercices impairs, disponibles à l'adresse suivante :



Solutions détaillées
des exercices impairs



www.lienmini.fr/05526-reponses-ex

À propos de l'auteur

Le Professeur Howell est professeur émérite de l'université du Vermont. Après avoir obtenu son Ph.D. à l'université de Tulane en 1967, il a rejoint l'université du Vermont jusqu'en 2002, année où il prit sa retraite alors qu'il était président du département de la psychologie. Il a également passé à deux reprises une année dans deux universités de Grande-Bretagne en tant que professeur invité.

Le Professeur Howell est l'auteur de plusieurs livres et de nombreux articles et chapitres de livre. Même émérite, il continue d'écrire et est co-éditeur, avec Brian Everitt, de l'encyclopédie des statistiques en sciences comportementales, publiée par Wiley. Il a rédigé une série de chapitres dans divers livres à propos des statistiques et des designs de recherche.

Le Professeur Howell vit maintenant dans le Colorado où il apprécie la neige et s'adonne à sa passion pour le ski et la randonnée.



Compléments numériques

Cet ouvrage offre divers compléments numériques, disséminés au fil des pages.

Pour les étudiants :



+ de 450 exercices supplémentaires ;



+ de 50 défis ;



180 flash-cards pour réviser ;



18 résumés audio ;



+ de 50 sites Internet pour aller plus loin dans votre étude.

Pour y accéder,

**Flashez le QR code avec votre
téléphone ou votre tablette**



OU

Tapez l'URL dans votre navigateur



Pour les professeurs :



Des supports de cours.

Pour y accéder, rendez-vous sur le site de l'éditeur à l'adresse suivante :

www.deboecksuperieur.com/site/305526

Concepts de base

MOTS-CLÉS

Analyse exploratoire de données, données catégorielles, données de mesure, données qualitatives, données quantitatives, échantillon, échantillon aléatoire, échelle d'intervalles, échelle de rapports, échelle nominale, échelle ordinale, inférence statistique, paramètre, population, répartition aléatoire, statistique, statistique descriptive, validité externe, validité interne, variable, variable indépendante, variable continue, variable dépendante, variable discrète

SOMMAIRE

1. Termes importants	2
2. Statistique descriptive et inférence statistique.....	6
3. Échelles de mesure.....	7
4. Utilisation des ordinateurs	11
5. À propos de cette nouvelle édition.....	12

OBJECTIF

Examiner le type de problèmes présentés dans ce livre et les questions liées à la sélection d'une procédure statistique.

Le stress est un problème qui se pose à chacun de nous au cours de notre vie. Il naît dans nos interactions quotidiennes avec notre entourage, dans nos interactions avec l'environnement, à l'approche d'un examen et, pour de nombreux étudiants, lors de la prise de conscience de la nécessité de suivre un cours de statistique. Si la plupart d'entre nous apprennent à réagir et à s'adapter au stress, le processus d'apprentissage est souvent lent et pénible. Ce préambule quelque peu morose peut ne pas sembler constituer une introduction très engageante pour un cours sur les statistiques, mais il conduit à la description d'un projet de recherche sur le stress qui illustre une série de concepts statistiques importants. J'ai moi-même été associé à un projet très similaire il y a un certain nombre d'années. Cet exemple est donc très loin d'être hypothétique.

EXEMPLE

PROGRAMME DE GESTION DU STRESS



Un groupe d'éducateurs a mis sur pied un cours destiné à apprendre à des lycéens à maîtriser leur stress et l'effet de la gestion du stress sur l'estime de soi. Les éducateurs ont toutefois besoin d'un examinateur externe qui soit à même de les renseigner sur le bon fonctionnement du cours, et en particulier de leur dire si les étudiants qui ont suivi le cours ont une meilleure estime d'eux-mêmes que ceux qui ne l'ont pas suivi. Supposons pour le moment que nous soyons chargés de mettre au point une évaluation de leur programme. Notre étude ne sera pas complète, mais elle illustrera quelques-uns des problèmes liés à la mise au point et à l'analyse d'expériences et quelques-uns des concepts statistiques avec lesquels vous devez vous familiariser.

1. Termes importants

Même si le programme de maîtrise du stress a été conçu pour l'ensemble des lycéens de France, il est évidemment impossible de le leur appliquer. Tout d'abord, ces lycéens sont beaucoup trop nombreux. En outre, il ne sert à rien d'appliquer un programme à tout le monde avant d'être renseigné sur son utilité. Plutôt que de nous concentrer sur l'ensemble de la population des lycéens, nous allons prélever de cette population un échantillon à qui nous appliquerons le programme. Mais nous n'allons pas prélever n'importe quel échantillon. Nous souhaitons prélever un **échantillon aléatoire**, même si j'expliquerai bientôt que les échantillons véritablement aléatoires sont en règle générale peu pratiques, voire impossibles. Pour prélever un échantillon aléatoire, nous suivrons un ensemble spécifique de procédures visant à faire en sorte que chaque élément (étudiant) de la population ait une chance égale d'être choisi. (L'exemple typique pour illustrer un échantillon aléatoire est le fait de mettre des noms dans un chapeau et de tirer au sort à l'aveugle. Bien que presque personne ne fasse jamais exactement cela, c'est une bonne illustration de ce que nous avons à l'esprit). Lorsque nous aurons prélevé notre échantillon de lycéens, nous répartirons

ceux-ci de manière aléatoire en deux groupes : l'un, comprenant la moitié des lycéens, bénéficiera du programme de maîtrise du stress, l'autre n'en bénéficiera pas.

Cette description a déjà introduit plusieurs concepts qui doivent être précisés : population, échantillon, échantillon aléatoire et répartition aléatoire. Une **population** est l'ensemble des événements (scores des étudiants, revenus des personnes, vitesses de course des rats, etc.) qui vous intéressent. Ainsi, si vous vous intéressez aux scores obtenus en matière de maîtrise du stress par tous les lycéens de France, l'ensemble de tous les scores des lycéens formerait une population, qui comprendrait dans ce cas plusieurs millions d'éléments. Par contre, si vous vous intéressez uniquement aux scores obtenus en matière de maîtrise du stress par les étudiants de terminale du lycée d'Olivet, en Mayenne (un village de moins de 500 habitants), la population ne comprendra qu'une quinzaine d'éléments.

Population

Il faut en conclure que les populations peuvent être de toutes tailles. Elles peuvent aller d'un ensemble de nombres relativement réduit, et donc facile à rassembler, à un ensemble de nombres important, mais fini, qui, en pratique, serait compliqué à rassembler dans son entièreté. En fait, il peut s'agir d'un ensemble de nombres infini, comme l'ensemble de tous les dessins que pourraient théoriquement faire les étudiants, population qui est impossible à rassembler. Malheureusement, les populations qui nous intéressent sont généralement très vastes. Il en résulte en pratique que nous ne mesurons (presque) jamais de populations entières. Nous sommes obligés de prélever de la population un **échantillon** d'observations, que nous utiliserons en vue d'inférer quelque chose à propos des caractéristiques de cette population.

Échantillon

En supposant que l'échantillon soit vraiment aléatoire, nous pouvons non seulement estimer certaines caractéristiques de la population, mais encore nous faire une bonne idée de la précision de nos estimations. Dans la mesure où l'échantillon n'est pas aléatoire, nos estimations peuvent être crédibles ou non, puisque l'échantillon peut ou non refléter avec précision l'ensemble de la population. Le caractère aléatoire comporte au moins deux aspects que nous devons prendre en considération. Le premier concerne la question de savoir si l'échantillon est représentatif de la population dont il est extrait. Ceci implique essentiellement l'échantillonnage aléatoire de la population et conduit à ce que l'on appelle la **validité externe**. La validité externe fait référence à la question de savoir si l'échantillon reflète la population. Un échantillon prélevé dans une petite ville d'Alsace ne donnerait pas une estimation valable du pourcentage de la population nord-africaine en France, pas plus qu'un échantillon prélevé uniquement en Bretagne. En revanche, un échantillon d'une petite ville d'Alsace pourrait nous fournir une estimation raisonnable du temps de réaction des gens face à des stimuli présentés sur un écran. Nous nous trouvons là face à l'un des problèmes que pose l'étude de l'échantillonnage aléatoire. Un échantillon non aléatoire de sujets ou de participants peut toujours nous être utile si nous pouvons nous convaincre et convaincre les autres qu'il ressemble beaucoup à ce que nous obtiendrions si nous pouvions prélever un échantillon véritablement aléatoire. Cependant, si notre échantillon non aléatoire n'est pas représentatif de ce que nous obtiendrions avec un échantillon véritablement aléatoire, notre capacité à tirer des conclusions est compromise et nos résultats pourraient alors être très trompeurs.

Validité externe

Avant de poursuivre, éclaircissons un point qui est à l'origine de nombreuses confusions. En fait, l'échantillon choisi par quelqu'un peut correspondre à la population choisie par quelqu'un d'autre. Par exemple, si je devais mener une étude sur l'efficacité de ce livre en tant qu'outil pédagogique, je pourrais considérer les scores

obtenus par une classe donnée à un examen comme un échantillon non aléatoire de la population composée des scores de tous les étudiants utilisant ou susceptibles d'utiliser ce livre. Par contre, le professeur de cette classe ne s'intéresse sans doute pas énormément à ce livre et se soucie uniquement de ses étudiants. Pour lui, ce même ensemble de scores constituerait donc une population. Par ailleurs, quelqu'un qui s'intéresse à l'enseignement des statistiques pourrait considérer ma population (tous les utilisateurs de mon livre) comme un échantillon non aléatoire prélevé d'une population plus large (tous les utilisateurs d'un livre de statistique). La définition d'une population dépend donc de ce qu'on souhaite étudier.

Dans notre étude sur le stress, il est très peu probable que nous puissions envisager sérieusement de prélever un échantillon réellement aléatoire des lycéens de France afin de leur appliquer le programme de maîtrise du stress. C'est tout simplement irréalisable. Alors, comment tirer profit des méthodes et procédures statistiques basées sur la condition d'application de l'échantillonnage aléatoire ? La seule façon d'y parvenir est d'être attentif à n'appliquer ces méthodes et procédures que lorsque nous sommes convaincus que nos résultats sont généralement représentatifs de la population étudiée. Si nous ne pouvons pas établir cette hypothèse, nous devons revoir la conception de notre étude. La question n'est pas tant une question de raffinement statistique qu'une question de bon sens. Dans la mesure où, selon nous, notre échantillon n'est pas représentatif des lycéens de France, nous devons être prudents lors de l'interprétation des résultats. Dans la mesure où notre échantillon est représentatif de la population, nos estimations seront valides.

Répartition aléatoire

Le deuxième aspect du caractère aléatoire concerne la **répartition aléatoire**. Si la sélection aléatoire touche la *source* de nos données et est importante pour la généralisation des résultats de notre étude à l'ensemble de la population, la répartition aléatoire des individus (une fois sélectionnés) dans les groupes de traitement est essentielle pour l'intégrité de notre expérience. Nous faisons ici référence à ce que l'on appelle la **validité interne**. Nous voulons nous assurer que les résultats que nous obtenons sont le fruit des différences dans la façon dont nous traitons nos groupes, et non le résultat des personnes que nous plaçons dans ces groupes. Si, par exemple, nous plaçons tous les étudiants timides de notre échantillon dans un même groupe et tous les étudiants sûrs d'eux dans un autre, il est très probable que nos résultats soient tout autant, voire plus, une fonction de la répartition des groupes que des traitements que nous avons appliqués à ces groupes. Laisser les étudiants se répartir eux-mêmes dans des groupes présente des risques similaires et est généralement une mauvaise idée, à moins que cela ne soit la nature du phénomène étudié. Dans la pratique, la répartition aléatoire est généralement bien plus importante que le fait de prélever un échantillon aléatoire. Dans les études médicales, la répartition aléatoire peut devenir extrêmement complexe en raison de la nature de l'étude, et il y a parfois des personnes dont le travail consiste simplement à contrôler cet aspect de l'expérience.

Validité interne

Une fois abordées la sélection des individus et leur répartition dans des groupes de traitement, il est temps de considérer la façon dont nous traiterons chaque groupe et comment nous caractériserons les données qui en résulteront. Comme nous souhaitons étudier la capacité des individus à maîtriser le stress et à conserver une bonne estime de soi sous différents types de traitements, et comme la réaction au stress est fonction de plusieurs variables, le choix des variables à étudier constitue un aspect essentiel de la planification de l'étude. Une **variable** est une propriété d'un objet ou

Variable

événement qui peut prendre différentes valeurs. Ainsi, la couleur des cheveux est une variable parce qu'il s'agit de la propriété d'un objet (les cheveux) et qu'elle peut prendre différentes valeurs (cheveux bruns, blonds, roux, gris, etc.). Pour ce qui est de notre évaluation du programme de gestion du stress, les facteurs tels que le traitement que nous utilisons, la confiance en soi de l'étudiant, le soutien social, le sexe, le degré de contrôle personnel et le groupe de traitement constituent tous des variables pertinentes.

En statistiques, nous distinguons les variables indépendantes des variables dépendantes. Dans notre exemple, l'appartenance à un groupe est une **variable indépendante**, car nous la contrôlons. Nous décidons quels seront les traitements et qui bénéficiera de chaque traitement. Nous décidons que ce groupe-ci recevra le traitement de gestion du stress et que ce groupe-là ne le recevra pas. Si nous avons comparé des hommes et des femmes, il est clair que nous ne contrôlons pas le sexe de la personne, mais nous décidons des sexes à étudier (ce qui n'est pas une décision difficile) et que nous voulons comparer des hommes et des femmes. D'autre part, les données – comme les scores d'estime de soi, les scores de sentiment de contrôle personnel, etc. qui en résultent – sont les **variables dépendantes**. Fondamentalement, l'étude porte sur les variables indépendantes, et les résultats de l'étude (les données) sont les variables dépendantes. Les variables indépendantes peuvent être quantitatives ou qualitatives, discrètes ou continues, alors que les variables dépendantes sont généralement, mais certainement pas toujours, quantitatives et continues, termes que nous allons définir à présent.¹

Variable indépendante

Variable dépendante

Nous faisons une distinction entre les **variables discrètes**, telles que le sexe ou la classe de lycée, qui ne prennent qu'un nombre limité de valeurs, et les **variables continues**, telles que l'âge et le score d'estime de soi, qui peuvent prendre, du moins en théorie, n'importe quelle valeur entre le point le plus bas et le point le plus haut de l'échelle.² Comme vous le verrez, cette distinction joue un rôle important dans la façon dont nous traitons les données.

Variable discrète

Variable continue

La distinction entre variables continues et variables discrètes est étroitement liée à la distinction entre données quantitatives et données catégorielles. Par **données quantitatives** (parfois appelées **données de mesure**), on entend les résultats de tout type de mesure, par exemple les points obtenus à un test, le poids des personnes, les scores figurant sur une échelle d'autosatisfaction, etc. Dans tous les cas, un certain type d'instrument (au sens le plus large) a été utilisé pour mesurer quelque chose, et nous sommes intéressés de savoir « à quel point » un objet possède une certaine propriété.

Données quantitatives

Données de mesure

Les **données catégorielles** (également appelées **données qualitatives**) sont quant à elles illustrées par des affirmations telles que « Notre étude comprend 34 femmes et 26 hommes » ou « Quinze personnes ont été classées « très angoissées », 33 « moyennement angoissées » et 12 « peu angoissées ». » Ici, nous classons des éléments par catégories, et nos données se composent de fréquences pour chaque catégorie (d'où le nom « données catégorielles »). Même si notre étude incluait des centaines

Données catégorielles

Données qualitatives

1. Beaucoup ont du mal à se rappeler quelle est la variable dépendante et quelle est la variable indépendante. Remarquez que les mots « dépendante » et « données » commencent tous les deux par « d ».

2. En fait, une variable continue est une variable pour laquelle *n'importe quelle* valeur comprise entre les extrêmes de l'échelle (ex. 32.485687...) est possible. Cependant, dans la pratique, nous considérons qu'une variable est continue si elle peut prendre de nombreuses valeurs différentes et qu'elle est discrète si elle ne peut prendre que quelques valeurs.

d'individus, les résultats (données) ne consisteraient qu'en deux ou trois nombres : le nombre d'individus appartenant à chaque catégorie d'anxiété. Par contre, si, au lieu de classer les personnes selon trois catégories (très, moyennement ou peu angoissées), nous avons assigné à chacune d'entre elles un score basé sur une échelle plus ou moins continue d'angoisse, nous serions confrontés à des données de mesure et ces données comprendraient les scores obtenus par chaque individu pour cette variable. Notez que dans les deux situations, la variable s'appelle *angoisse*. Tout comme la plupart des distinctions, la distinction entre données de mesure et données catégorielles peut sembler exagérée. Elle n'en est pas moins utile, et il est souvent aisé, dans la pratique, de savoir si une variable est une donnée de mesure ou une donnée catégorielle.

2. Statistique descriptive et inférence statistique

Revenons-en à notre programme d'intervention sur le stress. Une fois que nous avons choisi les variables à mesurer et que les écoles ont appliqué le programme à leurs étudiants, nous nous trouvons confrontés à une série de données brutes : les scores. Deux sous-domaines de la statistique s'occupent de la manière dont nous utilisons ces données.

Statistique descriptive

Lorsque notre objectif est uniquement de décrire un ensemble de données, nous avons recours à la **statistique descriptive**. Ainsi, l'une des premières choses à faire avec nos données est d'en tracer un graphique, de calculer les moyennes et d'autres indices, puis de rechercher les scores extrêmes ou les distributions peu courantes. Ces procédures relèvent de la statistique descriptive parce qu'elles visent essentiellement à décrire les données. Le domaine de la statistique descriptive était autrefois méprisé, considéré comme un domaine peu intéressant surtout apprécié par des statisticiens soucieux de tracer des graphes dénaturés pour des publications telles que le *Nouvel observateur*. Dans les années 70, toutefois, John Tukey a élaboré ce qu'il a appelé la statistique exploratoire, ou **analyse exploratoire de données**. Il a montré la nécessité d'accorder une attention toute particulière aux données et de les examiner en détail avant d'avoir recours à des procédures plus élaborées. Certaines des innovations de Tukey ont fait leur chemin dans le courant dominant de la statistique et seront étudiées dans les chapitres suivants ; d'autres n'ont pas été aussi bien accueillies. Cependant, l'accent mis par Tukey sur la nécessité d'examiner de près vos données a eu une grande influence, en partie en raison de la haute réputation de Tukey en tant que statisticien.

Analyse exploratoire de données

Inférence statistique

Lorsque nous aurons décrit nos données en détail et aurons bien compris ce que les nombres ont à nous dire visuellement, nous nous intéresserons plus particulièrement à ce qu'on appelle l'**inférence statistique**. En fait, la plus grande partie de ce livre sera consacrée à cette partie de la statistique qu'on appelle inférentielle. En mettant au point notre expérience sur l'effet du stress sur l'estime de soi, nous avons reconnu qu'il n'était pas possible de mesurer l'ensemble de la population, c'est pourquoi nous avons prélevé un échantillon de cette population. Toutefois, nos vraies questions portent sur la population proprement dite. Nous pourrions par exemple nous demander quel est le score moyen (quelle est la tendance centrale) d'autosatisfaction de l'ensemble de la population des étudiants qui auraient pu suivre notre programme, alors que nous ne disposons que du score moyen obtenu par un échantillon d'étudiants ayant réellement suivi le programme.

Une mesure, comme le score moyen d'autosatisfaction, qui se réfère à l'ensemble d'une population est appelée **paramètre**. Cette même mesure est appelée **statistique** lorsqu'elle est calculée à partir d'un échantillon de données que nous avons rassemblées. Les paramètres sont les entités *réelles* d'intérêt, tandis que les statistiques correspondantes sont en fait des *suppositions*. Si la plus grande partie de ce livre traite de statistiques d'échantillon (donc de suppositions), cela ne doit pas nous faire oublier que ce qui nous intéresse réellement, c'est le paramètre correspondant de la population. Nous voulons *inférer* quelque chose sur les caractéristiques de la population (paramètres) à partir de ce que nous savons des caractéristiques de l'échantillon (statistiques). Dans notre étude hypothétique, nous souhaitons savoir si le score moyen d'autosatisfaction d'une population d'étudiants potentiellement intéressés de participer à notre programme est supérieur ou inférieur au score moyen d'autosatisfaction des autres étudiants. De nouveau, ce problème relève de l'inférence statistique, puisque nous inférons des caractéristiques propres à des populations à partir de caractéristiques propres à des échantillons.

Paramètre
Statistique

3. Échelles de mesure

Les échelles de mesure sont considérées par certains auteurs comme un sujet crucial et par d'autres comme un sujet inutile. Si je penche plutôt pour la deuxième attitude, il n'en est pas moins important pour vous de vous familiariser quelque peu avec le problème en général (Vous ne devez pas nécessairement approuver quelque chose pour trouver que cela vaut la peine de l'examiner. Après tout, les médecins se targuent de savoir beaucoup de choses sur les maladies même s'ils en défendent rarement l'utilité). Cet examen présente en outre l'avantage de vous montrer que la statistique n'est pas un ensemble de faits immuables, mais plutôt un ensemble de faits rassemblés permettant une grande diversité d'interprétations et d'opinions.



FOCUS SUR...

L'IMPORTANCE DES ÉCHELLES DE MESURE SELON S.S. STEVENS

S.S. Stevens a sans doute été le principal défenseur de l'importance des échelles de mesure dans le choix de procédures statistiques.³ Zumbo et Zimmerman (2000) ont discuté les échelles de mesure en profondeur et montré que le système de Stevens devait être replacé dans son contexte historique. Dans les années 1940 et 1950, Stevens tentait de défendre la recherche en psychologie contre certains chercheurs des « sciences dures » qui avaient une

vision restrictive de la mesure en sciences. Il entendait rendre la psychologie « respectable. » Stevens consacra la plus grande partie de sa prestigieuse carrière scientifique à développer des échelles de mesure dans le domaine de la psychophysique, faisant avancer nos connaissances de manière importante. Toutefois, en dehors de ce secteur, peu d'efforts ont été consentis en psychologie pour développer le genre d'échelle que Stevens avait à

l'esprit, et l'intérêt ne s'est pas révélé très grand. Les critiques que Stevens trouvait tellement menaçantes se sont, pour la majeure partie, évaporées et, avec elles, la conviction que les échelles de mesure influencent de manière déterminante les procédures statistiques qu'il convient de mettre en œuvre. Mais les termes qui décrivent ces mesures restent importants.

3. Le chapitre 1 du *Handbook of Experimental Psychology* de Stevens (1951) constitue une excellente référence pour ceux qui souhaitent examiner les questions mathématiques importantes sous-tendant cette position.

3.1. Les échelles nominales

Échelle nominale

Dans un sens, les **échelles nominales** ne sont pas des échelles ; elles ne classent pas les éléments sur une échelle selon une dimension, mais leur collent plutôt une étiquette. Les variables telles que le sexe et l'adhésion à un parti politique sont des variables nominales. Ces données catégorielles sont généralement mesurées sur une échelle nominale parce qu'on se contente d'assigner des étiquettes de catégories (ex. masculin ou féminin ; droite, gauche ou centre) à des observations. L'ensemble des numéros de dossard assignés à des athlètes donne un exemple numérique d'échelle nominale. Souvent, ces nombres n'ont d'autre signification que celle d'étiquettes pratiques servant à distinguer les athlètes. On pourrait tout aussi bien utiliser des lettres ou des photos d'animaux.

3.2. Les échelles ordinales

Échelle ordinale

Parmi les vraies échelles, les plus simples sont les **échelles ordinales**, qui classent les personnes, objets ou événements le long d'un continuum. Les grades dans la marine française en constituent un excellent exemple. Un capitaine de frégate n'a pas le prestige d'un capitaine de vaisseau, qui est à son tour moins haut placé qu'un vice-amiral. Toutefois, rien ne porte à croire que la *différence* de prestige entre un capitaine de frégate et un capitaine de vaisseau soit la même qu'entre un capitaine de vaisseau et un vice-amiral. En psychologie, prenons l'exemple de l'échelle de Holmes et Rahe (1967) portant sur le stress de la vie. À l'aide de cette échelle, on compte (parfois avec des pondérations différentielles) le nombre de changements (mariage, déménagement, nouvel emploi, etc.) intervenus au cours des six dernières années de la vie de quelqu'un. Une personne obtenant un score de 20 est supposée avoir été confrontée à davantage de stress qu'une personne ayant un score de 15, et cette dernière est quant à elle supposée avoir connu davantage de stress qu'une personne obtenant un score de 10. Les personnes sont donc classées, en termes de stress, en fonction du nombre de changements intervenus récemment dans leur vie. Il s'agit d'un exemple d'échelle ordinale parce qu'aucune implication n'est faite concernant les différences entre les points de l'échelle. On ne suppose pas, par exemple, que la différence entre les points 10 et 15 représente la même différence de stress que la différence entre les points 15 et 20. Les distinctions de ce type sont liées aux échelles d'intervalles.

3.3. Les échelles d'intervalles

Échelle d'intervalles

Une **échelle d'intervalles** est une échelle de mesure pour laquelle on peut à juste titre parler de différences entre les points de l'échelle. L'échelle de température Celsius en constitue un bon exemple puisqu'une différence de 5 points a la même signification sur toute l'échelle. La différence de température entre 5° et 10° est la même que la différence entre 30° et 35°. Remarquez que cette échelle possède également les propriétés des deux précédentes. Par contre, les échelles d'intervalles ne nous permettent pas de parler utilement de rapports. Ainsi, on ne peut par exemple dire qu'une température de 10° est deux fois moins élevée qu'une température de 20°, ou deux fois plus élevée qu'une température de 5°. Pour cela, il faut utiliser une échelle de rapports (À ce propos, il est utile de remarquer que quand on effectue des

conversions tout à fait correctes d'une échelle d'intervalles à une autre – par exemple, de l'échelle Celsius à Fahrenheit –, on ne conserve même pas les mêmes rapports. Le rapport entre 41° et 77° sur l'échelle Fahrenheit n'est donc pas égal au rapport entre 5° et 25° sur l'échelle Celsius, même si ces valeurs renvoient à des températures analogues. Cela souligne le caractère arbitraire des rapports lorsqu'il est question d'échelles d'intervalles).

3.4. Les échelles de rapports

Une **échelle de rapports** est une échelle possédant un vrai point zéro. Notez que le zéro doit être un vrai point zéro et non un point arbitraire tel que 0°F ou même 0 °C (Un vrai point zéro est le point correspondant à l'absence de la chose mesurée. Comme 0°F et 0 °C ne représentent pas l'absence de température ou de mouvement moléculaire, on ne peut parler de vrai point zéro). Comme exemples d'échelles de rapports, citons les échelles physiques habituelles que sont la longueur, le volume, le temps, etc. Avec ces échelles, non seulement nous avons les propriétés des échelles précédentes, mais nous pouvons également parler de rapports. On peut dire qu'en termes physiques, une durée de 10 secondes est deux fois plus longue qu'une durée de 5 secondes, qu'un poids d'1 kg est deux fois moins lourd qu'un poids de 2 kg, etc.

Échelle de rapports

On pourrait penser que le type d'échelle avec lequel on travaille est évident. Malheureusement, en particulier avec le type de mesures rassemblées en psychologie, c'est rarement le cas. Le type d'échelle que vous utilisez dépend souvent de l'objectif que vous poursuivez en l'utilisant. Parmi des adultes de même corpulence, le poids est une échelle de rapports de combien ils pèsent, mais probablement au mieux une échelle ordinale de la forme physique ou de la santé.

EXEMPLE

TYPE D'ÉCHELLE DE LA TEMPÉRATURE D'UNE PIÈCE SELON SON UTILISATION



Comme exemple de type de mesure ayant une échelle qui dépend de son utilisation, prenons la température d'une maison. On considère généralement les degrés Celsius comme une échelle d'intervalles. C'est l'exemple qui vient d'être cité, et pour un physicien, il ne fait pas de doute que la différence entre 16 °C et 18 °C est exactement la même que la différence entre 24 °C et 26 °C. Par contre, si l'on mesure la température comme étant un indicateur du *confort* plutôt que de l'activité moléculaire, ces mêmes nombres ne constituent plus une échelle d'intervalles. Si la température d'une pièce passait de 16 °C à 18 °C, une personne assise dans cette pièce ne manquerait pas de le remarquer (et ne s'en plaindrait pas). Par contre, si la température passait de 24 °C à 26 °C, on ne pourrait en dire autant. Il est donc important de souligner que *c'est la variable sous-jacente que nous mesurons (par exemple le confort), et non les nombres eux-mêmes, et que c'est elle qui importe dans la définition de l'échelle*. En termes de confort, les degrés Celsius ne constituent pas une échelle d'intervalles. Ils ne forment même pas une échelle ordinale parce que le confort est susceptible d'augmenter avec la température jusqu'à un certain point et puis de commencer à diminuer.

3.5. Le rôle des échelles de mesure

Comme je l'ai dit, les auteurs ne sont pas d'accord sur l'importance attribuée aux échelles de mesure. Certains ignorent complètement le problème tandis que d'autres y consacrent des manuels entiers. Un point de vue raisonnable (en d'autres termes, mon point de vue) considère que l'essentiel est d'impérativement distinguer les

nombres qu'on rassemble des objets ou événements auxquels ils se réfèrent. Nous en avons parlé dans l'exemple de la température d'une pièce: l'échelle (d'intervalles ou ordinaire) dépendait de notre choix de mesurer un attribut physique de la température ou encore ses effets sur les personnes (par ex. en termes de confort). Une différence de 2 °C est *physiquement* la même à n'importe quel endroit de l'échelle, mais une différence de 2 °C lorsqu'il fait déjà chaud dans la pièce *paraît* moins importante que lorsqu'il fait relativement frais. Autrement dit, nous avons une échelle d'intervalles des unités physiques, mais uniquement une échelle ordinaire du confort (à nouveau, jusqu'à un certain point).

Comme les tests statistiques utilisent des nombres sans tenir compte des objets ou événements auxquels ces nombres renvoient, on peut effectuer n'importe quelle opération mathématique standard (addition, multiplication, etc.) quelle que soit la nature de l'échelle sous-jacente. Lord (1953) a rédigé à ce sujet un excellent article, divertissant et à conseiller, intitulé « On The Statistical Treatment of Football Numbers »; selon lui, ces nombres se prêtent à n'importe quel traitement puisque « les nombres ne se rappellent pas leur origine » (p. 751).

Le problème se pose lorsqu'il s'agit d'interpréter les résultats d'un certain type de manipulation statistique. À ce stade, il faut se demander si les résultats statistiques sont liés de manière pertinente aux objets ou événements en question. Il n'est plus question ici d'un problème statistique, mais méthodologique. Aucune procédure *statistique* ne peut dire si le fait qu'un groupe ait obtenu des scores plus élevés qu'un autre à un questionnaire sur l'angoisse est révélateur des différences entre les groupes pour ce qui est des niveaux d'angoisse sous-jacents. En outre, se féliciter du fait que le questionnaire fournisse une échelle de rapports des *scores* d'angoisse (un score de 50 est deux fois plus élevé qu'un score de 25), c'est perdre de vue le fait que notre but était de mesurer l'angoisse, qui peut ne pas augmenter de manière cohérente proportionnellement aux scores. Nos tests statistiques ne peuvent s'appliquer qu'aux nombres obtenus, et la validité des affirmations relatives aux objets ou événements que nous pensons mesurer dépend essentiellement de notre connaissance de ces objets ou événements et non de l'échelle de mesure. Nous faisons de notre mieux pour garantir que nos mesures soient le plus proche possible de ce que nous souhaitons mesurer, mais nos résultats ne sont finalement que les nombres que nous obtenons et notre foi en la relation existant entre ces nombres et les objets ou événements sous-jacents.⁴

La conclusion qui s'impose de cette discussion – et que ce livre retiendra – est que l'échelle de mesure sous-jacente n'est pas essentielle dans notre choix de techniques statistiques. Évidemment, l'interprétation des résultats de ces manipulations statistiques requiert un certain bon sens. Il ne serait pas très intelligent de conclure qu'une peinture jugée excellente par une personne et détestable par une autre doit pour cela être qualifiée d'assez médiocre.

4. Comme l'a souligné Cohen (1965), « Thurstone a dit un jour qu'en psychologie, nous mesurons les hommes via leur ombre. En effet, en psychologie clinique, nous mesurons souvent les hommes via leur ombre pendant qu'ils dansent dans une salle de bal illuminée par les reflets d'un vieux miroir polyédrique tournant » (p. 102).

4. Utilisation des ordinateurs

Quand j'ai écrit la première édition de ce livre, il y a cinquante ans, la plupart des analyses statistiques étaient effectuées sur des calculatrices, et les livres étaient rédigés en conséquence. Cependant, les méthodes ont changé, et la plupart des calculs sont maintenant effectués par ordinateur.

Ce livre s'adapte à la disponibilité accrue des ordinateurs en les incluant dans la discussion. L'aspect informatique gagnera considérablement en importance à mesure que nous avancerons dans le livre et que les calculs deviendront plus laborieux. Pour les procédures assez simples, les formules de calcul nous informent à propos de la définition du concept. Par exemple, la formule d'un écart-type ou d'un test t définit et permet de comprendre ce qu'est réellement un écart-type ou un test t . Dans ce cas, l'accent est mis sur le calcul manuel, même si le livre donne également des exemples de solutions par ordinateur. Dans la suite du livre, lorsque nous abordons par exemple la régression multiple ou les techniques de traitement des données manquantes, les formules nous apprennent beaucoup moins. La formule de dérivation des coefficients de régression avec cinq variables prédictives ne va probablement pas améliorer votre compréhension des statistiques qui en résultent. En effet, elle ne ferait qu'embrouiller les choses. Dans de tels cas, nous aurons presque uniquement recours à des solutions par ordinateur.

Mais quel logiciel statistique allons-nous utiliser ? De nombreux logiciels sont actuellement disponibles pour les chercheurs ou étudiants effectuant des analyses statistiques. Les principaux logiciels statistiques, qui permettront d'effectuer pour ainsi dire toutes les analyses dont vous aurez besoin dans le cadre de ce livre, sont Minitab®, SAS®, SPSS™, Stata® et S-PLUS. Ces logiciels sont tout à fait fiables et relativement simples à utiliser, et le centre informatique de toute université en possède généralement au moins un. (Mais les experts ne s'accordent pas sur la question de savoir lequel est le meilleur logiciel). De nombreux exemples de sorties SPSS apparaîtront dans ce livre. À l'origine, j'utilisais une variété de logiciels, mais des analystes m'ont suggéré de me limiter à un seul, et SPSS est de loin le plus courant dans les sciences comportementales. Vous pouvez trouver d'excellents fichiers de code SAS à l'adresse ci-contre, et beaucoup d'entre eux sont liés aux exemples de ce livre.

Si vous souhaitez télécharger un logiciel peu coûteux qui fera presque tout ce dont vous aurez besoin, je vous recommande OpenStat (voir ci-contre). Je trouve qu'il est un peu compliqué pour lire les fichiers de données, mais à part cela, il fonctionne bien. Vous pouvez l'ouvrir et saisir des exemples que vous rencontrez, ou vous pouvez voir ce qui se passe lorsque vous modifiez mes exemples.

Je dois mentionner un environnement de programmation appelé R, qui est un logiciel du domaine public disponible gratuitement. C'est un programme que j'adore (ou que je déteste, suivant ce que j'ai oublié depuis ma dernière utilisation). R est très flexible, de nombreux textes donnent le code R pour certaines de leurs analyses, et il a des centaines de fonctions spéciales que des personnes ont écrites et mises à la disposition du public. R devient étonnamment populaire dans de nombreuses disciplines différentes et je m'attends à ce que cela continue. Vous pouvez télécharger R à l'adresse ci-contre. Tout au long de ce livre, je ferai parfois référence à des pages Web que j'ai écrites et qui contiennent du code R permettant d'effectuer

 Code SAS



www.lienmini.fr/05526-sas

 OpenStat



www.lienmini.fr/05526-open

 R



www.lienmini.fr/05526-r-pro



Programmer en R


www.lienmini.fr/05526-progr-r

des calculs importants. Ces pages peuvent facilement être collées dans R. Une page Web qui vous permettra de commencer à utiliser R est disponible ci-contre.

Il me faut mentionner des tableurs largement disponibles tels que Excel. Ces programmes peuvent effectuer un certain nombre de calculs statistiques. Ils donnent également d'assez bons graphiques et constituent un excellent moyen de réaliser des calculs à la main. Ils vous obligent à calculer de façon logique et à retenir toutes les étapes intermédiaires en prévision d'examens ultérieurs. Par le passé, de tels programmes ont été critiqués quant à la précision de leurs résultats. Des extensions récentes qui ont été écrites pour eux ont considérablement augmenté la précision ; ne sous-estimez donc pas leur valeur. Les programmes comme Excel présentent également l'avantage d'être installés sur la plupart des ordinateurs.

5. À propos de cette nouvelle édition

À différents égards, cette édition représente une rupture par rapport aux éditions précédentes. Au fil des années, chaque édition a proposé une quantité raisonnable de nouveau matériel et la discussion a connu des modifications de manière à refléter les changements dans la théorie et la pratique. Dans cette édition, j'ai l'intention d'introduire un certain nombre de sujets supplémentaires que les étudiants doivent comprendre, et je ne peux pas réellement faire cela sans supprimer une partie du matériel plus ancien. Cependant, grâce à Internet, la matière sera en grande partie déplacée plutôt que supprimée. Par exemple, si vous préférez voir l'ancienne version, plus complète, du chapitre sur les tests de comparaison multiple ou de celui sur les modèles log-linéaires, vous pouvez télécharger la version précédente sur mon site Web. Il en va de même pour d'autres suppressions majeures.

La première édition de ce livre a été écrite en 1973-1974, alors que j'étais en congé sabbatique à l'Université de Durham, en Angleterre. À cette époque, la plupart des cours de statistique en psychologie s'adressaient aux psychologues expérimentaux, et une grande partie de ces travaux concernait l'analyse de la variance. L'ouvrage de B. J. Winer, *Statistical Principles in Experimental Design* (1962, 1971), a dominé une grande partie du domaine des statistiques psychologiques et a influencé la façon dont les expériences étaient conçues, menées et analysées. Ma première édition suivait en grande partie cette tendance, et à juste titre. Depuis lors, l'analyse statistique des données a connu des changements considérables. Bien que l'analyse de la variance reste une technique dominante, suivie de près par la régression multiple, de nombreux autres sujets ont fait leur entrée en scène. Par exemple, les mesures de taille d'effet sont beaucoup plus importantes qu'elles ne l'étaient à l'époque, la méta-analyse est passée de pratiquement rien à un moyen important de résumer un domaine, et le traitement des données manquantes a considérablement changé grâce à l'introduction des modèles linéaires mixtes et des méthodes d'imputation des données. Nous ne pouvons plus former des étudiants prêts à mener des recherches de qualité sans aborder ces questions.

Mais pour inclure ces sujets et d'autres similaires, il faut en mettre d'autres de côté si l'on veut que le texte reste d'une taille gérable. Par le passé, les procédures de comparaisons multiples, par exemple, ont fait l'objet d'une attention considérable avec de nombreuses approches concurrentes. Mais si vous regardez ce que les gens font réellement dans la littérature, la plupart des approches concurrentes présentent

un intérêt pour les théoriciens davantage que pour les praticiens. Je pense qu'il est bien plus utile d'écrire sur le traitement des données manquantes que d'énumérer toutes les alternatives pour effectuer des comparaisons entre moyennes. C'est d'autant plus vrai qu'il est désormais facile de mettre ces anciens documents à la disposition de tous ceux qui souhaiteraient les lire.

À quels changements pouvez-vous donc vous attendre ? Les sujets les plus importants à inclure, à ce jour, comprennent un traitement plus approfondi des données manquantes, l'utilisation de modèles linéaires mixtes pour traiter les mesures répétées, une plus grande importance accordée aux intervalles de confiance, et une introduction à la méta-analyse et aux études de cas unique, qui sont très importants dans les interventions cliniques. Je ne peux pas couvrir toute cette matière de manière aussi approfondie que je le souhaiterais, mais je peux au moins vous mettre sur la voie et vous indiquer les sources pertinentes. Dans de nombreux cas, je dispose de matériel supplémentaire que j'ai créé et qui est disponible sur mon site (voir ci-contre). Sur ce site, vous avez également accès à tous les fichiers de données, aux réponses détaillées des exercices impairs, aux erreurs et corrections, ainsi qu'à mes propres manuels sur SPSS. C'est à ce site que je fais référence lorsque je mentionne « le site Web de ce livre ». Suivez simplement les liens que vous y trouverez.

Le plus grand changement à venir dans le domaine des méthodes statistiques est peut-être le « rééchantillonnage ». Au cours des cent dernières années, nous avons résolu des problèmes statistiques en calculant une certaine statistique de test, puis en nous disant : « si la population dans laquelle j'ai prélevé mon échantillon présente certaines caractéristiques, alors je peux équitablement comparer ma statistique à la valeur appropriée trouvée dans un tableau ». Mais que faire si la population ne présente pas ces caractéristiques ? Eh bien, il reste à espérer qu'elle le fasse.

Grâce à la vitesse de calcul très élevée d'un ordinateur portable standard, nous pouvons éviter certaines de ces hypothèses sur la population. Nous pouvons déterminer à quoi ressemblera une statistique lorsqu'elle sera tirée de presque n'importe quelle population. Ces procédures de rééchantillonnage sont extrêmement puissantes et commencent à jouer un rôle important. Dans un avenir proche, les approches traditionnelles conserveront leur statut, ne serait-ce que parce que c'est ce que les gens ont appris à utiliser. Mais cela est en train de changer. Chaque fois que vous voyez une case à cocher sur l'un des écrans de votre logiciel qui indique quelque chose comme « simuler » ou « bootstrap », c'est le signe que le crépuscule des méthodes traditionnelles approche.

Pour anticiper ce changement, ce livre introduit des méthodes de rééchantillonnage en plus des approches habituelles. Ces méthodes ne vous compliqueront pas la vie, car elles sont faciles à comprendre. Tout ce que je veux que vous reteniez, ce sont les concepts sous-jacents. Je vous fournirai même de courts programmes que vous pourrez utiliser pour effectuer l'échantillonnage vous-même, mais vous pouvez les ignorer si vous le souhaitez. Les nouvelles méthodes vous aideront à mieux comprendre en quoi consistaient les anciennes.

Enfin, cette édition française propose à la figure 1.1 une synthèse des différentes procédures possibles selon certaines dimensions, telles que le type de données, le nombre de variables ou encore les questions que l'on se pose. Ce schéma peut, dans une certaine mesure, être utilisé dans le choix d'un test statistique, avec toutes les précautions d'usage qui s'imposent notamment en termes de vérification de conditions d'application.



Site Web du livre



www.lienmini.fr/05526-howell

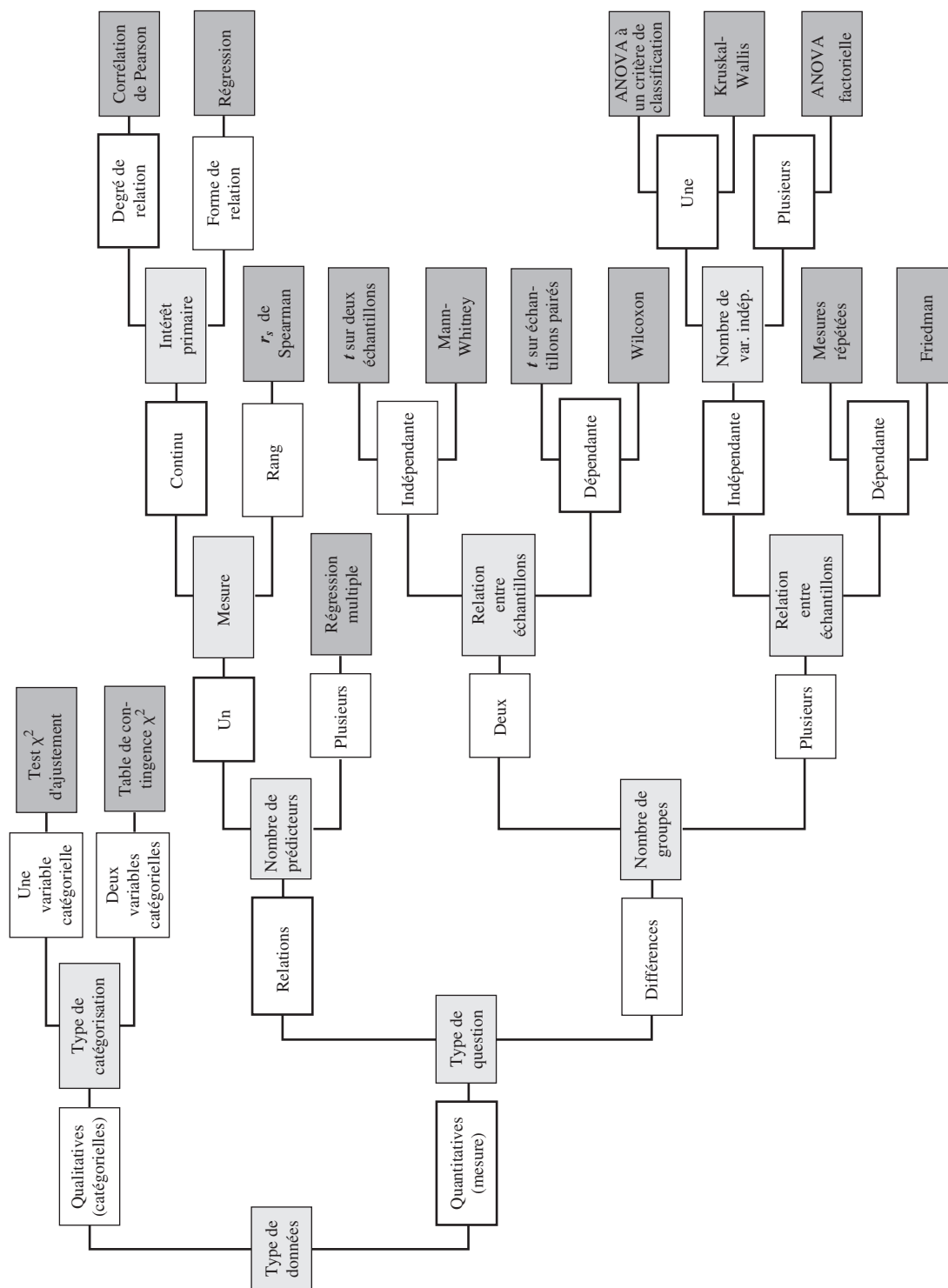


FIGURE 1.1. Arbre de décision

L'essentiel



Résumé

On s'intéresse généralement à une population d'unités d'intérêt. En pratique, on a rarement accès à l'ensemble de cette population, mais plutôt à un échantillon, c'est-à-dire un sous-ensemble d'unités issues de la population d'intérêt. Pour chaque unité, au moins une caractéristique a été mesurée, appelée variable. Les variables peuvent être réparties en fonction de leur nature (qualitative ou quantitative) ou du rôle qu'elles joueront dans l'analyse (variable dépendante ou indépendante). Lorsque les données de l'échantillon sont récoltées, elles peuvent être analysées avec deux sous-domaines de la statistique : la statistique descriptive, pour décrire l'ensemble de données, et l'inférence statistique, pour essayer de tirer sur base de l'échantillon des conclusions à propos de la population. L'échelle de mesure (nominale, ordinale, d'intervalles ou de rapports) des variables qui ont été mesurées est à prendre en compte pour ces analyses.



Résumé audio



www.lienmini.fr/05526-audio1



Exercices

1. Dans quelles conditions l'ensemble des étudiants de votre université serait-il considéré comme une population ?
2. Dans quelles conditions l'ensemble des étudiants de votre université serait-il considéré comme un échantillon ?
3. Si l'ensemble des étudiants de votre université était considéré comme un échantillon, comme dans l'exercice 2, s'agirait-il d'un échantillon aléatoire ou non ? Pourquoi ?
4. Pourquoi en 1950 le choix de noms à partir de l'annuaire téléphonique d'une ville ne produirait-il pas un échantillon aléatoire des résidents de cette ville ? Qui serait sous-représenté et qui serait sur-représenté ?
5. Donnez deux exemples de variables indépendantes et deux exemples de variables dépendantes.
6. Écrivez une phrase décrivant une expérience avec une variable indépendante et une variable dépendante.
7. Donnez trois exemples de variables continues.

8. Donnez trois exemples de variables discrètes.
9. Donnez un exemple d'étude dans laquelle on s'intéresse à l'estimation du score moyen (tendance centrale) d'une population.
10. Donnez un exemple d'étude qui ne s'intéresse pas à la valeur numérique réelle de la population, mais qui cherche à savoir si la moyenne d'une population est supérieure à la moyenne d'une autre population.

Testez-vous en ligne !



Plus d'exercices



www.lienmini.fr/05526-ex1



Défis



www.lienmini.fr/05526-defis1



Flash-cards



www.lienmini.fr/05526-cards1

Description et exploration des données

MOTS-CLÉS

Analyse exploratoire des données, asymétrie, boîte à moustaches ou *boxplot*, centrage, coefficient de variation (CV), décile, degrés de liberté (*dl*), diagramme en tiges et feuilles, diagramme de densité estimée par noyau, distribution bimodale, distribution de fréquences, distribution symétrique, distribution unimodale, écart absolu moyen, écart interquartile, écart-type (s ou σ), échantillon Winsorisé, échantillon tronqué, estimateurs non biaisés, étendue, médiane, mesures de position, mesures de tendance centrale, mode, moyenne, position médiane, percentile, quantile, quartile, reflet, score standardisé, standardisation, transformations linéaires, valeur aberrante, valeur attendue, valeur extrême, variance de l'échantillon, variance de la population, voussure

SOMMAIRE

1. Représentation graphique de données	19
2. Histogrammes	21
3. Ajustement des courbes aux données	23
4. Diagrammes en tiges et feuilles	27
5. Description de distributions	30
6. Notation	34
7. Mesures de tendance centrale	36
8. Mesures de variabilité	40
9. Diagrammes en boîte : représentations graphiques de dispersions et scores extrêmes	53
10. Obtention de mesures de dispersion en utilisant SPSS	58
11. Percentiles, quartiles et déciles	58
12. L'effet de transformations linéaires sur les données	60

Si les comparaisons s'effectuent en parallèle, le nombre de chiffres inclus dans l'ensemble de comparaison ne devrait pas avoir d'influence).

Même si l'objectif de Sternberg était de comparer les données dans les différentes

conditions, nous pouvons directement nous faire une idée de celles-ci en considérant l'ensemble complet des temps de réaction, quelle que soit la condition du stimulus. Les données du tableau 2.1 ont été recueillies au cours d'une expérience

analogue à celle de Sternberg, mais ne comprenant qu'un participant: moi. Aucune correction des réponses n'était autorisée de sorte que les données présentées ici proviennent uniquement d'essais corrects.

Stimuli de comparaison*	Temps de réaction, en centièmes de seconde																
1O	40	41	47	38	40	37	38	47	45	61	54	67	49	43	52	39	46
	47	45	43	39	49	50	44	53	46	64	51	40	41	44	48	50	42
	90	51	55	60	47	45	41	42	72	36	43	94	45	51	46	52	
1N	52	45	74	56	53	59	43	46	51	40	48	47	57	54	44	56	47
	62	44	53	48	50	58	52	57	66	49	59	56	71	76	54	71	104
	44	67	45	79	46	57	58	47	73	67	46	57	52	61	72	104	
3O	73	83	55	59	51	65	61	64	63	86	42	65	62	62	51	62	72
	55	58	46	67	56	52	46	62	51	51	61	60	75	53	59	56	50
	43	58	67	52	56	80	53	72	62	59	47	62	53	52	46	60	
3N	73	47	63	63	56	66	72	58	60	69	74	51	49	69	51	60	52
	72	58	74	59	63	60	66	59	61	50	67	63	61	80	63	60	64
	64	57	59	58	59	60	62	63	67	78	61	52	51	56	95	54	
5O	39	65	53	46	78	60	71	58	87	77	62	94	81	46	49	62	55
	59	88	56	77	67	79	54	83	75	67	60	65	62	62	62	60	58
	67	48	51	67	98	64	57	67	55	55	66	60	57	54	78	69	
5N	66	53	61	74	76	69	82	56	66	63	69	76	71	65	67	67	55
	65	58	64	65	81	69	69	63	68	70	80	68	63	74	61	85	125
	59	61	74	76	62	83	58	72	65	61	95	58	64	66	66	72	

* 1, 3 et 5 représentent le nombre de chiffres constituant l'ensemble de comparaison, et les lettres O et N indiquent si le chiffre était, oui (O) ou non (N), inclus dans la comparaison.

TABLEAU 2.1.

Données relatives aux temps de réaction enregistrés lors de l'expérience sur l'identification de chiffres

1. Représentation graphique de données

Comme vous pouvez le voir, le tableau 2.1 comprend tout simplement trop de nombres pour que nous puissions les interpréter en un coup d'œil. L'une des méthodes les plus simples de réorganiser les données pour les rendre plus intelligibles est d'en faire un graphique. Il y a plusieurs manières courantes de représenter graphiquement les données. Nous évoquerons tour à tour les distributions de fréquences, les histogrammes et les diagrammes en tiges et feuilles. Je crois fermement qu'il faut simplifier le plus possible les graphiques afin de ne pas rendre le message confus par des éléments inutiles. Cependant, si vous voulez voir un exemple remarquable de la façon dont la représentation graphique des données peut révéler des informations importantes que vous ne verriez pas autrement, la vidéo ci-contre est très impressionnante.



TED Talk :
Hans Rosling



www.lienmini.fr/05526-rosling

1.1. Distributions de fréquences

Distribution de fréquences

En guise de première étape, nous pouvons élaborer une **distribution de fréquences** afin d’organiser les données en un certain ordre logique. Pour notre exemple, nous pourrions compter combien de fois chaque temps de réaction possible a été enregistré. Par exemple, le participant a répondu cinq fois en 50 centièmes de seconde et douze fois en 51 centièmes de seconde. Il lui est arrivé une fois de s’embrouiller et de répondre en 1.25 seconde (125 centièmes de seconde). La distribution de fréquences de ces données est présentée dans le tableau 2.2, qui indique combien de fois chaque temps de réaction a été enregistré.

La distribution illustrée dans le tableau 2.2 et le graphique présenté à la figure 2.1 mettent en évidence une distribution très large des temps de réaction, qui vont de 36 centièmes de seconde à 125 centièmes de seconde. Les données ont tendance à se grouper autour de 60 centièmes, la plupart des scores étant compris entre 40 et 90 centièmes de seconde. Cette tendance n’apparaissait pas dans le tableau 2.1, qui ne comprenait que des données non organisées.

TABLEAU 2.2. Distribution de fréquences des temps de réaction

Temps de réaction en centièmes de seconde		Temps de réaction en centièmes de seconde	
	Fréquence		Fréquence
36	1	71	4
37	1	72	8
38	2	73	3
39	3	74	6
40	4	75	2
41	3	76	4
42	3	77	2
43	5	78	3
44	5	79	2
45	6	80	3
46	11	81	2
47	9	82	1
48	4	83	3
49	5	84	0
50	5	85	1
51	12	86	1
52	10	87	1
53	8	88	1
54	6	89	0
55	7	90	1
56	10	91	0
57	7	92	0
58	12	93	0
59	11	94	2
60	12	95	2
61	11	96	0
62	14	97	0
63	10	98	1
64	7	99	0
65	8	...	...
66	8	...	...
67	14	104	2
68	2	...	...
69	7	125	1
70	1		

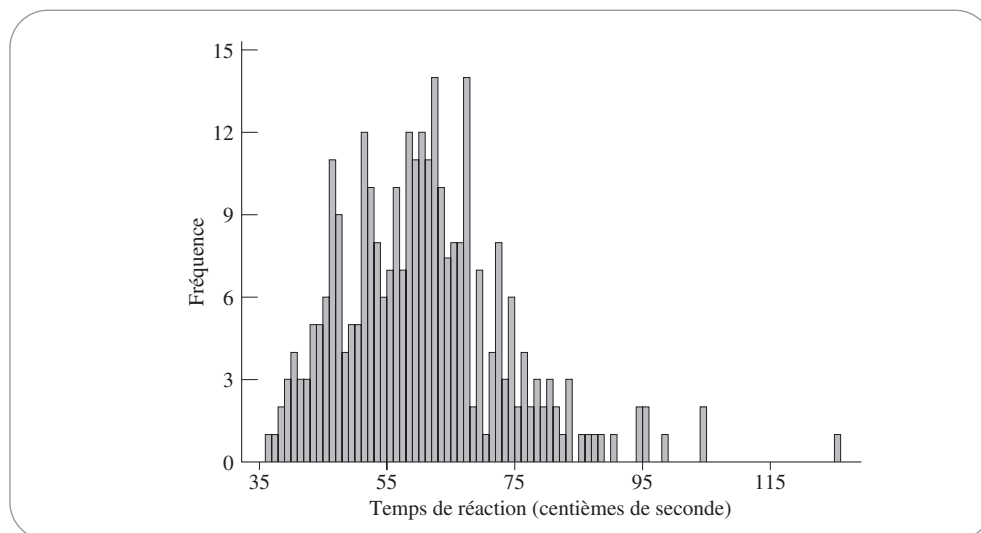


FIGURE 2.1. Graphique des temps de réaction en fonction de la fréquence

2. Histogrammes

D'après la distribution donnée dans le tableau 2.1, nous pouvons facilement représenter graphiquement les données, tel que dans la figure 2.1. Toutefois, lorsque, comme ici, nous sommes confrontés à une variable présentant de nombreuses valeurs différentes, chaque valeur est généralement peu fréquente, et nous notons souvent une fluctuation importante des fréquences des valeurs adjacentes. Remarquez par exemple qu'il y a quatorze 67, mais seulement deux 68. Dans des cas comme ceux-ci, il est plus judicieux de regrouper les valeurs adjacentes en un **histogramme**.¹ Notre objectif est de masquer le « désordre » aléatoire tout en préservant les tendances importantes qui caractérisent les données. Nous pourrions ainsi grouper les données en blocs de 5 centièmes de seconde en combinant les fréquences de tous les résultats situés entre 35 et 39, entre 40 et 44, etc. Le tableau 2.3 donne un exemple de cette distribution.

Histogramme

Le tableau 2.3 donne les limites supérieure et inférieure des intervalles comme étant des nombres entiers, pour la simple raison que cela améliore la lisibilité du tableau. Il faut toutefois se rendre compte que les limites réelles de l'intervalle (désignées sous le nom de limite inférieure réelle et de limite supérieure réelle) sont des valeurs décimales qui se situent entre le haut d'un intervalle et le bas du suivant. La **limite inférieure réelle** d'un intervalle est la plus petite valeur qui serait classée comme faisant partie de l'intervalle. De même, la **limite supérieure réelle** est la plus grande valeur qui serait classée comme faisant partie de l'intervalle. Par exemple, si nous avons enregistré des temps de réaction au millième de seconde près, au lieu du centième de seconde près, l'intervalle 35-39 inclurait toutes les valeurs situées

Limite inférieure réelle
Limite supérieure réelle

1. Différentes personnes semblent évoquer des choses différentes lorsqu'elles parlent d'un « histogramme. » Certaines utilisent ce terme pour les distributions de données, que des catégories aient ou non été combinées (elles diraient que la figure 2.1 est un histogramme), et d'autres réservent ce terme pour les cas où des catégories adjacentes ont été combinées. Vous avez sans doute déjà remarqué que je ne suis pas un maniaque de ce genre de distinctions et que je compte utiliser « histogramme » et « distribution de fréquence » de façon plus ou moins interchangeable. J'espère que personne ne trouvera cela confus.

entre 34.5 et 39.5, parce que les valeurs comprises entre ces points seraient arrondies dans cet intervalle (Les gens se demandent souvent avec inquiétude ce qu'ils feraient si une personne obtenait un score exactement égal à 39.50000000, ce qui la situerait juste au point de rupture entre deux intervalles. Il n'y a pas de quoi s'inquiéter. D'abord, cela n'arrive pas très souvent. Ensuite, il est toujours possible de jouer à pile ou face. Enfin, il y a beaucoup de choses plus importantes qui doivent nous préoccuper. Il suffit que vous optiez pour une manière de faire dans ce genre de situations et que vous vous y teniez. Il s'agit ici de l'un des « faux problèmes » qui incitent les gens à trouver l'étude des statistiques déroutante, ennuyeuse ou encore les deux.

Points milieu

Les **points milieu** énumérés dans le tableau 2.3 sont les moyennes des limites supérieure et inférieure et sont présentés par souci de commodité. Dans une représentation graphique des données, nous traçons souvent les points comme s'ils se situaient chacun au milieu de son intervalle.

Le tableau 2.3 énumère également les fréquences liées à chaque intervalle. Par exemple, sept temps de réaction se situaient entre 35 et 39 centièmes de seconde. La figure 2.2 présente la distribution du tableau 2.3 sous forme d'histogramme.

EXEMPLE

NOMBRE IDÉAL D'INTERVALLES DANS UN HISTOGRAMME



Les gens se demandent souvent quel est le nombre idéal d'intervalles à utiliser lorsqu'on groupe les données. Bien qu'il n'y ait pas de réponse correcte à la question des intervalles à utiliser, on opte généralement pour 10-12 intervalles. Keen (2010) donne au moins six règles possibles pour déterminer le nombre optimal d'intervalles, mais ces règles sont principalement destinées à ceux qui écrivent des logiciels et qui ont besoin de règles pour gérer le cas général. Si vous avez affaire à un ensemble particulier de données, vous pouvez tout à fait ignorer les règles formelles et modifier les paramètres jusqu'à ce que le résultat semble cohérent.²

Intervalle	Point milieu	Fréquence	Fréquence cumulée	Intervalle	Point milieu	Fréquence	Fréquence cumulée
35–39	37	7	7	85–89	87	4	291
40–44	42	20	27	90–94	92	3	294
45–49	47	35	62	95–99	97	3	297
50–54	52	41	103	100–104	102	2	299
55–59	57	47	150	105–109	107	0	299
60–64	62	54	204	110–114	112	0	299
65–69	67	39	243	115–119	117	0	299
70–74	72	22	265	120–124	122	0	299
75–79	77	13	278	125–129	127	1	300
80–84	82	9	287				

TABLEAU 2.3. Distribution de fréquences groupées

2. Pour choisir un nombre idéal d'intervalles, il peut être intéressant de prendre le nombre entier le plus proche de $\sqrt{N}$, où N est le nombre d'observations. Si l'on applique cette méthode ici, on obtient $\sqrt{N} = \sqrt{300} = 17.32 = 17$ intervalles, soit un nombre assez proche de celui que nous avons choisi (19).

Dans cet exemple, j'ai utilisé 19 intervalles parce que les nombres se répartissaient naturellement de cette manière, et parce que les observations étaient nombreuses. En général, et dans la mesure du possible, il vaut mieux utiliser les divisions naturelles du système de nombres (ex. 0-9, 10-19,... ou 100-119, 120-139) plutôt que de diviser l'étendue en exactement 10 intervalles définis de manière arbitraire. Toutefois, il faut adopter un autre type de limites si celles-ci permettent de faciliter l'interprétation des données. Ne l'oubliez pas, l'objectif est de faire en sorte que les données soient utiles, et non de suivre un ensemble de règles rigides élaborées par quelqu'un qui n'a jamais vu le problème en question.

Notez, à la figure 2.2, que les données relatives aux temps de réaction sont généralement centrées sur 50-70 centièmes de seconde, que la distribution monte et descend de façon régulière et qu'elle s'estompe vers la droite. On ne s'étonnera pas de voir ces temps s'estomper vers la droite (on parle d'asymétrie positive, ou d'asymétrie à droite) puisque la *rapidité* de réaction de la personne est limitée, alors que la *lenteur* possible de sa réaction ne l'est pas. Remarquez également la valeur extrême que constitue la valeur 125 centièmes. Cette valeur est appelée **valeur extrême** parce qu'elle est largement séparée du reste des données. Il arrive souvent que les valeurs extrêmes représentent des erreurs d'enregistrement des données, on parle alors de **valeurs aberrantes**. Mais dans ce cas-ci, il s'agissait juste d'un essai au cours duquel le participant hésitait entre les deux boutons.

Valeur extrême

Valeur aberrante

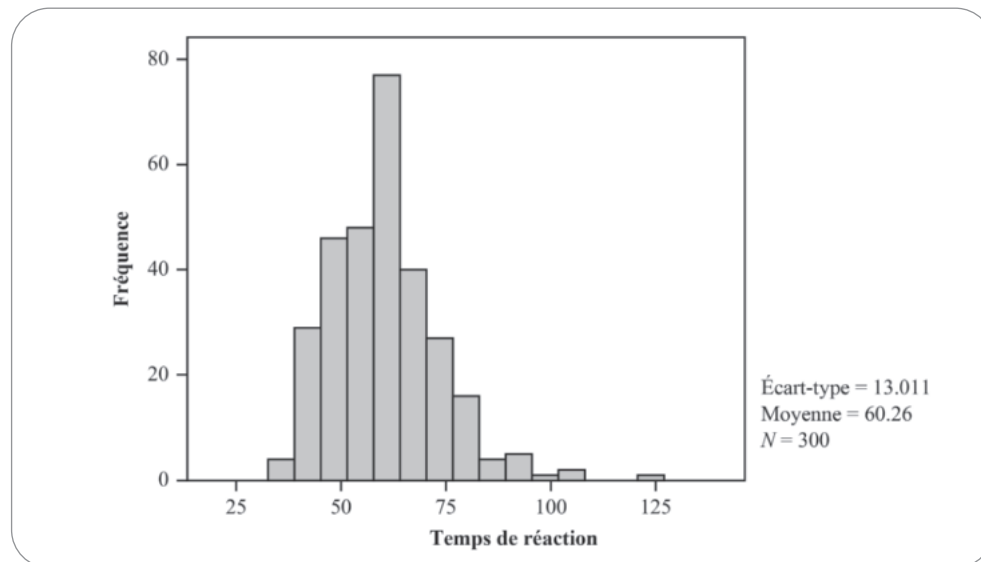


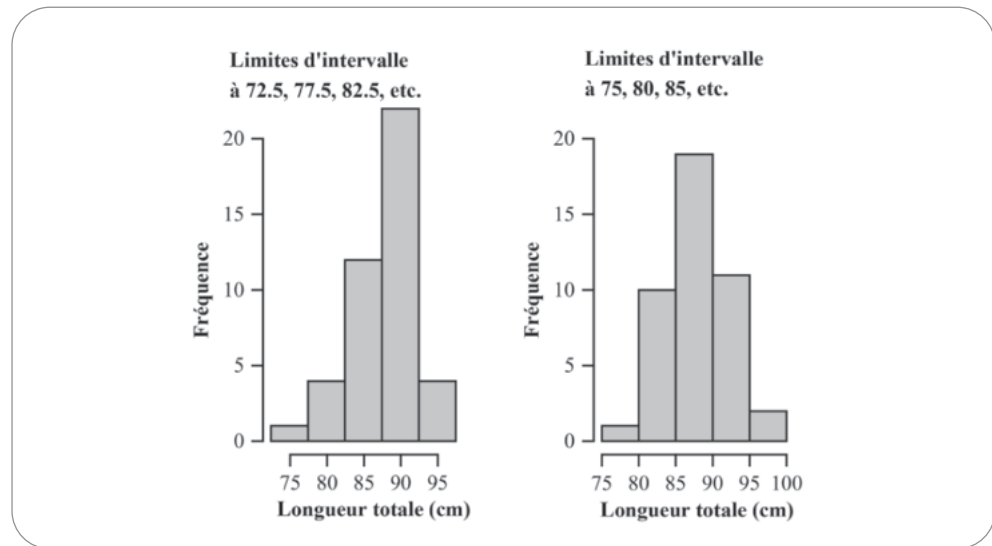
FIGURE 2.2. Distribution groupée des temps de réaction

3. Ajustement des courbes aux données

Les histogrammes tels que ceux des figures 2.1 et 2.2 peuvent souvent être utilisés pour afficher des données d'une manière qui fait sens, mais ils présentent leurs propres problèmes. Un certain nombre de personnes ont fait remarquer que les histogrammes, aussi courants soient-ils, ne permettent pas toujours de décrire clairement les données. Cela est particulièrement vrai pour les échantillons de petite taille, où des changements mineurs dans l'emplacement ou la largeur de l'intervalle peuvent marquer une différence notable dans la forme de la distribution. Wilkinson

(1994) a rédigé un excellent article sur ce sujet et d'autres problèmes connexes. Maindonald et Braun (2007) donnent l'exemple illustré à la figure 2.3, qui représente les longueurs des opossums. Le premier regroupe les données dans des cases avec des limites d'intervalles à 72.5, 77.5, 82.5... Le second utilise des limites d'intervalles à 70, 75, 80... Les personnes qui s'intéressent aux opossums, et il y en a, pourraient tirer des conclusions très différentes de ces deux graphiques en fonction des limites d'intervalles utilisées. Les données sont assez symétriques dans l'histogramme de droite, mais présentent une queue notable vers la gauche dans l'histogramme de gauche.

FIGURE 2.3. Deux histogrammes différents représentant les mêmes données sur la longueur des opossums



La figure 2.2 est en fait une représentation assez juste des temps de réaction, mais nous pouvons souvent l'améliorer en ajustant une courbe lissée aux données, seule ou en complément de l'histogramme. Je vais aborder deux des nombreuses approches de l'ajustement de courbes, dont l'une superpose une distribution normale (qui sera abordée plus en détail dans le chapitre suivant) et l'autre utilise ce que l'on appelle un « diagramme de densité estimée par noyau ».

3.1. Ajuster une courbe normale

Même si vous n'avez pas encore lu le chapitre 3, vous devriez, en principe, être familier avec une courbe normale. Elle est souvent appelée « courbe en cloche » et est symétrique autour du centre de la distribution, s'amenuisant aux deux extrémités. La distribution normale a une définition spécifique, mais nous y reviendrons au chapitre suivant. Pour l'instant, nous nous limiterons à dire que nous supposons souvent que nos données sont normalement distribuées, et la superposition d'une distribution normale sur l'histogramme nous donnera une idée du caractère raisonnable de cette hypothèse.³

La figure 2.4 a été produite par SPSS et vous pouvez voir que, bien que les données soient grossièrement décrites par la distribution normale, la distribution réelle est quelque peu tronquée sur la gauche et présente un nombre d'observations supérieur

3. Ce n'est pas la meilleure manière d'évaluer si la distribution est normale ou non, comme nous le verrons dans le chapitre suivant. Cependant, c'est une manière courante de procéder.

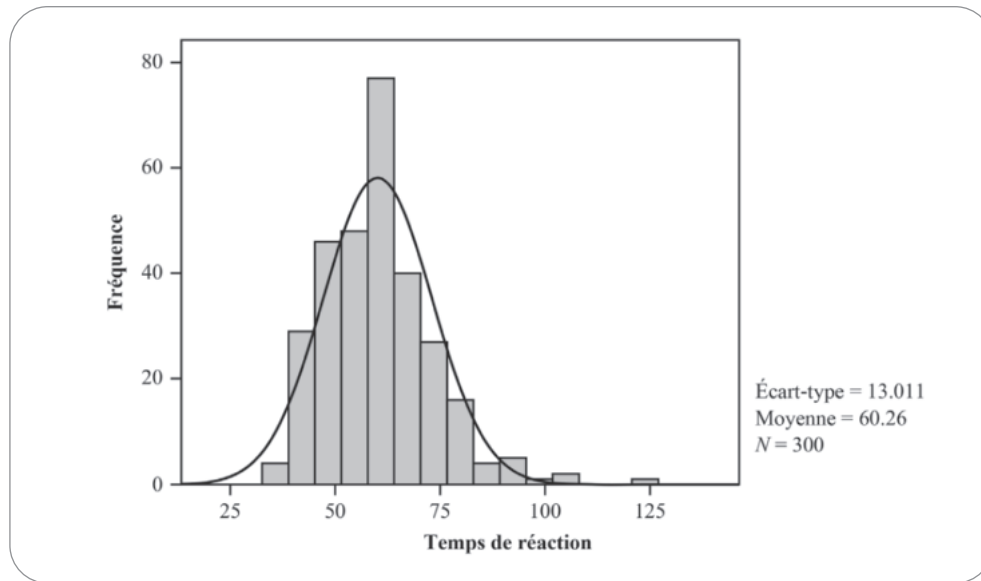


FIGURE 2.4. Histogramme des données de temps de réaction avec une courbe normale superposée

au nombre attendu sur l'extrême droite. La courbe normale n'est pas un mauvais ajustement, mais nous pouvons faire mieux. Une approche alternative consisterait à créer ce que l'on appelle un « graphique de densité estimée par noyau ».

3.2. Graphiques de densité estimée par de noyau

Dans la figure 2.4, nous avons superposé une distribution théorique aux données. Cette distribution n'a utilisé que quelques caractéristiques des données, à savoir leur moyenne et leur écart-type, et n'a fait aucun effort pour ajuster la courbe à la forme réelle de la distribution. En fait, nous pouvons superposer la distribution normale en calculant uniquement la moyenne et l'écart-type (que nous aborderons plus tard dans ce chapitre) des données. Les données individuelles et leur distribution ne jouent aucun rôle dans le tracé de cette distribution. Les **diagrammes de densité estimée par noyau** font pratiquement le contraire. Ils essaient en fait d'ajuster une courbe lissée aux données, tout en tenant compte du fait qu'il y a beaucoup de bruit dans les observations qui ne devrait pas trop déformer la courbe. Les diagrammes de densité estimée par noyau ne tiennent pas compte de la moyenne et de l'écart-type des observations.

Diagramme de densité estimée par noyau

L'idée derrière un graphique de densité estimée par noyau est que chaque observation aurait pu être légèrement différente. Par exemple, lors d'un essai où le temps de réaction de la personne interrogée était de 80 centièmes de seconde, le score aurait pu raisonnablement être de 79 ou 82. Il est même concevable que le score ait pu être de 73 ou 86, mais il n'est pas du tout probable que le score ait été de 20 ou 100. En d'autres termes, il existe une distribution de possibilités alternatives autour de toute valeur obtenue, et ceci est vrai pour toutes les valeurs obtenues. Nous allons utiliser ce fait pour produire une courbe globale qui s'adapte généralement assez bien aux données.

Les estimations par noyau peuvent être illustrées graphiquement en prenant un exemple tiré d'Everitt & Hothorn (2006). Les auteurs ont utilisé un ensemble très simple de données avec les valeurs suivantes pour la variable dépendante (X).

X	0.0	1.0	1.1	1.5	1.9	2.8	2.9	3.5
-----	-----	-----	-----	-----	-----	-----	-----	-----

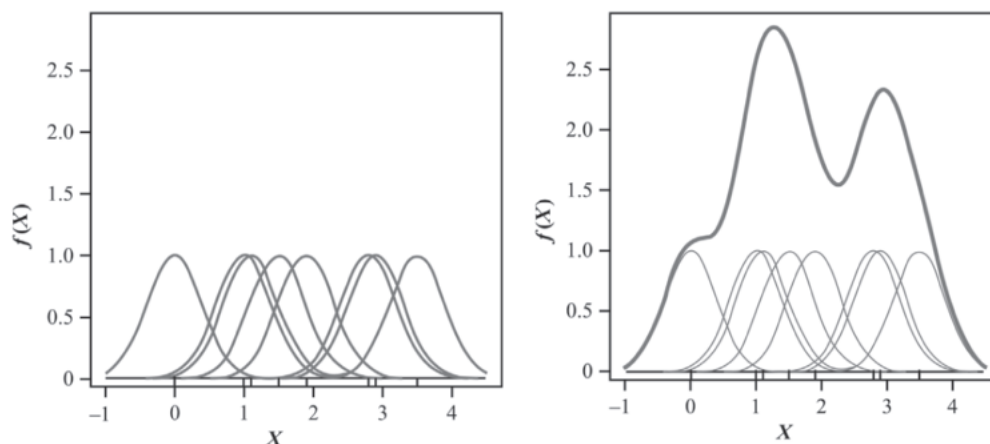
Si vous tracez ces points le long de l'axe X et que vous superposez de petites distributions représentant les valeurs alternatives qui auraient pu être obtenues à la place des valeurs réelles dont vous disposez, vous obtenez la distribution illustrée à gauche sur la figure 2.5. Everitt et Hothorn donnent un nom technique à ces petites distributions : les « bosses ». Remarquez que ces bosses sont des distributions normales, mais j'aurais pu spécifier une autre forme si je pensais qu'une distribution normale était inappropriée.

Nous allons à présent littéralement additionner ces bosses verticalement. Par exemple, supposons que nous nommions chaque bosse par le score sur lequel elle est centrée. Au-delà d'une valeur de 3.8 sur l'axe des X , vous avez une petite quantité de bosse_2.8, un peu plus de bosse_2.9 et une bonne partie de bosse_3.5. Vous pouvez additionner les hauteurs de ces trois bosses à $X = 3.8$ afin d'obtenir la densité estimée par noyau de la courbe globale à cette position. Vous pouvez faire de même pour toutes les autres valeurs de X . Si vous procédez ainsi, vous obtenez la distribution représentée à droite dans la figure 2.5. Au-dessus des bosses, nous avons une distribution ondulée (pour utiliser un autre terme technique) qui représente notre meilleure estimation de la distribution sous-jacente aux données desquelles nous sommes partis.

Nous pouvons maintenant revenir aux données de temps de réaction et superposer la fonction de densité estimée par noyau à cet histogramme. (Je laisse de côté les bosses car elles sont trop nombreuses pour être lisibles.) Le graphique qui en résulte est illustré à la figure 2.6. Remarquez que cette courbe représente beaucoup mieux les données que la distribution normale superposée. Plus particulièrement, elle s'adapte très bien aux queues de la distribution.

SPSS ajuste les diagrammes de densité estimée par noyau à l'aide de la syntaxe, et vous pouvez les ajuster à l'aide de SAS et S-Plus (ou son cousin proche, R). Il est assez facile de trouver des exemples de ces programmes sur Internet. Au fur et à mesure que la psychologie s'étend à d'autres domaines, et en particulier aux neurosciences et aux sciences de la santé, des techniques comme les diagrammes de densité estimée par noyau deviennent de plus en plus courantes. Ces tracés comportent un certain nombre d'aspects techniques, par exemple la forme des bosses et la largeur de bande utilisée pour les créer, mais vous disposez maintenant des informations de

FIGURE 2.5. Illustration de la fonction de densité estimée par noyau pour X



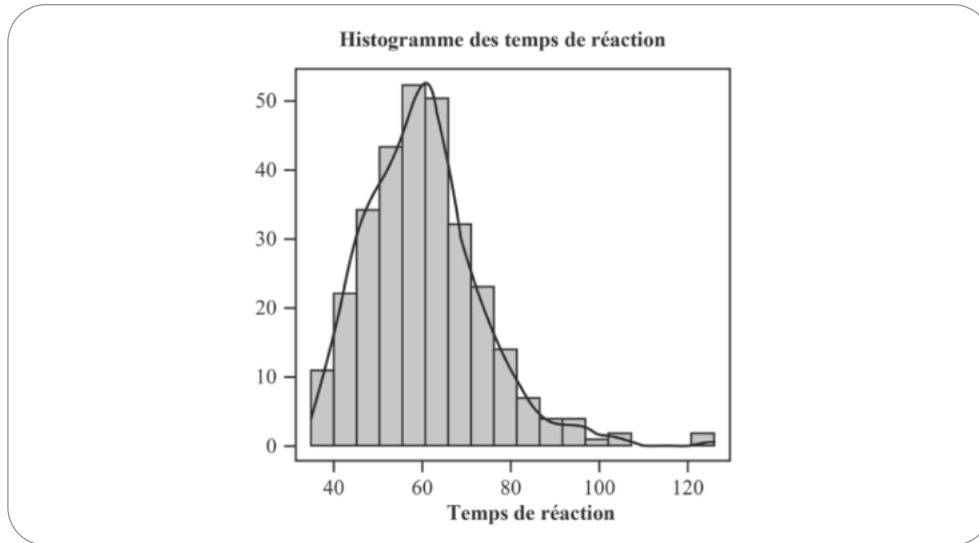


FIGURE 2.6. Graphique de densité estimée par noyau pour des données sur le temps de réaction

base qui vous permettront de comprendre et de travailler avec ces tracés. Une page Web décrivant comment calculer des diagrammes de densité estimée par noyau dans SPSS et dans SAS est disponible ci-contre ; un programme dans R est également disponible sur le site Web du livre à l'adresse ci-contre.

4. Diagrammes en tiges et feuilles

Bien que les histogrammes, les distributions de fréquences et les fonctions de densité estimées par noyau soient des méthodes couramment utilisées pour présenter les données, elles ne sont pas pour autant dépourvues d'inconvénients. Comme les histogrammes représentent souvent des observations qui ont été regroupées en intervalles, ils perdent régulièrement les valeurs numériques réelles des scores individuels dans chaque intervalle. Quant aux distributions de fréquences, elles conservent les valeurs des différentes observations, mais peuvent s'avérer difficiles à utiliser lorsqu'elles ne résument pas suffisamment les données. Il existe une méthode alternative permettant d'éviter ces deux inconvénients : le **diagramme en tiges et feuilles**.

John Tukey (1977) a développé, dans le cadre de son approche générale de l'analyse des données, connue sous le nom d'**analyse exploratoire des données**, différentes méthodes permettant de présenter les données sous une forme visuellement utile. L'une des méthodes les plus simples est la représentation en tiges et feuilles. C'est une méthode employée par la plupart des logiciels statistiques disponibles. Il est impossible de l'évoquer ici au moyen des données relatives aux temps de réaction, parce que cet exemple nécessiterait une représentation légèrement plus complexe vu le grand nombre d'observations. Nous utiliserons plutôt un ensemble hypothétique de données à propos du temps (en minutes par semaine) consacré aux jeux électroniques par 100 étudiants pris séparément. La figure 2.7 montre certaines données brutes. Le côté gauche de la figure présente une partie des données (les données d'étudiants jouant entre 40 et 80 minutes par semaine), et le côté droit donne la représentation complète en tiges et feuilles qui en découle.

 Diagrammes de densité dans SPSS et SAS



www.lienmini.fr/05526-densite

 Diagrammes de densité dans R



www.lienmini.fr/05526-plots

Diagramme en tiges et feuilles

Analyse exploratoire des données

Les données brutes présentées à la figure 2.7 montrent qu'il y a plusieurs scores dans les 40, une autre série dans les 50, deux scores dans les 60 et quelques-uns dans les 70. Nous qualifierons les chiffres des dizaines – ici, 4, 5, 6 et 7 – de chiffres principaux pour ces scores. Ces chiffres principaux forment la tige, ou axe vertical, de notre représentation. Parmi les 14 scores situés dans les 40, on note un 40, deux 41, un 42, deux 43, un 44, aucun 45, trois 46, un 47, un 48 et deux 49. Les chiffres des unités 1, 2, 3, etc. sont appelés chiffres secondaires. Ils forment les **feuilles** – les éléments horizontaux – de notre représentation.⁴

FIGURE 2.7. Représentation en tiges et feuilles : données relatives aux jeux électroniques

Données brutes	Tiges	Feuilles
.	0	00000000000233566678
.	1	2223555579
.	2	33577
40 41 41 42 43	3	22278999
43 44 46 46 46	4	01123346667899
47 48 49 49	5	24557899
52 54 55 55 57	6	37
58 59 59	7	1556689
63 67	8	34779
71 75 75 76 76	9	466
78 79	10	23677
.	11	3479
.	12	2557899
.	13	89

Le côté droit de la figure 2.7 montre qu'à côté de l'entrée 4 de la tige, il y a un 0, deux 1, un 2, deux 3, un 4, trois 6, un 7, un 8 et deux 9. Ces valeurs de feuilles correspondent aux chiffres des unités des données brutes. De même, remarquez que les feuilles se trouvant en face de la valeur 5 de la tige correspondent aux chiffres des unités de toutes les réponses situées dans les 50. La représentation en tiges et feuilles permet de réorganiser complètement les données brutes utilisées pour cette représentation. On peut par exemple dire que 11 étudiants ont consacré 0 minute aux jeux électroniques, un étudiant y a consacré deux minutes, deux étudiants y ont consacré trois minutes, etc. En outre, la forme de la représentation ressemble à un histogramme vu de côté ; il procure donc aussi tous les avantages de cette méthode de représentation graphique des données.

Cette représentation simple en tiges et feuilles présente apparemment l'inconvénient d'entraîner, pour certaines données, un groupement trop imprécis pour nos objectifs. En fait, c'est la raison pour laquelle il a fallu avoir recours à des données hypothétiques pour cet exemple introductif. En essayant d'utiliser les données relatives aux temps de réaction, je me suis rendu compte que la tige de 50 (c.-à-d. 5) se trouvait face à 88 feuilles, ce qui est un peu ridicule. Ce n'est pas grave : Tukey a imaginé un moyen habile de contourner ce problème.

Si le problème vient des nombreuses valeurs situées entre 50 et 59, nous devrions peut-être diviser cet intervalle en intervalles plus petits. Nous *pourrions* essayer d'utiliser les intervalles 50-54, 55-59, etc. Mais dans ce cas, nous ne pourrions pas

4. Il n'est pas toujours vrai que les chiffres des dizaines forment la tige et les chiffres des unités les feuilles. Par exemple, si les données allaient de 100 à 1000, les chiffres des centaines formeraient la tige, les chiffres des dizaines les feuilles, et nous ne considérerions pas les chiffres des unités.

Données brutes	Tiges	Feuilles
36 37 38 38 39 39 39 40	3s	67
40 40 40 41 41 41 42 42	3.	88999
42 43 43 43 43 44 44	4*	0000111
44 44 44 45 45 45 45 45	4t	22233333
45 46 46 46 46 46 46 46	4f	44444555555
46 46 46 46 47 47 47 47	4s	6666666666677777777
47 47 47 47 47 48 48 48	4.	888899999
48 49 49 49 49 49 50 50	5*	0000011111111111111
50 50 50 51 51 51 51 51	5t	2222222222333333333
51 51 51 51 51 51 51 52	5f	444444555555555
52 52 52 52 52 52 52 52	5s	6666666666677777777
52 53 53 53 53 53 53 53	5.	888888888889999999999
53 54 54 54 54 54 54 55	6*	000000000001111111111
55 55 55 55 55 55	6t	22222222222222333333333
...	6f	44444445555555555
	6s	66666666677777777777777
	6.	889999999
	7*	01111
	7t	222222222333
	7f	44444455
	7s	666677
	7.	88899
	8*	00011
	8t	2333
	8f	5
	8s	67
	8.	8
	9*	0
	9t	
	9f	4455
	9s	
	9.	8
	High	104; 104; 125

FIGURE 2.8. Représentation en tiges et feuilles : données relatives aux temps de réaction

choisir simplement 5 comme tige, parce que cela ne permettrait pas de distinguer les deux intervalles. Tukey a suggéré d'utiliser « 5* » pour représenter 50-54 et « 5. » pour représenter 55-59. Mais notre problème n'en est pas pour autant résolu, parce que les catégories restent trop imprécises. Tukey a donc suggéré un schéma alternatif, où « 5* » représente 50-51, « 5t » représente 52-53, « 5f » représente 54-55, « 5s » représente 56-57 et « 5. » représente 58-59 (Les lettres sont choisies en fonction de l'initiale du chiffre en anglais : t pour *two* et *three*, f pour *four* et *five*, s pour *six* et *seven*). Si nous appliquons ce schéma aux données sur les temps de réaction, cela donne les résultats illustrés à la figure 2.8. Le choix du nombre de tiges à utiliser pose un problème analogue à la sélection du nombre de catégories d'un histogramme. De nouveau, il s'agit d'opter pour une solution intelligente, qui soit susceptible de donner des informations utiles. Une seule restriction : les tiges doivent avoir la même largeur. L'une ne peut représenter 50-54 si l'autre représente 60-69.

Notez que la figure 2.8 ne représente pas les valeurs extrêmes comme les autres. J'ai préféré utiliser le mot *high* (élevé) à la place de la tige puis insérer les valeurs réelles. Cela avait pour but de mettre en évidence la présence de valeurs extrêmes, ainsi que de gagner de la place.

Les représentations en tiges et feuilles peuvent s'avérer particulièrement utiles pour comparer deux distributions différentes. Il est possible de procéder à cette

comparaison en représentant graphiquement les deux distributions sur les côtés opposés de la tige. La figure 2.9 présente la distribution des notes obtenues par des hommes et des femmes dans un cours que j’ai enseigné sur les méthodes expérimentales et qui incluait un important volet statistique. Il s’agit de données réelles. Notez le recours à des tiges telles que 6* (pour 60-64) et 6. (pour 65-69). Notez en outre le code en bas du tableau, qui indique la façon dont les entrées se traduisent en scores bruts. Ce code dit que |4*|,1 représente 41 et non 4.1 ou 410. Remarquez pour terminer que la figure illustre la différence de performance entre étudiants et étudiantes.

FIGURE 2.9. Résultats réels (pourcentage) obtenus à un cours de méthodes expérimentales, représentés séparément par genre

Hommes	Tiges	Femmes
6	3*	1
	3.	
	4*	
	4.	
	5*	
2	5.	03
	6*	
	6.	
32200	7*	
88888766666655	7.	
4432221000	8*	555556666788899
766666555	8.	000001111222234444
422	9*	55666666666667788888899
	9.	000000000133
		56
Code 4* 1 = 41		

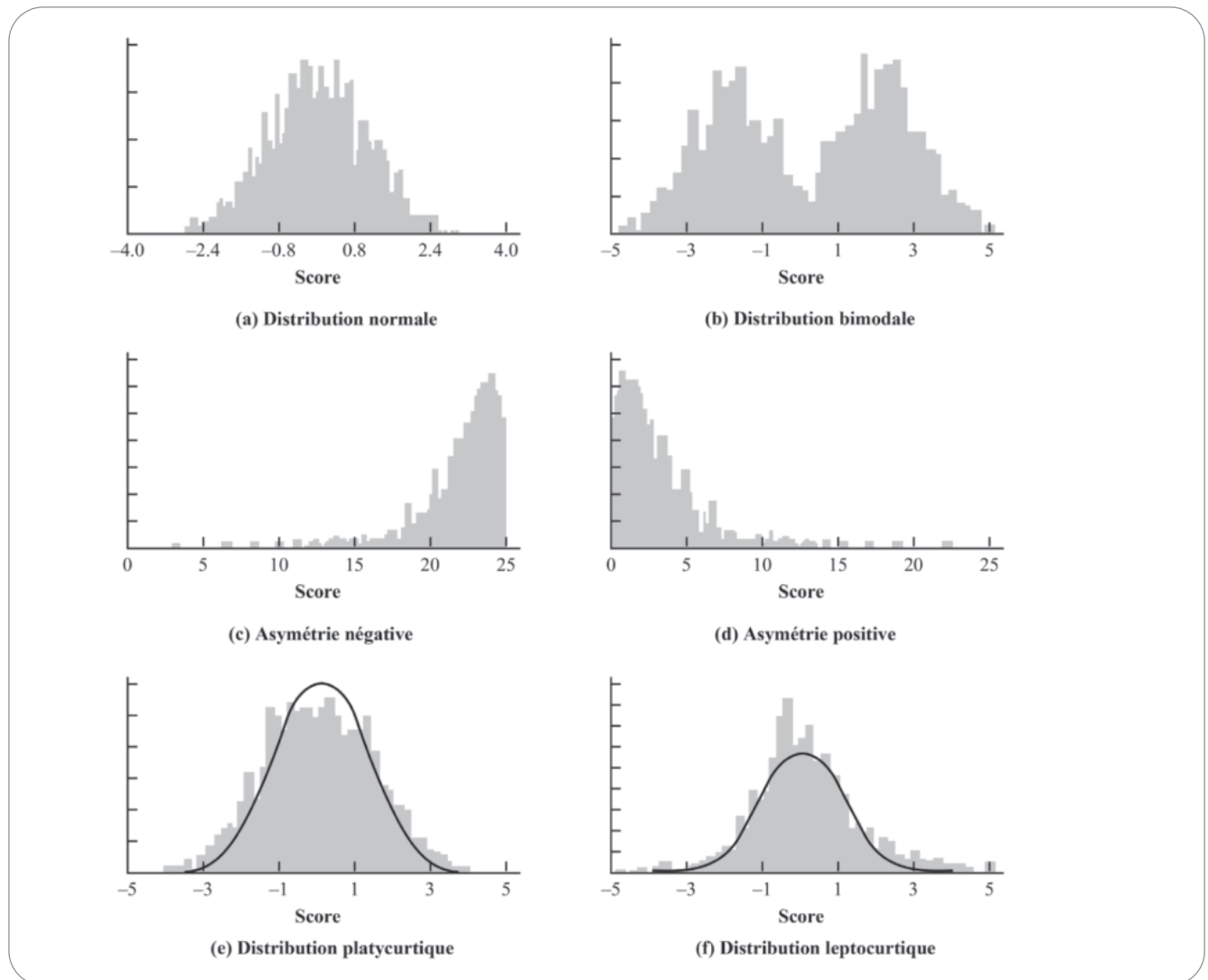
5. Description de distributions

Les distributions des scores illustrées aux figures 2.1 et 2.2 présentent une forme plus ou moins régulière : elles atteignent leur maximum avant de diminuer de manière progressive – même si on ne peut pas dire que ces figures sont parfaitement symétriques. Toutefois, toutes les distributions n’ont pas leur point culminant en leur centre et ne descendent pas de façon identique sur leurs côtés (voir la représentation en tiges et feuilles de la figure 2.7), et il est important de comprendre les termes utilisés pour décrire les différentes distributions. Observons les deux distributions illustrées aux figures 2.10 (a) et (b). Ces diagrammes représentent des données générées par ordinateur ; elles proviennent de populations présentant une forme particulière. Ces diagrammes, ainsi que les quatre autres de la figure 2.10, se basent sur des échantillons de 1000 observations, et les légères irrégularités constatées s’expliquent uniquement par la variabilité due au hasard. Les deux **distributions** illustrées aux figures 2.10 (a) et (b) sont dites **symétriques** parce qu’elles présentent la même forme de part et d’autre du centre.

La distribution de la figure 2.10 (a) provient de ce que nous appellerons par la suite une distribution normale. La distribution de la figure 2.10 (b) est qualifiée de **bimodale** étant donné ses deux sommets. Ce terme est utilisé pour désigner toute distribution comportant deux sommets prédominants, que ceux-ci aient ou non

Distribution symétrique

Distribution bimodale



exactement la même hauteur. Si une distribution ne présente qu'un seul sommet majeur, elle est dite **distribution unimodale**.

Observons ensuite les figures 2.10 (c) et (d). Ces deux distributions ne sont manifestement pas symétriques. La distribution de la figure 2.10 (c) présente une queue vers la gauche, tandis que celle de la figure 2.10 (d) se caractérise par une queue vers la droite. On parle d'asymétrie négative dans le premier cas et d'asymétrie positive dans le second (*Un conseil*: pour pouvoir les distinguer, il suffit de se rappeler que les asymétries négatives pointent vers les nombres négatifs (les petits nombres) et que les asymétries positives pointent du côté positif de l'échelle). Il existe des mesures statistiques du degré d'**asymétrie**, ou de dissymétrie, mais elles ne sont pas communément utilisées en sciences du comportement.

La figure 2.11 donne un exemple intéressant et authentique d'une distribution asymétrique positive et légèrement bimodale. Ces données ont été recueillies par Bradley (1963), qui a demandé aux participants d'appuyer le plus vite possible sur un bouton chaque fois qu'une petite lumière s'allumait. La plupart des points correspondant aux données sont distribués de façon régulière entre 7 et 17 centièmes

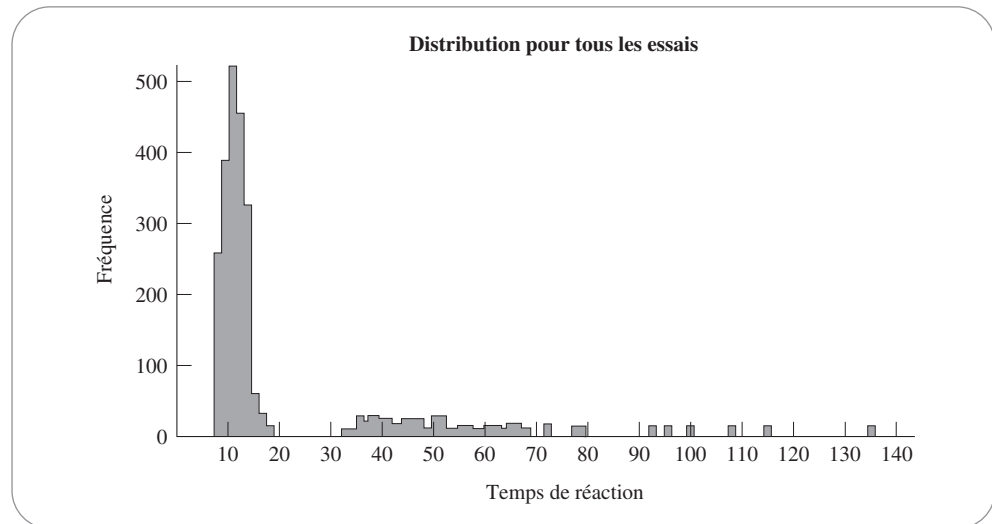
FIGURE 2.10. Formes de distributions de fréquences: (a) normale, (b) bimodale, (c) asymétrie négative, (d) asymétrie positive, (e) platycurtique et (f) leptocurtique

Distribution unimodale

Asymétrie

de seconde environ, mais on note un deuxième groupe de points, réduit, mais perceptible, entre 30 et 70 centièmes, qui s'estompe vers la droite. Ce deuxième groupe de points provient essentiellement d'essais au cours desquels le participant a manqué le bouton la première fois. Leur inclusion dans les données influence sensiblement la forme de la distribution. Un expérimentateur obtenant ce type de données pourrait sérieusement envisager de traiter séparément les temps supérieurs à un certain maximum, parce que ces temps reflètent davantage l'exactitude d'une réponse psychomotrice que la vitesse de cette réponse. Même si nous parvenions à améliorer cette distribution, nous devrions toujours nous demander si ces réponses manquées ont leur place dans les données que nous analysons.

FIGURE 2.11. Distribution de fréquences des données de Bradley relatives aux temps de réaction



Il est important de prendre en considération les différences existant entre les données de Bradley, illustrées à la figure 2.11, et les données que j'ai recueillies, présentées aux figures 2.1 et 2.2. Les deux distributions sont positivement asymétriques, mais mes temps de réaction sont plus longs et la distribution n'est pas bimodale. Notons une différence : dans le cas de mes données, il fallait décider sur quel bouton appuyer, alors que les participants de Bradley devaient uniquement appuyer sur un seul bouton chaque fois que la lumière s'allumait. La prise de décision prend du temps. En outre, le programme que j'utilisais pour présenter les stimuli enregistrait uniquement les données provenant des réponses correctes et non des erreurs. Aucune correction n'était possible, ce qui excluait l'éventualité de rater le bouton au premier essai et de devoir appuyer une seconde fois. Si je souligne ces différences, c'est pour montrer que la manière dont les données sont recueillies peut influencer sensiblement la nature des données obtenues.

Nous allons encore examiner une dernière caractéristique des distributions, à savoir la voussure (curtosis) ou l'aplatissement. La **vousure** a une définition mathématique spécifique, mais elle renvoie fondamentalement à la concentration relative de scores au centre, dans les extrémités supérieure et inférieure (les queues) et au niveau des épaules (entre le centre et les queues) d'une distribution. Dans les figures 2.10 (e) et (f), j'ai superposé une distribution normale au-dessus du diagramme des données pour permettre des comparaisons claires. Une distribution normale (qui fera l'objet d'une description détaillée au chapitre 3) est dite

mésocurtique. Ses queues ne sont ni trop fines ni trop épaisses, et il n'y a ni trop ni trop peu de scores concentrés au centre. Si, en partant d'une distribution normale, on déplace les scores du centre et des queues vers les épaules, la courbe s'aplatit et est dite platycurtique. La figure 2.10 (e) illustre clairement ce cas : la portion centrale de la distribution est beaucoup trop plate. Si, par contre, on déplace les scores des épaules vers le centre et les queues, la courbe devient plus pointue et ses queues s'épaississent. Ce type de courbe est appelé leptocurtique ; la figure 2.10 (f) en donne un exemple. Notons que dans cette distribution, les scores sont trop nombreux *à la fois* au centre et dans les queues.⁵

Sans être des mesures communément utilisées en sciences sociales, l'asymétrie et la voussure n'en constituent pas moins des caractéristiques bien pratiques pour décrire des distributions. Toute personne instruite devrait savoir ce qu'est une asymétrie positive, même s'il est peu probable qu'elle souhaite un jour calculer un indice numérique de l'asymétrie. Il faut cependant se rendre compte que ces mesures importent aux statisticiens, qui pourraient à juste titre s'énervier de me voir les considérer d'une façon aussi cavalière. Si je minimise l'importance de ces mesures, c'est parce qu'elles sont moins utiles à l'utilisateur de statistiques qu'au professionnel.

Il est important de reconnaître qu'il faut des échantillons très substantiels de données pour se faire une bonne idée de la forme d'une distribution, et en particulier de sa voussure. Si la taille des échantillons n'est que d'environ 30, on peut, au mieux, raisonnablement s'attendre à voir si les données ont plutôt tendance à s'empiler dans les queues de la distribution ou présentent une asymétrie prononcée d'un côté ou de l'autre.

Jusqu'ici, nous n'avons pratiquement pas fait mention des chiffres eux-mêmes. Nous avons vu comment les données peuvent être organisées et présentées sous forme de distributions, et nous avons discuté d'un certain nombre de façons de caractériser les distributions : la symétrie ou son absence (l'asymétrie), l'aplatissement et le caractère uni ou bimodal. Aussi utiles que ces informations puissent être dans certaines situations, elles sont insuffisantes dans d'autres. Nous ne connaissons toujours pas la vitesse moyenne du temps de réaction d'une décision simple, ni la similitude ou la dissemblance des temps de réaction pour des essais individuels. Pour obtenir ces connaissances, il nous faut réduire les données à un ensemble de mesures qui comportent les informations dont nous avons besoin. Les questions à poser concernent l'emplacement, ou la tendance centrale, et la dispersion, ou la variabilité, des distributions le long de l'échelle sous-jacente. Les mesures de ces caractéristiques sont examinées dans les sections 7 et 8. Mais avant de passer à ces sections, nous devons mettre en place un système de notation que nous pourrions utiliser dans cette présentation.

5. Je voudrais remercier Karl Wuensch de la East Carolina University pour l'aide précieuse qu'il m'a apportée dans le domaine de l'asymétrie et de la voussure. Ses idées sont reflétées dans ce passage, même si je ne suis pas certain qu'il approuverait mes commentaires sur la voussure. Sauf dans les cas extrêmes, il est peu probable que la plupart des gens, y compris les statisticiens, soient en mesure de regarder une distribution de données d'échantillon et de dire si elle est platykurtique ou leptokurtique sans calculs supplémentaires.

6. Notation

Tout examen de techniques statistiques requiert un système de notation à même d'exprimer les opérations mathématiques. Vous serez sans doute surpris d'apprendre qu'aucun système de notation standard n'a été adopté. Si de nombreuses tentatives ont été faites pour formuler une politique générale, il n'en reste pas moins qu'il est impossible de trouver deux manuels utilisant la même notation.

Les systèmes de notation courants vont du très complexe au très simple. Les systèmes plus complexes privilégient la précision aux dépens de la clarté, tandis que les systèmes les plus simples sacrifient la précision sur l'autel de l'intelligibilité. Comme la perte de précision est généralement mineure comparée au gain de compréhension, ce livre adoptera un système de notation extrêmement simple.

6.1. Notation des variables

En règle générale, une majuscule, souvent X ou Y , servira à représenter une variable. La lettre et un indice permettront de représenter une valeur individuelle de cette variable. Supposons par exemple que nous ayons les cinq scores suivants pour le temps pendant lequel des enfants de troisième année peuvent retenir leur souffle : [45, 42, 35, 23, 52]. Cet ensemble de scores sera appelé X . Le premier nombre de cet ensemble (45) peut être appelé X_1 , le deuxième (42) X_2 , etc. Lorsque nous voulons faire référence à un seul score sans préciser lequel, nous parlerons de X_i , où i peut prendre n'importe quelle valeur entre 1 et 5. En pratique, l'utilisation d'indices est souvent gênante ; ils seront donc omis si aucune confusion n'est possible.

6.2. Notation de la somme

L'un des symboles les plus courants en statistique est la majuscule grecque sigma (Σ), le symbole standard de sommation. Il se traduit généralement par « sommer, ou additionner, ce qui suit ». Ainsi, $\sum X_i$ se lit « additionner les X_i ». Pour être parfaitement correct, la notation correspondant à la somme de toutes les N valeurs de X doit être $\sum_{i=1}^N X_i$, qui se traduit par « additionner tous les X_i de $i = 1$ à $i = N$ ». En pratique, il est rarement nécessaire de spécifier aussi précisément ce qu'il faut faire, et dans la plupart des cas, tous les indices sont supprimés et la somme des X_i se note tout simplement $\sum X$.

Il convient de relever et de bien comprendre plusieurs extensions de ce cas très simple qu'est $\sum X$. L'une d'elles est $\sum X^2$, qui se lit « additionner les valeurs au carré de X » (donc, $45^2 + 42^2 + 35^2 + 23^2 + 52^2 = 8247$). Notez que ceci est très différent de $(\sum X)^2$, qui nous dit de sommer les X et de mettre le résultat au carré. Ceci serait égal à $(45 + 42 + 35 + 23 + 52)^2 = (197)^2 = 38809 = (\sum X)^2$. La règle toujours applicable, est d'effectuer les opérations entre parenthèses avant les autres opérations. Ainsi, pour $(\sum X)^2$, on additionne les valeurs de X avant de mettre le résultat au carré, alors que pour $\sum X^2$, on met les X au carré avant d'additionner.

Une autre expression courante, lorsqu'on dispose de données sur deux variables (X et Y), est $\sum XY$, qui signifie « additionner les produits des valeurs correspondantes de X et Y ». L'exemple suivant illustrera l'utilisation de ces termes et d'autres encore.

Imaginons une expérience simple pour laquelle nous enregistrons les scores d'anxiété (X) de cinq étudiants ainsi que le nombre de jours du dernier semestre durant lesquels ils ont manqué un test pour cause d'absence (Y). Les données et les opérations simples d'addition effectuées sur ces données sont illustrées dans le tableau 2.4. Certaines de ces opérations ont d'ores et déjà été examinées ; d'autres le seront dans les chapitres suivants.

	Score d'anxiété (X)	Tests manqués (Y)	X^2	Y^2	$X - Y$	XY
	10	3	100	9	7	30
	15	4	225	16	11	60
	12	1	144	1	11	12
	9	1	81	1	8	9
	10	3	100	9	7	30
Somme	56	12	650	36	44	141

TABEAU 2.4. Illustration d'opérations impliquant le symbole de sommation

$$\begin{aligned}
 \sum X &= (10 + 15 + 12 + 9 + 10) = 56 \\
 \sum Y &= (3 + 4 + 1 + 1 + 3) = 12 \\
 \sum X^2 &= (10^2 + 15^2 + 12^2 + 9^2 + 10^2) = 650 \\
 \sum Y^2 &= (3^2 + 4^2 + 1^2 + 1^2 + 3^2) = 36 \\
 \sum (X - Y) &= (7 + 11 + 11 + 8 + 7) = 44 \\
 \sum XY &= (10)(3) + (15)(4) + (12)(1) + (9)(1) + (10)(3) = 141 \\
 (\sum X)^2 &= 56^2 = 3136 \\
 (\sum Y)^2 &= 12^2 = 144 \\
 (\sum (X - Y))^2 &= 44^2 = 1936 \\
 (\sum X)(\sum Y) &= (56)(12) = 672
 \end{aligned}$$

6.3. Double indice

Il est courant d'utiliser deux indices ou plus pour préciser exactement à quelle valeur de X on fait référence. Supposons par exemple que nous analysons les données illustrées dans le tableau 2.5. Si nous voulons préciser l'entrée située dans la $i^{\text{ième}}$ ligne et la $j^{\text{ième}}$ colonne, nous noterons X_{ij} . Le score du troisième essai du Jour 2 est donc $X_{2,3} = 13$. Certains systèmes de notation utilisent $\sum_{i=1}^2 \sum_{j=1}^5 X_{ij}$, qui se traduit par « additionner les X_{ij} où i prend les valeurs 1 et 2 et j prend toutes les valeurs allant de 1 à 5 ». Il faut connaître ce système de notation puisqu'il est utilisé par certains manuels. Toutefois, ce livre préfère utiliser la notation plus simple, mais moins précise $\sum X$ chaque fois que c'est possible, n'avoir recours à la notation $\sum X_{ij}$ que quand c'est absolument nécessaire, et ne jamais employer $\sum \sum X_{ij}$.

Il est essentiel de bien comprendre les principes de notation pour apprendre même les techniques statistiques les plus élémentaires. Étudiez le tableau 2.4 jusqu'à ce que vous soyez sûr de bien comprendre toutes les procédures.

TABLEAU 2.5. Données hypothétiques illustrant la notation

		Essai					Total
		1	2	3	4	5	
Jour	1	8	7	6	9	12	42
	2	10	11	13	15	14	63
Total		18	18	19	24	26	105

7. Mesures de tendance centrale

Nous avons vu comment représenter des données de manière à pouvoir commencer à tirer des conclusions sur ce qu’elles peuvent signifier. La représentation graphique de données montre la forme générale de la distribution et donne une image de la grandeur générale des nombres en question. Cette section-ci abordera plusieurs statistiques pouvant être utilisées pour représenter le « centre » de la distribution. Elles sont appelées mesures de tendance centrale. La section suivante poursuivra en examinant les mesures liées à la dispersion des observations autour de cette tendance centrale.

L’expression **mesures de tendance centrale**, ou encore **mesures de position**, fait référence à l’ensemble des mesures liées à l’endroit où la distribution est centrée sur l’échelle. Ces mesures diffèrent quant à l’utilisation qu’elles font des données, en particulier des valeurs extrêmes, mais elles tentent toutes de nous dire quelque chose à propos de l’endroit où se situe le centre de la distribution. Les trois principales mesures de tendance centrale sont le mode, qui se base sur quelques points de données seulement, la médiane, qui fait abstraction de la plupart des données, et la moyenne, qui se calcule à partir de toutes les données. Ces mesures seront abordées tour à tour. Nous commencerons par ce qui est probablement la mesure la moins utilisée (et aussi la moins utile), le mode.

7.1. Le mode

Le **mode** peut se définir simplement comme le score le plus courant, c’est-à-dire le score obtenu par le plus grand nombre de participants. Ainsi, le mode est la valeur de *X* qui correspond au point le plus élevé de la distribution. Si deux valeurs adjacentes apparaissent à une fréquence égale (qui est aussi la plus grande), on convient généralement de prendre la moyenne de ces deux valeurs et de l’appeler mode. Si, par contre, deux valeurs *non adjacentes* apparaissent à une fréquence égale (ou presque égale), la distribution est dite bimodale et on donnera très probablement les deux modes. Par exemple, la distribution du temps consacré aux jeux électroniques est plus ou moins bimodale (voir figure 2.7), présentant des sommets aux intervalles 0-9 minutes et 40-49 minutes (on parle dans ces cas d’intervalles modaux).

Mesures de tendance centrale
Mesures de position

Mode

7.2. La médiane

La **médiane** est le score qui correspond au point auquel ou sous lequel 50 % des scores sont inclus lorsque les données sont disposées en ordre numérique. Selon cette définition, la médiane est également appelée 50^{ième} percentile.⁶ Prenons par exemple les nombres (5, 8, 3, 7, 15). Si les nombres sont disposés en ordre croissant (3, 5, 7, 8, 15), on compte autant de scores inférieurs que de scores supérieurs à 7 ; cette valeur serait donc appelée médiane. Considérons un nombre pair de scores, par exemple (5, 11, 3, 7, 15, 14). En les ordonnant, on obtient (3, 5, 7, 11, 14, 15), et aucun score n'a 50 % des valeurs en dessous de lui et 50 % des valeurs au-dessus de lui. Ce point se situe en fait entre le 7 et le 11. Dans pareil cas, la moyenne (9) des deux scores situés au milieu (7 et 11) est généralement choisie comme médiane.⁷

Médiane

Nous aurons bientôt besoin du terme **position médiane**. La position médiane de N nombres se définit comme suit :

Position médiane

$$\text{Position médiane} = \frac{N + 1}{2}$$

Ainsi, pour cinq nombres, la position médiane = $(5 + 1)/2 = 3$, ce qui signifie tout simplement que la médiane est le troisième nombre dans une série ordonnée. Pour 12 nombres, la position médiane = $(12 + 1)/2 = 6.5$; la médiane se situe entre le sixième et le septième nombre et correspond à la moyenne de ces deux nombres.

Pour ce qui est des données relatives aux temps de réaction présentées dans le tableau 2.2, la position médiane = $(300 + 1)/2 = 150.5$. Lorsque les données sont classées par ordre, le 150^{ième} temps est 59 et le 151^{ième} est 60 ; donc, la médiane est $(59 + 60)/2 = 59.5$ centièmes de seconde. Vous pouvez en faire le calcul à partir du tableau 2.2. Les données relatives aux jeux électroniques comprennent 100 scores, et la position médiane est 50.5. La représentation en tiges et feuilles illustrée par la figure 2.7 montre que le cinquantième score est 44 et que le cinquante et unième score est 46. La médiane est égale à 45, c'est-à-dire la moyenne de ces deux valeurs.

7.3. La moyenne

La moyenne est la mesure la plus commune de tendance centrale ; elle nécessite peu d'explications. La **moyenne** ($\bar{X}$) est la somme des scores divisée par le nombre de scores et est souvent notée $\bar{X}$ (lire « X barre »).⁸ Si l'on utilise la notation de la somme donnée ci-dessus, elle se définit comme suit :

Moyenne

6. Le percentile spécifique se définit comme le point de l'échelle auquel ou sous lequel se situe un pourcentage spécifique des scores.

7. La définition de la médiane est une source de discordance chez les statisticiens. La définition donnée ici, qui définit la médiane comme un *point* dans une distribution de nombres, est la préférée de la plupart des statisticiens. Elle se conforme en outre à l'énonciation disant que la médiane est le 50^{ième} percentile. Par ailleurs, beaucoup affirment que la médiane est soit le *nombre* du milieu dans une série ordonnée (si N est impair), soit la moyenne des deux *nombres* du milieu (si N est pair). Ces controverses font penser à une réunion dont l'ordre du jour ne comprend aucun point essentiel. Moins la question est importante, plus on en parle.

8. L'American Psychological Association voudrait que l'on utilise M pour la moyenne au lieu de $\bar{X}$, mais j'ai utilisé $\bar{X}$ depuis tant d'années que j'aurais vraiment beaucoup de mal à y renoncer. Le reste du monde statistique est généralement d'accord avec moi et nous allons donc utiliser $\bar{X}$ dans le reste du texte.

$$\bar{X} = \frac{\sum X}{N}$$

où $\sum X$ est la somme de toutes les valeurs de X , et N est le nombre de valeurs X . Par exemple, la moyenne des nombres 3, 5, 12 et 5 est égale à

$$\frac{3 + 5 + 12 + 5}{4} = \frac{25}{4} = 6.25$$

En ce qui concerne les données relatives aux temps de réaction présentées dans le tableau 2.2, la somme des observations est égale à 18078. En divisant ce nombre par $N = 300$, on obtient $18078/300 = 60.26$. Notez que cette réponse s'accorde bien avec la médiane, qui est égale à 59.5. La moyenne et la médiane seront proches chaque fois que la distribution est presque symétrique (selon la définition donnée dans la section 5). De même, cette valeur s'accorde ici relativement bien avec l'intervalle modal (60-64).

7.4. *Avantages et inconvénients relatifs du mode, de la médiane et de la moyenne*

La moyenne et la médiane ne seront égales que si la distribution est symétrique, et les trois mesures ne seront équivalentes que si la distribution est symétrique et unimodale. Dans tous les autres cas, y compris presque toutes les autres situations dont nous devrons débattre, il faudra choisir une mesure de tendance centrale. Il n'existe pas de bonnes règles pour sélectionner une mesure de tendance centrale, mais il est possible de faire des choix intelligents parmi les trois mesures.

7.4.1. Le mode

Le mode est le score le plus couramment enregistré. C'est donc, par définition, un score qui a réellement été enregistré, alors que la moyenne et parfois la médiane peuvent correspondre à des valeurs absentes des données. Le mode présente également l'avantage évident de représenter le plus grand nombre de gens. Le gérant d'un petit magasin devrait s'axer sur le mode. Si 80 % des clients souhaitent acheter le grand format « familial » de détergent et que 20 % préfèrent le petit format « individuel », il ne serait pas particulièrement judicieux de s'appuyer sur une autre mesure du centre de la distribution et d'opter uniquement pour le format « standard ».

7.4.2. La médiane

Tout comme le mode, la médiane présente l'avantage principal de ne pas être influencée par les scores extrêmes. La médiane de (5, 8, 9, 15, 16) et de (0, 8, 9, 15, 206) est égale à 9. Cette caractéristique s'avère intéressante pour de nombreux expérimentateurs dans des études enregistrant quelquefois des scores extrêmes dépourvus de signification particulière. Par exemple, le rat moyen peut, après apprentissage, parcourir un petit trajet en 1 à 2 secondes environ. Une fois de temps en temps, ce même rat risque de s'arrêter sans raison à la mi-parcours, de commencer à se gratter, d'aller renifler les photocellules puis de s'endormir. Dans ce cas, que le rat mette 30 secondes ou 10 minutes pour parcourir le labyrinthe n'a pas d'importance pratique. Cela peut même dépendre du moment où l'expérimentateur abandonne et lui donne un petit coup de crayon pour le réveiller. Si les trois essais effectués par le rat

en un jour donnaient des temps de (1.2, 1.3 et 20 secondes), cela aurait pour nous la même signification – quant à la connaissance de la tâche par le rat – que des temps de (1.2, 1.3 et 136.4 secondes). Dans les deux cas, la médiane serait égale à 1.3, alors que, de toute évidence, sa moyenne quotidienne serait très différente (7.5 contre 46.3 secondes). Ce problème incite fréquemment les expérimentateurs à travailler avec le temps médian plutôt qu’avec le temps moyen par jour.

7.4.3. La moyenne

La moyenne est de loin la plus courante des trois principales mesures de tendance centrale. Il ne serait pas exagéré de dire que pour beaucoup, la statistique est presque synonyme d’étude de la moyenne.

Comme nous l’avons vu, la moyenne présente certains inconvénients. Elle est influencée par les scores extrêmes, sa valeur pourrait être absente des données et son interprétation pour la variable sous-jacente mesurée nécessite une certaine confiance dans le fait que les données présentent les propriétés d’une échelle d’intervalle. Vous allez peut-être dire que si la moyenne comporte réellement tous ces inconvénients, il vaut peut-être mieux l’oublier et l’abandonner définitivement, comme d’autres statistiques telles que le « rapport critique », concept statistique dont on ne parle plus depuis des années. Mais la moyenne est d’une autre trempe.

La moyenne se caractérise en effet aussi par plusieurs avantages importants qui compensent largement ses inconvénients. Le principal avantage d’un point de vue historique (mais peut-être pas de votre point de vue) est probablement la possibilité de manipuler algébriquement la moyenne. Autrement dit, on peut utiliser la moyenne dans une équation et la manipuler par les règles normales d’algèbre, en particulier parce qu’on peut écrire une équation qui définit la moyenne. Comme on ne peut pas écrire une équation standard du mode ou de la médiane, il n’est pas réellement possible de manipuler ces statistiques par les règles algébriques standard. Cet avantage de la moyenne explique en grande partie sa large utilisation en dépit de ses défauts. Elle présente encore d’autres avantages, à savoir plusieurs caractéristiques souhaitables quant à son utilisation comme estimation de la tendance centrale de la population. En particulier, si l’on prélève plusieurs échantillons d’une population, les moyennes de ces échantillons constituent des estimations plus stables (moins variables) de la tendance centrale de cette population que ne le sont les médianes ou les modes de ces échantillons. Si la moyenne d’un échantillon est si couramment utilisée, c’est en grande partie parce qu’elle donne généralement une meilleure estimation de la tendance centrale de la population que le mode ou la médiane.

7.5. Moyennes tronquées

Les moyennes tronquées sont des moyennes calculées sur des données parmi lesquelles nous avons écarté un certain pourcentage des points se trouvant à chaque extrémité de la distribution. Par exemple, si nous disposons d’un ensemble de 100 observations et que nous voulons calculer une moyenne tronquée de 10 %, nous éliminons simplement les 10 scores les plus élevés et les 10 scores les plus bas et nous prenons la moyenne des scores restants. Il s’agit d’une vieille idée qui revient à la mode, et son plus grand défenseur est peut-être Rand Wilcox (Wilcox, 2003, 2005).

Il y a plusieurs raisons de tronquer un échantillon. Comme je l’ai mentionné au chapitre 1, et comme je le répéterai tout au long du livre, l’un des principaux objectifs de déterminer la moyenne d’un échantillon est d’estimer la moyenne de la population à partir de laquelle cet échantillon a été prélevé. Si nous voulons une bonne estimation, nous attendons de celle-ci qu’elle varie peu d’un échantillon à l’autre. (Pour utiliser un terme que nous définirons dans les chapitres suivants, nous voulons une estimation avec une erreur standard minimale). Si notre échantillon est très dispersé, c’est-à-dire s’il comporte beaucoup de scores élevés et de scores bas, sa moyenne ne sera pas une très bonne estimation de la moyenne de la population. En éliminant les valeurs extrêmes de l’échantillon, notre estimation de la moyenne de la population est plus stable. C’est donc souvent un avantage, même si nous voulons être sûrs de ne pas éliminer des informations utiles.

Une autre raison de tronquer un échantillon est de contrôler les problèmes d’asymétrie. Si vous avez une distribution très asymétrique, les valeurs extrêmes vont tirer la moyenne vers elles et conduire à une moins bonne estimation de la moyenne de la population. Tronquer permet d’éliminer l’influence de ces résultats extrêmes. Mais considérez les données de Bradley (1963) sur les temps de réaction, présentées à la figure 2.10. Je conviens que les longs temps de réaction sont probablement dus au fait que la personne interrogée a manqué le bouton, et qu’ils ne sont donc pas liés à un temps de réaction au sens strict du terme, et pourraient légitimement être éliminés, mais voulons-nous vraiment éliminer le même nombre d’observations à l’autre extrémité de l’échelle ?

Wilcox a fait un travail considérable sur les problèmes liés au fait de tronquer une moyenne, et je respecte évidemment sa réputation bien méritée. En outre, je pense que les étudiants ont besoin de connaître les moyennes tronquées, car elles font l’objet de discussions dans la littérature. Mais je ne pense pas pouvoir aller aussi loin que Wilcox dans la promotion de leur utilisation. Ceci étant dit, je ne pense pas que ma réticence doive vous dissuader de considérer sérieusement la question ; je recommande d’ailleurs le livre de Wilcox (2003).

8. Mesures de variabilité

La section précédente a évoqué plusieurs mesures liées au centre d’une distribution. Toutefois, une valeur de tendance centrale de la distribution (qu’il s’agisse du mode, de la médiane ou de la moyenne) ne nous renseigne pas de façon complète. Il faut d’autres informations pour indiquer dans quelle mesure les différentes observations sont groupées autour de cette valeur de tendance centrale ou s’en écartent. Il se peut que la valeur de tendance centrale reflète la position générale de la plupart des scores, ou que les scores soient distribués sur une large étendue de valeurs, et que la « valeur de tendance centrale » ne soit pas très représentative de l’ensemble total d’observations. Tout le monde a connu des examens auxquels tous les étudiants obtenaient plus ou moins la même cote, et d’autres dont les résultats allaient d’« excellent » à « très mauvais ». C’est aux mesures faisant référence aux différences entre ces deux types de situation qu’on pense lorsqu’on parle de la **dispersion**, ou variabilité, autour de la médiane, du mode ou de tout autre point. En général, on parlera plus spécialement de la dispersion autour de la moyenne.

EXEMPLE

ILLUSTRATION DE L'AUGMENTATION DE LA VARIABILITÉ DU POIDS EN FONCTION DE L'ÂGE DES NOURRISSONS



Un exemple pour illustrer la variabilité nous vient de Weaver (1999) et s'appuie sur quelque chose qui vous paraîtra à coup sûr familier – un graphique portant sur la croissance attendue des nourrissons. La figure 2.12 fournit un tel graphique. La moitié inférieure du graphique reprend l'étendue normale du poids de filles entre la naissance et l'âge de 36 mois. La ligne en gras intitulée « 50 » au centre représente le poids médian à chaque âge. Les deux lignes de chaque côté représentent les limites dans lesquelles on peut espérer observer la moitié de la distribution. Les deux lignes suivantes de chaque côté incluent, respectivement, les 80 % et les 90 % des enfants. Cette figure permet de voir très facilement l'augmentation de la dispersion à mesure qu'augmente l'âge des nourrissons. Les poids des nouveaux-nés se situent à moins d'un kilo de la moyenne tandis que les poids des enfants de trois ans sont répartis sur environ 3 kilos de chaque côté de la moyenne. De toute évidence, la moyenne augmente également, même si nous nous attardons ici davantage sur la question de la dispersion.

EXEMPLE

VARIABILITÉ DES SCORES DE PERCEPTION DE LA BEAUTÉ DES VISAGES



En guise de second exemple, prenons des données intéressantes recueillies par Langlois et Roggman (1990) sur la perception de la beauté des visages. Réfléchissons un moment à certains visages que nous trouvons beaux. Ont-ils tendance à présenter des caractéristiques originales (nez proéminent, forme inhabituelle des sourcils) ou des traits plutôt ordinaires? Langlois et Roggman voulaient étudier ce qui rend les visages beaux. Dans ce but, ils ont montré à des étudiants des images de visages générées par ordinateur. Certaines de ces images avaient été créées via la fusion de photos de quatre personnes différentes en vue de produire une photo composite. Ces images seront désignées comme étant l'ensemble 4. D'autres images (l'ensemble 32) ont été créées via la fusion de 32 visages différents. Comme on pourrait s'y attendre, lorsqu'on fusionne quatre visages, une certaine individualité reste possible dans la photo composite. Ainsi, certaines photos peuvent présenter des visages fins et d'autres des visages ronds. Par contre, la fusion de 32 photos donne généralement des résultats très proches d'une « tendance centrale ». Les nez ne sont ni trop longs ni trop courts, les oreilles ne sont ni décollées ni collées à l'excès, etc. Des étudiants ont été invités à examiner les images ayant résulté de ces fusions et à les évaluer toutes sur une « échelle de beauté » graduée de un à cinq. Les auteurs souhaitaient essentiellement savoir si les visages de l'ensemble 4 étaient *en moyenne* moins bien évalués que les visages de l'ensemble 32. Et ce fut effectivement le cas, ce qui porte à croire que les visages dotés de caractéristiques distinctives sont perçus comme étant moins beaux que les visages plus classiques. Cette section porte sur la mesure de la variabilité. Intéressons-nous au degré de *similitude* des évaluations des visages. À mon avis, les images fusionnées de 32 visages devraient être plus homogènes, et entraîner par conséquent une évaluation plus similaire, que les fusions de quatre visages.

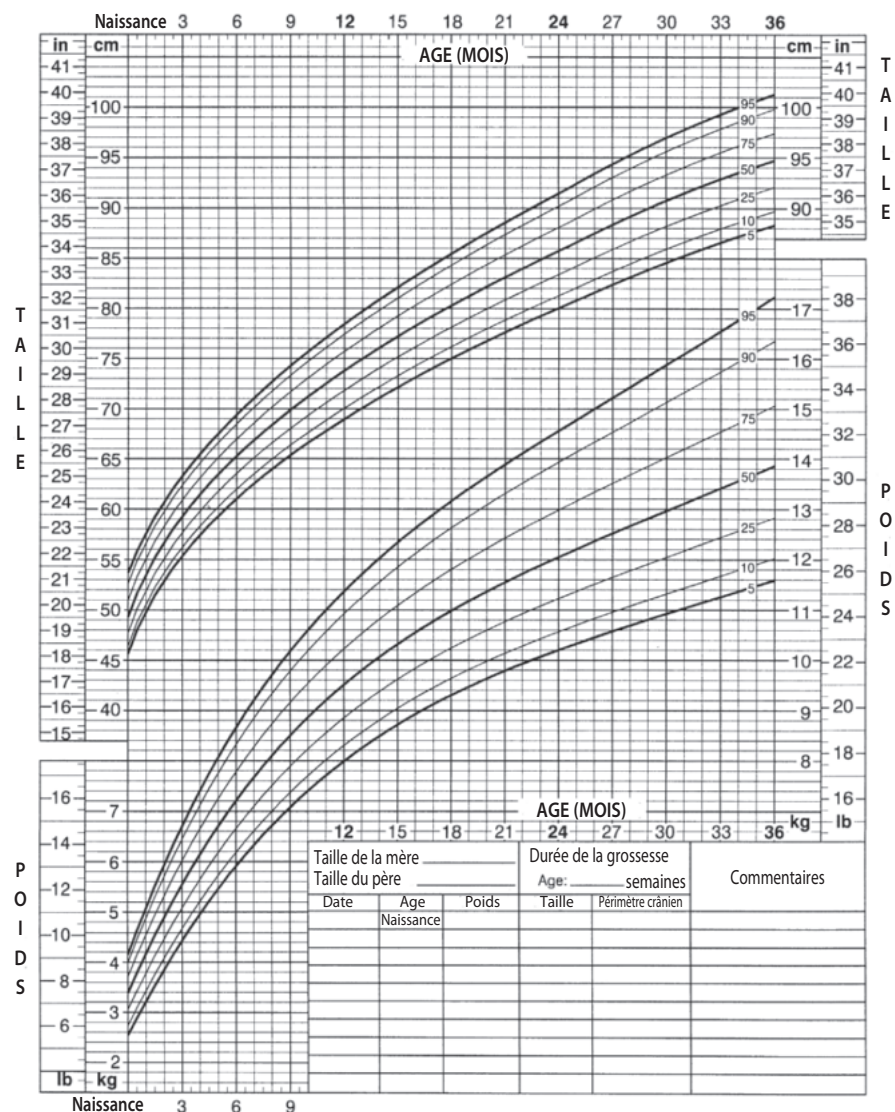
Les données figurent dans le tableau 2.6.⁹ Comme on le voit, Langlois et Roggman avaient raison de prédire que les visages de l'ensemble 32 seraient jugés plus beaux que ceux de l'ensemble 4 (Les moyennes étaient respectivement égales à 3.26 et 2.64). Notons toutefois aussi que les évaluations des images résultant de la fusion de 32 visages sont sensiblement plus homogènes que celles des images fusionnant 4 visages. La figure 2.13 présente ces deux ensembles de données sous la forme d'un histogramme standard.

Si la figure 2.13 met clairement en évidence une variabilité plus importante dans l'évaluation des fusions de quatre photos que dans l'évaluation des fusions de 32 photos, une mesure est nécessaire pour refléter cette différence de variabilité. On pourrait utiliser différentes mesures, qui seront évoquées l'une après l'autre, en commençant par les plus simples.

9. Ces données ne sont pas celles réellement récoltées par Langlois et Roggman, mais elles ont été générées de façon à obtenir exactement la même moyenne et le même écart-type que les données d'origine. Langlois et Roggman ont utilisé six photos composites par ensemble. Quant à moi, j'ai utilisé 20 photos par ensemble afin de mieux adapter les données au propos de ce chapitre. Ces données fourniront toutefois les mêmes conclusions que les données de Langlois et Roggman.

Distribution du poids et de la taille des nourrissons en fonction de l'âge

Courbes de croissance des filles de la naissance à 36 mois :
percentiles de la taille et du poids en fonction de l'âge



Revised November 21, 2000.

SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000). <http://www.cdc.gov/growthcharts>



Ensemble 4		Ensemble 32	
Image	Ensemble de 4 visages	Image	Ensemble de 32 visages
1	1.20	21	3.13
2	1.82	22	3.17
3	1.93	23	3.19
4	2.04	24	3.19
5	2.30	25	3.20
6	2.33	26	3.20
7	2.34	27	3.22
8	2.47	28	3.23
9	2.51	29	3.25
10	2.55	30	3.26
11	2.64	31	3.27
12	2.76	32	3.29
13	2.77	33	3.29
14	2.90	34	3.30
15	2.91	35	3.31
16	3.20	36	3.31
17	3.22	37	3.34
18	3.39	38	3.34
19	3.59	39	3.36
20	4.02	40	3.38
Moyenne = 2.64		Moyenne = 3.26	

TABLEAU 2.6. Beauté perçue des visages composites

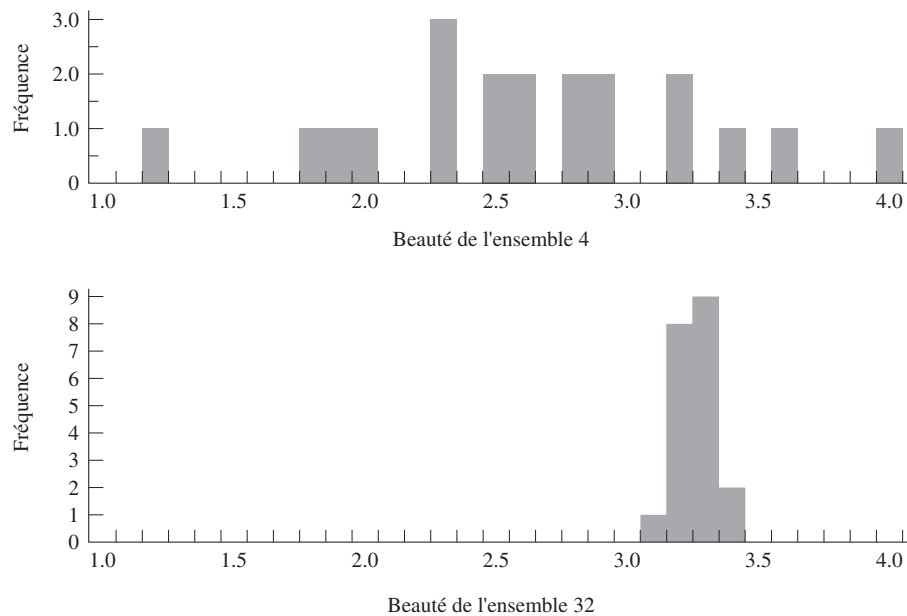


FIGURE 2.13. Distribution des scores de beauté des photos composites

8.1. L'étendue

Étendue

L'**étendue** est une mesure de distance, à savoir la distance entre le score le moins élevé et le score le plus élevé. Pour nos données, l'étendue est de $(4.02 - 1.20) = 2.82$ unités pour l'ensemble 4 et de $(3.38 - 3.13) = 0.25$ unité pour l'ensemble 32. L'étendue est une mesure de plus en plus courante ; nous y faisons référence dans la vie quotidienne lorsque nous disons par exemple que « le prix d'un hebdomadaire varie de quelques euros à une dizaine d'euros ». Cependant, l'étendue dépend uniquement des valeurs extrêmes ou, si ces valeurs sont exceptionnellement extrêmes, de ce qu'on appelle les *outliers* ou valeurs aberrantes. Il en résulte que l'étendue peut donner une image déformée de la variabilité.

8.2. Écart interquartile et autres statistiques liées à l'étendue

Écart interquartile

L'**écart interquartile** est une tentative visant à contourner le problème de l'énorme dépendance de l'étendue vis-à-vis des scores extrêmes. On obtient un écart interquartile en éliminant les 25 % supérieurs et les 25 % inférieurs de la distribution et en prenant l'étendue restante. Le point qui délimite les 25 % inférieurs de la distribution est appelé le premier quartile, et est habituellement noté Q_1 . De façon similaire, le point qui délimite les 25 % supérieurs de la distribution est appelé le troisième quartile et est noté Q_3 . (La médiane est le deuxième quartile, Q_2 .) La différence entre le premier et le troisième quartile ($Q_3 - Q_1$) est l'écart interquartile. On peut calculer l'écart interquartile des données relatives à la beauté des visages en négligeant les cinq scores les moins élevés et les cinq scores les plus élevés et en déterminant l'étendue restante. Dans ce cas, l'écart interquartile de l'ensemble 4 est égal à 0.58 tandis que celui de l'ensemble 32 n'atteint que 0.11. L'écart interquartile joue un rôle essentiel dans une méthode graphique très utile intitulée diagramme en boîte (*boxplot*). Cette méthode sera abordée à la section 10.

L'écart interquartile pose, à divers égards, des problèmes opposés à ceux qui caractérisent l'étendue. En particulier, il élimine une trop grande partie des données. Quand on veut savoir si les évaluations sont plus variables pour un ensemble de photos que pour un autre, il n'est pas très pertinent de rejeter les scores qui sont les plus extrêmes et donc varient le plus par rapport à la moyenne.

Il n'y a rien de sacrilège au fait d'éliminer les 25 % inférieurs et supérieurs de la distribution avant de calculer l'étendue. En fait, on peut éliminer n'importe quel pourcentage pour autant qu'on puisse justifier ce nombre pour soi et pour les autres. Ce qu'on veut vraiment faire, c'est éliminer les scores les plus susceptibles d'être des erreurs ou des scores dus à des événements inhabituels sans éliminer la variabilité qu'on cherche à étudier.

Dans une section précédente, nous avons abordé l'utilisation d'échantillons tronqués pour générer des moyennes tronquées. Le fait de tronquer peut s'avérer une approche intéressante lorsqu'on est confronté à des distributions asymétriques ou qui recèlent des valeurs aberrantes. Néanmoins, lorsque nous utilisons des **échantillons tronqués** pour estimer la variabilité, nous avons recours à une variation basée sur ce que nous appelons un **échantillon Winsorisé**. Nous créons, par exemple, un échantillon Winsorisé de 10 % en remplaçant les scores correspondant aux 10 % inférieurs par des copies du plus petit score restant, et en remplaçant les scores

Échantillon tronqué
Échantillon Winsorisé

correspondant aux 10 % supérieurs par des copies du plus grand score restant, et puis nous calculons la mesure de variabilité sur les données ainsi modifiées.

8.3. L'écart moyen

Au premier abord, il semblerait que si l'on veut mesurer le degré de dispersion des scores autour de la moyenne (c'est-à-dire leur écart par rapport à la moyenne), le plus logique soit d'obtenir tous les écarts (c'est-à-dire $X_i - \bar{X}$) et d'en faire la moyenne. On pourrait raisonnablement croire que plus les scores sont dispersés, plus les écarts sont importants et plus la moyenne des écarts est importante. Or, ici, le bon sens induit en erreur. Si l'on calcule les écarts par rapport à la moyenne, certains scores se situent au-delà de la moyenne et accuseront un écart positif, tandis que d'autres se situent en deçà de la moyenne et présenteront un écart négatif. Au bout du compte, les écarts positifs et négatifs se compenseront mutuellement et la somme des écarts sera nulle. On ne sera pas plus avancé.

8.4. L'écart absolu moyen

Si l'on pense à la difficulté d'essayer de retirer quelque chose d'utile de la moyenne des écarts, on pourrait être amené à suggérer de résoudre le problème en prenant les valeurs absolues des écarts. (La valeur absolue d'un nombre est la valeur de ce nombre indépendamment de son signe algébrique. La valeur absolue est indiquée par l'ajout de deux traits verticaux autour du nombre, p. ex. $|-3| = 3$). L'idée des valeurs absolues est judicieuse, parce qu'on veut savoir *dans quelle mesure* les scores s'écartent de la moyenne sans se préoccuper de savoir s'ils se trouvent au-delà ou en deçà de cette valeur. La mesure proposée ici est une mesure parfaitement légitime, qu'on appelle **écart absolu moyen** (EAM). La somme des écarts absolus est divisée par N (le nombre de scores) pour donner un écart moyen : EAM. En dépit de sa simplicité et de son attrait intuitif, l'écart absolu moyen ne joue pas un rôle majeur dans les méthodes statistiques. On lui préfère généralement des mesures beaucoup plus utiles, la variance et l'écart-type.

Écart absolu moyen

8.5. La variance

Les mesures dont nous avons parlé jusqu'à présent intéressent bien davantage les statisticiens que les psychologues et leurs pairs. Ce sont des notions que vous devez connaître, mais la variance, et l'écart-type qui suivra, occuperont la majeure partie de notre attention. La mesure qui va à présent nous occuper, la **variance de l'échantillon** (s^2), est une autre solution au problème des écarts dont la moyenne vaut zéro (Pour faire référence à la **variance de la population**, nous utilisons le symbole σ^2 [sigma minuscule au carré]). Dans le cas de la variance, nous tirons profit du fait que le carré d'un nombre négatif est positif. Nous additionnons donc les écarts *au carré* plutôt que les écarts absolus. Comme nous voulons une moyenne, nous divisons ensuite cette somme par une fonction de N , le nombre de scores. Nous pourrions penser à diviser cette somme par N , mais nous la diviserons en fait par $(N - 1)$. Nous utilisons $(N - 1)$ pour la *variance de l'échantillon uniquement* parce que, comme nous le verrons sous peu, cela donne une variance d'échantillon qui permet de mieux estimer la variance de la population correspondante (La variance de la population se calcule en divisant la somme des écarts au carré, pour chaque valeur de la

Variance de l'échantillon

Variance de la population

population, par N au lieu de $(N - 1)$. Cependant, il est rare qu'on calcule la variance d'une population ; on en fait presque toujours une estimation à partir de la variance d'un échantillon).

S'il est important de préciser la variable à laquelle s^2 se réfère, on place en indice la lettre qui représente la variable. Donc, si l'on note les données de l'ensemble 4 comme étant X , la variance peut se noter s_X^2 (On pourrait utiliser $s_{\text{ensemble 4}}^2$, mais il n'est pas pratique d'avoir de longs indices. On désigne généralement les variables par de simples lettres telles que X et Y).

Pour notre exemple, nous pouvons calculer comme suit les variances des échantillons de l'ensemble 4 et de l'ensemble 32¹⁰ :

$$\begin{aligned} \text{Ensemble 4}(X) \\ s_X^2 &= \frac{\sum (X - \bar{X})^2}{N - 1} \\ &= \frac{(1.20 - 2.64)^2 + (1.82 - 2.64)^2 + \cdots + (4.02 - 2.64)^2}{20 - 1} \\ &= \frac{8.1569}{19} = 0.4293 \end{aligned}$$

$$\begin{aligned} \text{Ensemble 32}(Y) \\ s_Y^2 &= \frac{\sum (Y - \bar{Y})^2}{N - 1} \\ &= \frac{(3.13 - 3.26)^2 + (3.17 - 3.26)^2 + \cdots + (3.38 - 3.26)^2}{20 - 1} \\ &= \frac{0.0903}{19} = 0.0048 \end{aligned}$$

Ces calculs montrent que la différence entre les variances reflète les différences entre les distributions.

Même si la variance constitue un concept d'une importance essentielle et l'une des statistiques les plus fréquemment utilisées, elle ne permet pas l'interprétation intuitive directe qu'on souhaiterait. Comme elle se base sur les écarts *au carré*, le résultat est en unités au carré. Pour l'ensemble 4, l'estimation moyenne de la beauté est donc de 2.64 et la variance est égale à 0.4293 unité *au carré*. Or, les unités au carré ne sont pas pratiques pour travailler et n'ont pas beaucoup de sens par rapport aux données. Heureusement, la solution à ce problème est toute trouvée : il suffit de prendre la racine carrée de la variance.

8.6. L'écart-type

Écart-type (s ou σ)

L'**écart-type** (s ou σ) se définit comme la racine carrée positive de la variance ; pour un échantillon, elle est symbolisée par s (avec un indice permettant d'identifier la variable si nécessaire) ou parfois SD pour standard deviation ou ET pour écart-type.¹¹

10. Pour ces calculs comme pour les autres calculs du livre, il se peut que vous obteniez des réponses légèrement différentes des miennes pour les mêmes données. Si c'est le cas, ces différences sont très probablement des différences d'arrondi. Si vous refaites mes calculs et arrivez à des réponses *similaires*, c'est suffisant.

11. L'American Psychological Association propose d'utiliser l'abréviation « SD » (standard deviation), mais la plupart des statisticiens préfèrent utiliser « s »

(La notation σ est utilisée pour faire référence à l'écart-type d'une *population*). L'écart-type est défini par la formule suivante :

$$s_X = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

Pour notre exemple,

$$s_X = \sqrt{s_X^2} = \sqrt{0.4293} = 0.6552$$

$$s_Y = \sqrt{s_Y^2} = \sqrt{0.0048} = 0.0689$$

Par souci pratique, nous arrondirons respectivement ces réponses à 0.66 et 0.07.

Si l'on examine la formule de l'écart-type, on voit que, tout comme l'écart absolu moyen, l'écart-type est fondamentalement une mesure de la moyenne des écarts de chaque score par rapport à la moyenne. Bien sûr, ces écarts ont été mis au carré, additionnés, etc., mais au fond, il s'agit toujours d'écarts. Et même si l'on a divisé par $(N - 1)$ au lieu de N , on a tout de même obtenu un résultat qui se rapproche très fort d'une moyenne de ces écarts. On peut donc dire sans trop déformer la vérité que l'estimation de la beauté de l'ensemble 4 présente en moyenne un écart de 0.66 unité par rapport à la moyenne, tandis que pour l'ensemble 32, l'écart est en moyenne de 0.07 unité seulement. Cette conception, qui revient à considérer l'écart-type comme une sorte d'écart moyen, contribue largement à lui donner un sens tout en évitant de dénaturer le concept.

Ces résultats permettent de tirer deux conclusions intéressantes quant à la beauté. D'abord, si vous faisiez partie des participants à cette expérience, le fait que la fusion par ordinateur de plusieurs visages produit des photos composites similaires se refléterait par la faible variabilité qui caractérise les évaluations de l'ensemble 32 : toutes ces images sont jugées fondamentalement semblables. Ensuite, le fait que la moyenne de l'ensemble 32 est plus élevée que celle de l'ensemble 4 montre que fusionner plusieurs visages donne des visages composites qui semblent plus beaux. Ces constatations correspondent-elles à votre expérience quotidienne ? Personnellement, j'aurais pensé, à tort, que les visages dotés de traits caractéristiques auraient été les plus appréciés. Réfléchissez à nouveau aux visages que vous trouvez beaux. Sont-ils réellement caractéristiques ? Si c'est le cas, pouvez-vous avancer une hypothèse supplémentaire pour expliquer les résultats ?

On peut également examiner l'écart-type du point de vue du nombre de scores qui se situent à moins d'un écart-type au-delà ou en deçà de la moyenne. Pour un large éventail de distributions raisonnablement symétriques et présentant une partie centrale arrondie, on peut dire que les deux tiers environ des observations se situent à moins d'un écart-type de la moyenne (et pour une distribution normale, dont nous parlerons au chapitre 3, il s'agit presque exactement des deux tiers). Cette règle est très utile même si elle admet sans doute des exceptions, en particulier pour les distributions asymétriques. Si je vous disais que pour les enseignants du primaire, le salaire annuel moyen devrait s'élever à \$ 39 259 avec un écart type de \$ 4 000, on ne se tromperait pas beaucoup en concluant que les deux tiers environ de ces enseignants gagneront entre \$ 35 000 et \$ 43 000. En outre, la plupart des enseignants (95 %) se situent à moins de 2 écarts-types de la moyenne. (Mais si vous vous

intéressez vraiment aux salaires, la médiane serait probablement une meilleure mesure que la moyenne).

8.7. Formules de calcul de la variance et de l'écart-type

Les expressions citées précédemment pour la variance et l'écart-type sont parfaitement correctes, mais très peu maniables pour toute quantité raisonnable de données. Elles sont également susceptibles d'entraîner des erreurs d'arrondi puisqu'elles impliquent une mise au carré d'écarts fractionnaires. Elles constituent d'excellentes formules de définition, mais nous allons à présent considérer un ensemble plus pratique de formules de calcul. D'un point de vue algébrique, ces formules sont équivalentes aux précédentes; elles donneront donc les mêmes réponses tout en nécessitant moins d'efforts.

La formule de définition de la variance d'un échantillon était la suivante :

$$s_X^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

En voici une formule de calcul plus pratique :

$$s_X^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}$$

De même, pour l'écart-type d'un échantillon :

$$\begin{aligned} s_X &= \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}} \\ &= \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}} \end{aligned}$$

Des personnes dont je respecte l'opinion ont récemment suggéré que je ne devais plus présenter de telles formules dans ce livre et, ce, parce que les gens ne calculent plus que très rarement les variances à la main. Même si ce dernier point est vrai et que je n'utilise moi-même que très rarement ces formules dans mes cours, beaucoup de collègues continuent de penser qu'il est important de pouvoir réaliser ces calculs. De manière plus importante, il se fait que ces formules vont surgir à plusieurs endroits dans ce livre sous des formes un peu différentes et le fait de comprendre ce que sont en fait ces autres formules permet de mieux comprendre ce qui se passe. En revanche, j'ai moi aussi des doutes sur l'utilité de présenter des formules très complexes. C'est pourquoi j'ai revu les éditions récentes de ce texte en me concentrant surtout sur les formes définitionnelles des formules.

Si nous appliquons la formule de calcul applicable à la variance de l'échantillon pour l'ensemble 4, nous obtenons :

$$\begin{aligned}
 s_x^2 &= \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1} \\
 &= \frac{1.20^2 + 1.82^2 + \dots + 4.02^2 - \frac{52.89^2}{20}}{19} \\
 &= \frac{148.0241 - \frac{52.89^2}{20}}{19} = 0.4293
 \end{aligned}$$

Notons que la réponse obtenue ici est identique à celle obtenue via la formule de définition. Notons également, comme nous l'avons déjà souligné, que $\sum X^2 = 148.0241$ n'est pas égal à $(\sum X)^2 = 52.89^2 = 2797.35$. Je vous laisse le soin de calculer la variance pour l'ensemble 32.

Pour vous rassurer, sachez que dans le reste du livre, vous n'aurez pas besoin de connaissances mathématiques supérieures à celles que nécessitent ces calculs-ci – non pas parce que je souhaite diminuer la difficulté de la matière, mais parce qu'une compréhension de la plupart des statistiques appliquées ne nécessite guère plus de connaissances des mathématiques avancées. (Comme je l'ai dit, il s'agit du niveau mathématique de l'enseignement secondaire.)

8.8. Influence des valeurs extrêmes sur la variance et sur l'écart-type

La variance et l'écart-type sont très sensibles à l'influence des scores extrêmes. En d'autres termes, les scores extrêmes jouent un rôle disproportionné dans la détermination de la variance. Prenons un ensemble de données allant de 0 à 10 environ, avec une moyenne de 5. La formule de définition de la variance nous apprend qu'un score de 5 (la moyenne) ne contribue en rien à la variance, puisque l'écart par rapport à la moyenne est 0. Un score de 6 contribue de $1/(N - 1)$ unités à s^2 , parce que $(X - \bar{X})^2 = (6 - 5)^2 = 1$. Toutefois, un score de 10 contribue de $25/(N - 1)$ unités à s^2 , parce que $(10 - 5)^2 = 25$. Donc, si 6 et 10 s'écartent respectivement de 1 et de 5 unités par rapport à la moyenne, leur contribution relative à la variance est de 1 et 25. C'est ce que nous voulons dire en affirmant que les écarts importants sont représentés de manière disproportionnée. Il serait bon de nous en souvenir si nous utilisons un instrument de mesure qui « convient parce qu'il est fiable pour tout sauf pour les extrêmes ». Ce sont justement ces extrêmes qui pourraient avoir les principales répercussions sur l'interprétation des données. C'est l'une des principales raisons pour lesquelles nous n'aimons pas particulièrement avoir des données asymétriques.

8.9. Le coefficient de variation

L'une des opérations les plus courantes en statistique consiste à comparer les moyennes de deux ou plusieurs groupes, ou même de deux ou plusieurs variables. Toutefois, comparer la variabilité de ces groupes ou variables est une activité tout aussi légitime et utile. Supposons par exemple que nous disposions de deux tests concurrents visant à évaluer la mémoire à long terme.

Le premier produit généralement des données caractérisées par une moyenne de 15 et un écart-type de 3.5. Le deuxième, tout différent, génère des données ayant

une moyenne de 75 et un écart-type de 10.5. Toutes autres choses égales par ailleurs, quel test convient le mieux pour évaluer la mémoire à long terme ? On pourrait être tenté d'affirmer que le deuxième test est meilleur dans la mesure où l'on veut une mesure qui tolère une variabilité suffisante pour permettre d'étudier les différences entre les personnes et que le deuxième test donne lieu au plus grand écart-type. N'oublions cependant pas que les deux tests diffèrent considérablement de par leur moyenne, et cette différence doit être prise en compte.

EXEMPLE

LIEN ÉTROIT ENTRE ÉCART-TYPE ET GRANDEUR DE LA MOYENNE



Si l'on réfléchit un moment au fait que l'écart-type se base sur les écarts par rapport à la moyenne, il semble logique de dire qu'il est plus facile de s'écarter sensiblement d'une grande moyenne que d'une petite moyenne. Par exemple, si l'on évalue l'efficacité pédagogique sur une échelle en 7 points ayant une moyenne de 3, il serait impossible d'obtenir un écart supérieur à 4. Par contre, si l'échelle comprenait 70 points pour une moyenne de 30, des écarts de 10 ou 20 ne seraient pas rares. D'une manière ou d'une autre, il faut tenir compte du fait qu'il est davantage possible d'obtenir des écarts importants dans le deuxième cas lorsqu'on compare la variabilité de ces deux mesures. En d'autres termes, lorsqu'on examine l'écart-type, il faut également garder à l'esprit la grandeur de la moyenne.

Coefficient de variation (CV)

Le moyen le plus simple de comparer des écarts-types pour des mesures présentant des moyennes très différentes est de diviser l'écart-type par la grandeur de la moyenne. C'est ce qu'on fait avec le **coefficient de variation (CV)**.¹² On définira simplement ce coefficient comme étant l'écart-type divisé par la moyenne.

$$CV = \frac{\text{Écart-type}}{\text{Moyenne}} = \frac{s_X}{\bar{X}} \times 100$$

(Nous multiplions par 100 pour exprimer le résultat en pourcentage.) Revenons-en à notre exemple de tâche de mémorisation. Pour la première mesure, $CV = (3.5/15) \times 100 = 23.3$. Pour la deuxième mesure, $CV = (10.5/75) \times 100 = 14$. Dans ce cas, le coefficient de variation de la première mesure est presque deux fois plus grand que celui de la seconde. Si je pouvais être sûr que le coefficient de variation supérieur de la première mesure n'est pas uniquement imputable à une imprécision des mesures, j'aurais tendance à préférer la première mesure à la deuxième.

Lorsqu'on utilise le coefficient de variation, il est important de garder à l'esprit la nature de la variable mesurée. Si l'échelle est arbitraire, on ne souhaitera peut-être pas trop faire confiance au coefficient. Mais on ne souhaitera peut-être pas non plus trop faire confiance à la variance. Sur ce point, il est très utile de faire preuve d'un peu de bon sens.

12. Je voudrais remercier Andrew Gilpin (communication personnelle, 1990) qui m'a rappelé l'utilité du coefficient de variation. Il s'agit d'une statistique importante que l'on a trop souvent tendance à négliger.

EXEMPLE

COEFFICIENT DE VARIATION ET ÉCART-TYPE



Pour prendre un deuxième exemple, Katz, Lautenschlager, Blackburn et Harris (1990) ont demandé à des étudiants de répondre à un ensemble de questions à choix multiples extraites du Scholastic Aptitude Test¹³ (SAT), un examen administré au niveau national aux États-Unis. Un groupe a lu les passages utiles avant de répondre aux questions. Un autre groupe a répondu aux questions sans avoir lu le passage sur lequel elles se basaient, ce qui revient à répondre à un QCM sur l'histoire de la Mongolie sans avoir suivi le cours. Les données sont les suivantes :

	Passage lu	Passage non lu
Moyenne	69.6	46.6
Écart-type	10.6	6.8
CV	15.0	14.6

Le rapport des deux écarts-types est égal à $10.6/6.8 = 1.56$, ce qui signifie que l'écart-type du groupe ayant lu le passage est supérieur de plus de 50 % à celui de l'autre groupe. Par contre, les coefficients de variation sont pratiquement les mêmes pour les deux groupes, ce qui porte à croire que la différence de variabilité observée entre les groupes peut s'expliquer par les scores plus élevés dans le premier groupe (Entre parenthèses, si les étudiants avaient répondu au hasard, on aurait constaté une moyenne de 20 avec un écart-type de 4. Même sans avoir lu le passage, les étudiants obtiennent des résultats largement supérieurs au hasard parce qu'ils parviennent à deviner intelligemment).

8.10. Estimateurs non biaisés

Comme cela a été souligné plus haut, nous utiliserons généralement des mesures telles que la moyenne et la variance dans le but de les employer comme *estimations* des valeurs correspondantes dans les populations. Les caractéristiques des échantillons sont appelées statistiques et désignées par des lettres romaines (p. ex. $\bar{X}$). Les caractéristiques des populations sont appelées paramètres et désignées par des lettres grecques. Ainsi, la moyenne de la population est symbolisée par la lettre μ (mu). Donc, en général, nous utilisons les statistiques comme des estimations de paramètres.

Si l'obtention d'une statistique a pour objectif de l'utiliser comme estimateur d'un paramètre, on ne devrait pas s'étonner de voir le choix d'une statistique (et même la définition qu'on lui donne) se baser en partie sur l'efficacité de cette statistique en tant qu'estimateur du paramètre en question. En fait, on préfère généralement la moyenne aux autres mesures de tendance centrale parce qu'elle constitue un bon estimateur de μ . Si la variance de l'échantillon (s^2) se définit comme telle, avec $(N - 1)$ dans le dénominateur, c'est en particulier en raison des nombreux avantages constatés lorsque s^2 est utilisé pour estimer la variance de la population (σ^2).

Les statisticiens définissent plusieurs propriétés différentes des estimateurs (l'exhaustivité, l'efficacité, la robustesse et le biais). Les trois premières ne sont pas particulièrement importantes pour nous dans ce livre, je ne les aborderai donc pas. Mais la distinction entre estimateurs biaisés et **estimateurs non biaisés** est importante : nous recherchons généralement (mais pas toujours) des estimateurs non biaisés.

Estimateurs non biaisés

Supposons que nous ayons une population dont nous connaissons la moyenne (μ), par exemple la taille de tous les joueurs du championnat de basket. Si nous

13. Ce test est maintenant connu sous le nom de SAT, ou, plus récemment, de SAT-I.

Valeur attendue

devions prélever un échantillon de cette population et calculer la moyenne de l'échantillon ($\bar{X}_1$), nous nous attendrions à ce que $\bar{X}_1$ soit assez proche de μ , en particulier si N est grand, puisqu'il s'agit d'un estimateur de μ . Ainsi, si la taille moyenne de cette population est égale à 2 mètres ($\mu = 2$ m), nous nous attendrions à ce qu'un échantillon de par exemple 10 joueurs ait également une taille moyenne de 2 mètres, même si elle ne serait sans doute pas exactement égale à 2 mètres. (Nous pouvons écrire que $\bar{X}_1 \approx 2$, le symbole $\approx$ signifiant « est approximativement égal à »). Supposons à présent que nous prélevions un autre échantillon et obtenions sa moyenne ($\bar{X}_2$). (L'indice est utilisé pour différencier les moyennes des échantillons successifs. Ainsi, la moyenne du 43^{ième} échantillon, s'il y en avait autant, serait notée $\bar{X}_{43}$). Il est probable que cette moyenne serait elle aussi assez proche de μ , mais nous ne nous attendrions pas à ce qu'elle soit exactement égale à μ ou à $\bar{X}_1$. Si nous devions poursuivre cette procédure et prélever des moyennes d'échantillon à l'infini, nous nous rendrions compte que la moyenne des moyennes des échantillons serait précisément égale à μ . Nous disons donc que la **valeur attendue** (c'est-à-dire la tendance centrale basée sur de très nombreux échantillons) de la moyenne de l'échantillon est égale à μ , la moyenne de la population qu'elle estime. Un estimateur dont la valeur attendue est égale au paramètre à estimer est appelé estimateur non biaisé; il est très important qu'une statistique présente cette propriété. La moyenne de l'échantillon et la variance de l'échantillon sont toutes deux des estimateurs non biaisés de leurs paramètres correspondants. Nous utilisons $N - 1$ comme dénominateur de la formule de la variance de l'échantillon précisément parce que nous voulons générer une estimation non biaisée.

8.11. La variance de l'échantillon en tant qu'estimateur de la variance de la population

La variance de l'échantillon illustre parfaitement ce qui a été dit à propos de l'absence de biais. Vous vous souvenez peut-être que j'ai substitué $N - 1$ à N comme diviseur dans le calcul de la variance et de l'écart-type. Il est temps d'en expliquer la raison. (Il se peut que vous ayez été parfaitement d'accord avec moi pour diviser par $N - 1$, mais je reçois un grand nombre de questions à ce propos et je suppose donc que vous ressentirez vous aussi le besoin de lire l'explication – ou alors passez au point suivant.)

Il y a plusieurs façons d'expliquer pourquoi, dans le cas de la variance de l'échantillon, le dénominateur doit être $N - 1$. La plus simple est peut-être liée à ce que nous avons dit concernant la variance de l'échantillon (s^2) en tant qu'estimateur non biaisé de la variance de la population (σ^2). Supposons pour le moment que nous ayons un nombre infini d'échantillons (chacun contenant N observations) d'une population et que nous en connaissions la variance. Supposons ensuite que nous fassions la bêtise de calculer la variance de chaque échantillon par la formule

$$\frac{\sum (X - \bar{X})^2}{N}$$

(Remarquez le dénominateur). Si nous prenons la moyenne des variances de ces échantillons, nous obtenons

$$\text{Moyenne } \frac{\sum (X - \bar{X})^2}{N} = E \left[\frac{\sum (X - \bar{X})^2}{N} \right] = \frac{(N-1)\sigma^2}{N}$$

où $E[\]$ se lit « la valeur attendue de (ce qui est entre parenthèses) ». Donc la valeur moyenne de $\sum (X - \bar{X})^2 / N$ n'est pas σ^2 . Il s'agit d'un estimateur biaisé.

La discussion qui précède revient plus ou moins à dire que nous divisons par $N-1$ parce que cela marche. Mais pourquoi cela marche-t-il ? Pour expliquer cela, nous devons commencer par considérer les **degrés de liberté (dl)**. Supposez que vous ayez devant vous les trois nombres 6, 8 et 10. Leur moyenne est égale à 8. On vous dit à présent que vous pouvez changer n'importe le(s)quel(s) de ces nombres pour autant que la moyenne reste égale à 8. Combien de nombres pouvez-vous changer ? Si vous les changez tous les trois à l'aveuglette, il y a très peu de chances que la moyenne soit encore égale à 8. Vous ne pouvez changer librement que deux de ces nombres si vous voulez que la moyenne reste constante. Par exemple, si vous changez le 6 en un 7 et le 10 en un 13, le nombre restant est fixé ; cela doit être un 4 si vous voulez que la moyenne soit égale à 8. Si l'on vous donnait les mêmes instructions pour un total de 50 nombres, vous ne pourriez en changer que 49 ; le 50^{ième} serait fixé.

Degrés de liberté (dl)

Revenons-en à présent aux formules de la variance des populations et des échantillons et voyons pourquoi nous avons perdu un degré de liberté en calculant la variance des échantillons

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \quad s^2 = \frac{\sum (X - \bar{X})^2}{N-1}$$

Dans le cas de σ^2 , μ est connu et ne doit pas être estimé à partir des données. Ainsi, aucun dl n'est perdu et le dénominateur est N . Par contre, dans le cas de s^2 , μ n'est pas connu et doit être estimé à partir de la moyenne de l'échantillon ($\bar{X}$). Une fois que nous avons estimé μ à partir de $\bar{X}$, nous l'avons fixé dans le but d'estimer la variabilité. Nous perdons dès lors le degré de liberté dont nous venons de parler, et il ne reste que $N-1$ dl ($N-1$ scores peuvent varier). Nous perdons ce degré de liberté *chaque fois* que nous estimons une moyenne. Il en résulte que le dénominateur (le nombre de scores servant de base à notre estimation) doit refléter cette restriction. Il représente le nombre de données indépendantes.

9. Diagrammes en boîte : représentations graphiques de dispersions et scores extrêmes

Nous avons vu précédemment que les représentations en tiges et feuilles permettent de représenter les données de plusieurs façons intéressantes simultanément. Elles combinent les données en un schéma très proche de l'histogramme, tout en conservant les valeurs particulières des observations. Outre la représentation en tiges et feuilles, John Tukey a élaboré d'autres moyens d'examiner les données. L'un de ces moyens souligne davantage la dispersion des données : il s'agit de la méthode appelée diagramme en boîte, ou **boxplot**, ou parfois **boîte à moustaches**.

Boxplot ou boîte à moustaches

EXEMPLE

DURÉE D'HOSPITALISATION DE BÉBÉS NÉS À TERME



Les données et leur représentation en tiges et feuilles illustrée, plus bas, dans le tableau 2.7 proviennent de bébés présentant un poids normal ou faible à la naissance et ayant participé à une étude sur le développement des nourrissons à l'Université du Vermont; il s'agit des données préliminaires sur la durée d'hospitalisation de 38 bébés dont le poids à la naissance était normal. Les données relatives à trois bébés manquent pour cette variable particulière; elles sont représentées par des astérisques (*) (Elles ont été incluses pour souligner qu'il ne faut pas se contenter d'ignorer les données manquantes). Comme les données varient de 1 à 10, à deux exceptions près, toutes les feuilles sont égales à 0. Les 0 ne font qu'occuper un certain espace pour permettre de générer une distribution semblable à un histogramme. L'examen des données telles qu'elles sont illustrées dans la représentation en tiges et feuilles révèle une asymétrie positive avec pour médiane une durée d'hospitalisation de trois jours. Vers le bas de la tige, on voit l'entrée HI et les valeurs 20 et 33. Il s'agit de valeurs extrêmes, ou aberrantes; elles sont ainsi mises en évidence dans le but de souligner leur existence. Le diagramme en boîte doit notamment traiter la question de savoir si ces valeurs sont suffisamment élevées pour éveiller des soupçons. La dernière ligne de la représentation en tiges et feuilles indique le nombre d'observations manquantes.

TABLEAU 2.7. Données et représentation en tiges et feuilles de la durée d'hospitalisation de bébés nés à terme (en jours)

Données			Représentation en tiges et feuilles	
2	1	7	1	000
1	33	2	2	000000000
2	3	4	3	00000000000
3	*	4	4	0000000
3	3	10	5	00
9	2	5	6	0
4	3	3	7	0
20	6	2	8	
4	5	2	9	0
1	*	*	10	0
3	3	4	HI	20, 33
2	3	4	Données manquantes = 3	
3	2	3		
2	4			

À l'origine, Tukey a défini les diagrammes en boîte en termes des mesures spéciales qu'il a conçues. La plupart des gens dessinent maintenant les diagrammes en boîte en utilisant des mesures plus traditionnelles et c'est cette approche que j'adopte dans cette édition.

Nous avons défini la position médiane d'un ensemble de N scores par $(N + 1)/2$. Lorsque la position médiane est un nombre entier, comme ce sera le cas lorsque N est impair, la médiane est tout simplement la valeur qui occupe cette position lorsque les données sont classées dans l'ordre. Lorsque la position médiane est un nombre fractionnaire (par exemple lorsque N est pair), la médiane est la tendance centrale des deux valeurs situées de part et d'autre de cette position. Pour les données présentées dans le tableau 2.7, la position médiane est égale à $(38 + 1)/2 = 19.5$, et la médiane est égale à 3. Pour construire un diagramme en boîte, nous allons également prendre le premier et le troisième quartile, tels que définis précédemment. La façon la plus facile de faire cela est de définir la position du quartile comme

$$\text{Position des quartiles} = \frac{\text{Position médiane} + 1}{2}$$

Si la position médiane est une valeur fractionnaire, les chiffres à droite de la virgule doivent être enlevés du numérateur lorsque nous calculons la position des quartiles. La position des quartiles est aux quartiles ce que la position médiane est à la médiane. Elle nous dit à quel endroit, dans une série ordonnée, se trouvent les valeurs des quartiles¹⁴. Pour les données relatives à l'hospitalisation, la position des quartiles est égale à $(19 + 1)/2 = 10$. Les quartiles seront donc les dixièmes scores à partir du bas et à partir du haut. Ces valeurs sont respectivement égales à 2 et 4. Pour les ensembles de données n'incluant aucun ex aequo, ou pour les grands échantillons, les quartiles grouperont les 50 % situés au milieu des scores.

Pour terminer la liste des concepts nécessaires à la compréhension des diagrammes en boîte, nous devons encore considérer trois termes supplémentaires : l'écart interquartile, les barrières intérieures et les valeurs adjacentes. Comme vu précédemment, l'écart interquartile est tout simplement l'écart entre les premier et troisième quartiles. Pour nos données, l'écart interquartile est égal à $4 - 2 = 2$. Une barrière intérieure est définie par Tukey comme un point qui se situe à 1.5 fois l'écart interquartile au-dessus ou en dessous de la charnière en question. Comme l'écart interquartile est égal à 2 pour nos données, la barrière intérieure se trouve $2 \times 1.5 = 3$ points plus loin que les quartiles. Comme nos quartiles sont égaux aux valeurs 2 et 4, les barrières intérieures seront à $2 - 3 = -1$ et $4 + 3 = 7$. Les valeurs adjacentes sont les valeurs, comprises dans les données, qui ne sont pas plus extrêmes (plus éloignées de la médiane) que les barrières intérieures. Comme notre plus petite valeur est égale à 1, il s'agit là de la valeur la plus proche de la barrière intérieure inférieure et elle est donc la valeur adjacente la moins élevée. La barrière intérieure supérieure est égale à 7, et comme nos données comprennent un 7, il s'agira de la valeur adjacente la plus élevée. Les calculs relatifs à tous les termes que nous venons de définir sont illustrés dans le tableau 2.8.

EXEMPLE**CONSTRUCTION PAS À PAS D'UN DIAGRAMME EN BOÎTE**

Les barrières intérieures et les valeurs adjacentes peuvent créer une certaine confusion. Imaginons un troupeau de vaches dispersées dans une prairie. (J'ai passé la majeure partie de ma vie dans le Vermont, c'est pourquoi l'exemple des vaches me vient naturellement à l'esprit). La barrière entourant la prairie représente la barrière intérieure du diagramme en forme de boîte. Les vaches les plus proches de cette barrière, mais toujours à l'intérieur, constituent les valeurs adjacentes. Ne vous préoccupez pas des vaches qui se sont échappées de la prairie et se promènent sur la route. À ce stade, elles ne sont pas comprises dans les calculs. (Ce seront des valeurs extrêmes.)

À présent, nous sommes prêts à tracer le diagramme en boîte. Nous commençons par tracer et par graduer une échelle qui couvre l'étendue totale des valeurs obtenues. Cela a été fait au bas du tableau 2.7. Nous traçons alors une boîte rectangulaire de Q_1 à Q_3 , en symbolisant la position de la médiane par une ligne verticale. Ensuite, nous traçons, à partir des quartiles, des lignes (appelées moustaches) qui rejoignent les valeurs adjacentes. Pour terminer, nous représentons la position de tous les points plus extrêmes que les valeurs adjacentes (ici par des astérisques).

Le tableau 2.8 nous montre plusieurs points importants. Tout d'abord, la portion centrale de la distribution est relativement symétrique. Ceci est mis en évidence par le fait que la médiane se situe au centre de la boîte et c'était déjà apparent dans la représentation en tiges et feuilles. Nous voyons également que la distribution est positivement asymétrique, parce que la moustache de droite est sensiblement plus longue que celle de gauche. Cela se voyait aussi sur la représentation en tiges et feuilles, bien que beaucoup moins clairement. Enfin, nous notons quatre

14. Tukey parle de « hinges » (charnières) plutôt que de quartiles, mais on ne perd rien à les considérer comme des quartiles.

valeurs extrêmes, une valeur extrême se définissant ici comme une valeur plus extrême que les moustaches (et dès lors plus extrême que les valeurs adjacentes). La représentation en tiges et feuilles ne montrait pas la position des valeurs extrêmes de façon aussi évidente que le diagramme en boîte.

TABEAU 2.8. Calcul et diagrammes en boîte à partir des données du tableau 2.7

$$\text{Position médiane} = (N + 1)/2 = (38 + 1)/2 = 19.5$$

$$\text{Médiane} = 3$$

$$\text{Position des quartiles} = (\text{position médiane}^\dagger + 1)/2 = (19 + 1)/2 = 10$$

$$Q_1 = 10^{\text{ième}} \text{ score le plus bas} = 2$$

$$Q_3 = 10^{\text{ième}} \text{ score le plus élevé} = 4$$

$$\text{Écart interquartile} = 4 - 2 = 2$$

$$\text{Écart interquartile} \times 1.5 = 2 \times 1.5 = 3$$

$$\text{Barrière intérieure inférieure} = Q_1 - 1.5 (\text{écart interquartile}) = 2 - 3 = -1$$

$$\text{Barrière intérieure supérieure} = Q_3 + 1.5 (\text{dispersion des quartiles}) = 4 + 3 = 7$$

$$\text{Valeur adjacente inférieure} = \text{plus petite valeur} \geq \text{barrière inférieure} = 1$$

$$\text{Valeur adjacente supérieure} = \text{plus grande valeur} \leq \text{barrière supérieure} = 7$$



[†] Ne pas prendre en compte les valeurs fractionnaires.

Les valeurs extrêmes méritent une attention toute particulière. Une valeur extrême pourrait représenter une erreur de mesure, d'enregistrement des données ou d'encodage des données (on parle dans ce cas de valeur aberrante), tout comme il pourrait s'agir d'une valeur légitime qui est tout simplement – et par hasard – extrême. Par exemple, nos données représentent la durée d'hospitalisation, et un bébé né à terme pourrait présenter, à la naissance, une imperfection physique nécessitant une hospitalisation prolongée. Comme il s'agissait de données réelles, il s'est avéré possible de consulter les registres hospitaliers et d'examiner avec plus d'attention ces quatre cas extrêmes. L'examen a montré que les deux scores les plus extrêmes étaient dus à des erreurs d'encodage et pouvaient être directement rectifiés. Les deux autres scores extrêmes s'expliquaient par l'existence de problèmes physiques chez les bébés. Ici, le directeur de projet a dû décider si les problèmes étaient suffisamment graves pour justifier d'exclure les bébés de l'étude (Tous deux ont finalement été conservés). Les deux valeurs rectifiées étaient respectivement 3 et 5 au lieu de 33 et 20; la figure 2.14 donne le diagramme en boîte des données corrigées. Ce diagramme est identique à celui du tableau 2.8 à l'exception de l'écartement et des deux plus grandes valeurs (Vérifiez par vous-même que les données rectifiées donneraient effectivement ce diagramme).

Les diagrammes en boîte constituent des outils extrêmement précieux pour examiner les données du point de vue de leur dispersion. Je les trouve particulièrement utiles pour détecter les erreurs au sein des données et pour mettre en lumière des problèmes potentiels avant la réalisation d'autres analyses. Les diagrammes en boîte sont souvent présentés, dans la suite de ce livre, comme une aide visuelle pour l'analyse des données.

Les diagrammes en boîte s'avèrent réellement utiles lorsqu'on compare plusieurs groupes. Nous allons utiliser l'exemple avec lequel nous avons ouvert le chapitre,



FIGURE 2.14. Diagramme en boîte : données rectifiées du tableau 2.7

FOCUS SUR...

DIFFÉRENCES ENTRE DIAGRAMMES EN BOÎTE SELON LE LOGICIEL UTILISÉ POUR LES GÉNÉRER



Prenez garde : la définition des divers éléments de ces diagrammes peut varier en fonction des différents programmes informatiques consacrés aux statistiques (Voir Frigge, Hoaglin et Iglewicz (1989) pour un examen détaillé de ce problème). Deux programmes différents pourraient générer des diagrammes en boîte légèrement différents pour le même ensemble de

données. Ils pourraient même identifier des valeurs extrêmes différentes. Néanmoins, les diagrammes en boîte sont normalement utilisés comme des outils heuristiques informels, et les différences subtiles de définition ne posent que rarement, voire jamais, de problème. Si j'évoque ces divergences potentielles, c'est uniquement pour expliquer pourquoi les analyses

que vous effectuez d'après les données de ce livre pourraient donner des résultats légèrement différents selon le programme informatique que vous utilisez. (Un codage simple pour créer des diagrammes en boîte dans R est accessible sur le site Web sous le nom de Boxplots.R).

lorsque nous avons enregistré le temps de réaction des réponses à la question de savoir si un chiffre spécifique était présent sur l'écran précédent et, ce, en fonction du nombre de stimuli présentés sur cet écran. Les diagrammes en boîte de la figure 2.15, obtenues avec SPSS, montrent les temps de réaction pour les cas où le stimulus était effectivement présent, séparément selon le nombre de stimuli dans l'ensemble initial. Les valeurs extrêmes sont indiquées par leur numéro d'identification qui est ici le même que le numéro de l'essai au cours duquel le stimulus a été

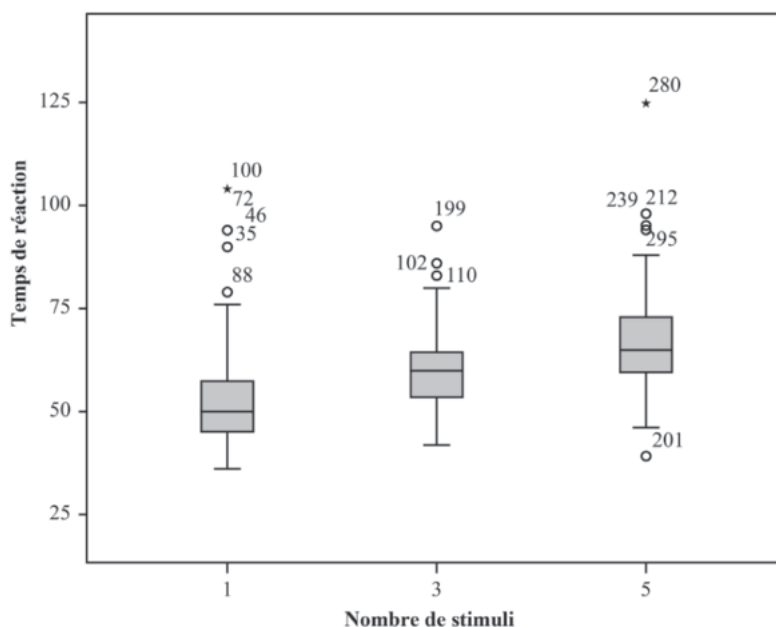
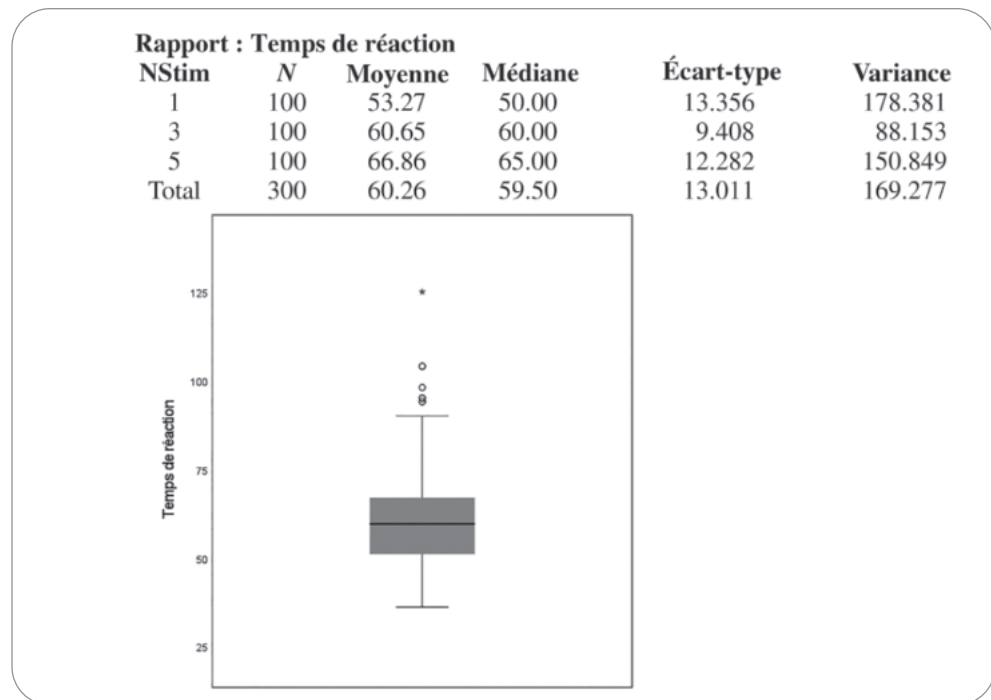


FIGURE 2.15. Diagramme en boîte pour les temps de réaction en fonction du nombre de stimuli dans l'ensemble initial

ENCART 2.1. Analyse SPSS des données de temps de réaction

présenté. La conclusion la plus claire que l'on peut tirer sur base de cette figure est qu'une augmentation du nombre de stimuli dans l'ensemble initial fait en sorte que le temps de réaction augmente, tout comme la dispersion. On peut également voir, en se limitant à l'analyse des boîtes et des moustaches, que les distributions sont relativement symétriques (les boîtes sont à peu près centrées autour de la médiane et il y a peu de valeurs extrêmes, par contre, ces dernières correspondent quasiment toutes à des temps de réaction élevés).

10. Obtention de mesures de dispersion en utilisant SPSS

Nous pouvons également utiliser SPSS pour calculer les mesures de tendance centrale et de dispersion, comme le montre l'encart 2.1, qui est basé sur nos données de l'expérience sur les temps de réaction. J'ai utilisé le menu *Analyse/Comparer les moyennes/Moyennes* car je voulais obtenir les statistiques descriptives séparément pour chaque niveau de NStim (le nombre de stimuli présentés). Remarquez que vous avez également ces statistiques pour les trois groupes. Le menu *Graphiques/Boîtes de dialogue ancienne version/Boîte à moustaches* a produit le diagramme en boîte repris à l'encart 2.1. Comme vous avez déjà vu le diagramme en boîte ventilé par NStim à la figure 2.13, je ne présente ici que les données combinées. On remarquera que les valeurs extrêmes ressortent clairement.

11. Percentiles, quartiles et déciles

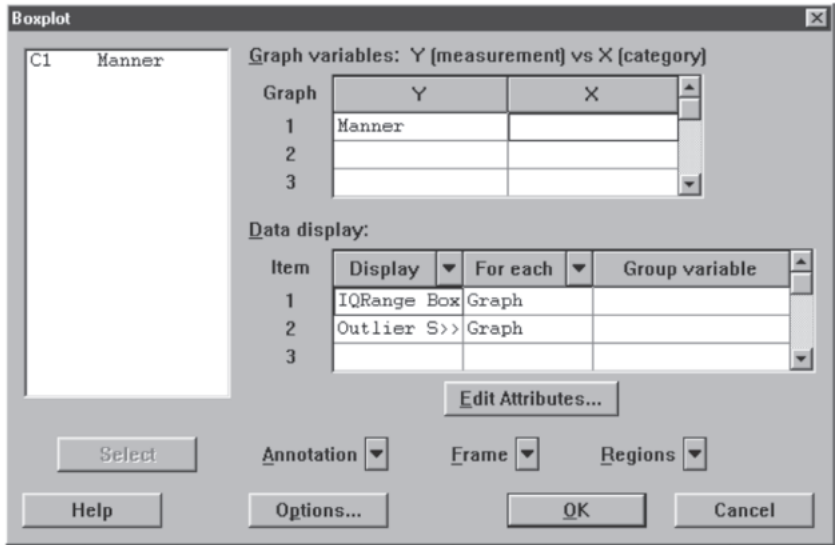
Une distribution possède bien d'autres propriétés que sa position et sa dispersion. Nous en avons brièvement évoqué une lorsque nous avons abordé les diagrammes

en boîte pour lesquels nous avons eu recours aux quartiles : les **quartiles** ont été définis comme les valeurs qui divisent la distribution en quarts. Ainsi, le premier quartile isole les 25 % inférieurs, le deuxième quartile isole les 50 % inférieurs et le troisième quartile isole les 75 % inférieurs (Notons que le deuxième quartile correspond également à la médiane). Ces percentiles ont été montrés de façon très claire sur le graphique de croissance à la figure 2.12. Si l'on souhaite examiner des gradations plus précises de la distribution, on peut s'attacher aux **déciles**, qui divisent la distribution en dixièmes : le premier décile isole les 10 % inférieurs, le deuxième décile isole les 20 % inférieurs, etc. Enfin, la plupart d'entre vous connaissent les **percentiles**, c'est-à-dire les valeurs qui divisent la distribution en centièmes. Ainsi, le 81^{ième} percentile est le point de la distribution en dessous duquel se situent 81 % des scores.

Quartile

Décile

Percentile



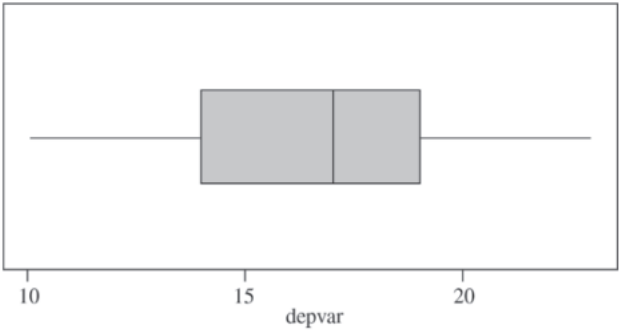
ENCART 2.2. Analyse Minitab des données à propos du nombre de tics

Statistiques descriptives

Variable	N	Mean	Median	TrMean	StDev	SE Mean
dépvvar	15	16.400	17.000	16.385	3.562	0.920

Variable	Minimum	Maximum	Q1	Q3
dépvvar	10.000	23.000	14.000	19.000

Boxplot of depvar



Quantile

Les quartiles, déciles et percentiles sont les trois exemples les plus courants d'une catégorie générale de statistiques connue sous le nom générique des **quantiles**. Ce livre ne s'étendra pas sur les quantiles, mais ceux-ci font généralement l'objet d'analyses détaillées dans des textes d'introduction (p. ex. Howell, 2008a). Les quantiles sont également essentiels dans diverses techniques d'analyse exploratoire des données préconisées par Tukey.

12. L'effet de transformations linéaires sur les données

Transformations linéaires

Nous souhaitons fréquemment transformer des données d'une façon ou d'une autre. Par exemple, nous pourrions souhaiter convertir des pouces en centimètres, des degrés Fahrenheit en degrés Celsius, des cotes d'examens basées sur 79 questions en cotes basées sur une échelle en 100 points, des revenus de quatre ou cinq chiffres en revenus d'un ou deux chiffres, etc. Heureusement, toutes ces transformations sont comprises dans un ensemble appelé **transformations linéaires**, dans lesquelles nous multiplions chaque X par une constante (peut-être 1) et y ajoutons une constante (peut-être 0).

$$X_{\text{nouv}} = bX_{\text{anc}} + a$$

où a et b sont nos constantes (Les transformations utilisant exposants, logarithmes, fonctions trigonométriques, etc., sont classées comme des transformations non-linéaires). Comme exemple de transformation linéaire, citons la formule de conversion des degrés Celsius en degrés Fahrenheit :

$$F = 9/5(C) + 32$$

Tant que nous nous limitons aux transformations linéaires, un ensemble de règles simples définit la moyenne et la variance des observations sur une nouvelle échelle en fonction de leur moyenne et de leur variance sur l'ancienne :

1. Additionner (ou soustraire) une constante à (ou d')un ensemble de données mène à additionner (ou à soustraire) cette même constante à (ou de) la moyenne :

$$\text{Pour } X_{\text{nouv}} = X_{\text{anc}} \pm a: \quad \bar{X}_{\text{nouv}} = \bar{X}_{\text{anc}} \pm a$$

2. Multiplier (ou diviser) un ensemble de données par une constante mène à multiplier (ou à diviser) la moyenne par cette même constante :

$$\text{Pour } X_{\text{nouv}} = bX_{\text{anc}}: \quad \bar{X}_{\text{nouv}} = b\bar{X}_{\text{anc}}$$

$$\text{Pour } X_{\text{nouv}} = X_{\text{anc}}/b: \quad \bar{X}_{\text{nouv}} = \bar{X}_{\text{anc}}/b$$

3. Additionner ou soustraire une constante à (ou d')un ensemble de scores ne modifie ni la variance ni l'écart-type :

$$\text{Pour } X_{\text{nouv}} = X_{\text{anc}} \pm a: \quad s_{\text{nouv}}^2 = s_{\text{anc}}^2$$

4. Multiplier (ou diviser) un ensemble de scores par une constante mène à multiplier (ou à diviser) la variance par le carré de la constante et l'écart-type par la constante :

$$\text{Pour } X_{\text{nouv}} = bX_{\text{anc}} : s_{\text{nouv}}^2 = b^2 s_{\text{anc}}^2 \text{ et } s_{\text{nouv}} = b s_{\text{anc}}$$

$$\text{Pour } X_{\text{nouv}} = X_{\text{anc}}/b : s_{\text{nouv}}^2 = s_{\text{anc}}^2/b^2 \text{ et } s_{\text{nouv}} = s_{\text{anc}}/b$$

Ces règles sont illustrées par l'exemple suivant. Dans chacun des cas, la constante utilisée est 3.

Addition d'une constante :

Anciennes données	$\bar{X}$	s^2	s	Nouvelles données	$\bar{X}$	s^2	s
4, 8, 12	8	16	4	7, 11, 15	11	16	4

Multiplication par une constante :

Anciennes données	$\bar{X}$	s^2	s	Nouvelles données	$\bar{X}$	s^2	s
4, 8, 12	8	16	4	12, 24, 36	24	144	12

12.1. Le centrage

Il est de plus en plus fréquent de voir le terme « **centrage** » utilisé en lien avec les données. L'idée de base est très simple. Il s'agit de soustraire la moyenne de l'échantillon de toutes les observations. Cela signifie que la nouvelle moyenne sera de 0.00, mais que l'écart-type et la variance ne seront pas affectés. Nous verrons plusieurs exemples de ce phénomène, mais je vais déjà en glisser un rapide ici. Supposons que nous ayons écrit une équation pour prédire l'attrait sexuel à partir de la taille. Cette équation aurait ce que l'on appelle une ordonnée à l'origine, c'est-à-dire la prédiction de l'attractivité lorsque la taille est égale à 0. Mais bien sûr, personne ne mesure 0 centimètre, si bien que cette statistique n'a pas de raison d'être. Mais si nous centrons les hauteurs en soustrayant la taille moyenne de nos données de taille, l'ordonnée à l'origine, qui correspond à la valeur de l'attrait lorsque la taille est égale à 0, représente maintenant l'attrait prédit pour une personne dont la taille est égale à la moyenne, ceci étant une statistique beaucoup plus informative. Le centrage est utilisé dans un certain nombre de situations, mais cet exemple illustre bien l'objectif général.

Centrage

12.2. Le reflet comme transformation

Une transformation aussi courante qu'utile consiste à inverser l'ordre d'une échelle. Par exemple, supposons que nous demandions à des personnes d'indiquer sur une échelle en 5 points leur degré d'accord avec chacune des différentes propositions avancées. Pour empêcher les participants de cocher le même point de l'échelle tout au long de la page sans réfléchir, nous avons formulé la moitié des questions de manière positive et l'autre moitié de manière négative. Ainsi, étant donné une échelle en 5 points où le 5 représente « tout à fait d'accord » et le 1 « pas du tout d'accord », un 4 pour la proposition « Je suis généreux » serait comparable à un 2 pour « Je suis

Reflét

égoïste». Si nous voulons que les scores soient comparables, nous devons transformer les propositions négatives (par exemple) et convertir un 5 en un 1, un 4 en un 2, etc. Cette procédure est appelée **reflét** (par analogie avec le miroir) et s'opère très simplement via une transformation linéaire. Il suffit d'écrire $X_{\text{nouv}} = 6 - X_{\text{anc}}$. La constante (6) est simplement la plus grande valeur de l'échelle plus 1. Il est évident que lorsque nous « reflétons » une échelle, nous en « reflétons » également la moyenne, mais la transformation n'a aucun effet sur la variance ou sur l'écart-type. Ce principe s'applique en vertu de la règle 3 de la liste qui précède.

12.3. La standardisation

Score standardisé

Standardisation

Une transformation linéaire courante utilisée pour transformer des données implique de soustraire la moyenne de chaque observation. Les observations ainsi transformées ne sont rien d'autre que des écarts, et la transformation en elle-même comprend le centrage puisque la moyenne est centrée à 0. Le centrage est utilisé très couramment dans la régression, dont nous parlerons plus loin dans ce livre. Une transformation encore plus courante consiste à créer des scores d'écart et à diviser ensuite les écarts par l'écart-type. Ces écarts sont alors appelés **scores standardisés**, ou scores centrés réduits, et le processus est désigné sous le nom de **standardisation**. Fondamentalement, les scores standardisés sont simplement des observations transformées mesurées en unités d'écart-type. Par exemple, un score standardisé de 0.75 est un score qui se situe à 0.75 écart-type au-dessus de la moyenne ; un score standardisé de -0.43 est un score qui se situe à 0.43 écart-type en dessous de la moyenne. Nous reviendrons très largement sur les scores standardisés lorsque nous aborderons la distribution normale au chapitre 3. Si j'introduis le sujet ici de façon spécifique, c'est pour montrer que nous pouvons calculer des scores standardisés indépendamment du fait que nous ayons ou non une distribution normale (définie au chapitre 3). Les gens associent souvent les scores standardisés à une distribution normale, mais c'est loin d'être obligatoire. La standardisation est une transformation linéaire simple des données brutes et, en tant que telle, ne modifie pas la forme de la distribution.

12.4. Transformations non-linéaires

Alors que les transformations linéaires sont habituellement utilisées pour convertir les données dans un format plus sensé – comme lorsqu'il s'agit d'exprimer les données sur une échelle de 0 à 100, de les mettre sous forme standardisée, et ainsi de suite, les transformations non-linéaires sont généralement exploitées pour modifier la forme d'une distribution. Comme nous l'avons vu, les transformations linéaires ne changent pas la forme sous-jacente d'une distribution. Les transformations non-linéaires, par contre, peuvent faire en sorte qu'une distribution asymétrique paraisse davantage symétrique et peuvent réduire l'impact des valeurs extrêmes.

Certaines transformations non linéaires sont si communes que nous ne les considérons normalement pas comme des transformations. Everitt (dans Hand, 1994) a rapporté les poids avant et après traitement de 29 filles recevant une thérapie cognitivo-comportementale pour l'anorexie. Une mesure logique serait le poids de la personne après l'intervention (Y). Une autre serait le gain de poids entre la période pré et post-intervention, mesuré par $(Y - X)$. Une troisième option serait de regarder le gain de poids en fonction du score initial. Cela correspondrait à $(Y - X)/X$. Nous

Le manuel de référence des statistiques en sciences humaines entièrement mis à jour

David C. Howell, PhD, est professeur émérite et ancien président du département de Psychologie de l'Université du Vermont. Il est co-éditeur, avec Brian Everitt, de *l'Encyclopédie des statistiques en sciences comportementales*.

Vincent Yzerbyt est professeur de Psychologie à l'Université catholique de Louvain (Belgique). Membre de l'Association for Psychological Science, il enseigne la méthodologie et les statistiques en sciences comportementales.

Yves Bestgen est professeur à l'Université catholique de Louvain et chercheur qualifié au Fonds National de la Recherche Scientifique, rattaché à l'Unité de psychologie cognitive.

Nathalie Lefèvre est docteure en Psychologie et statisticienne. Elle enseigne les statistiques à l'Université catholique de Louvain, dont elle dirige la plateforme de Support en Méthodologie et Calcul Statistique.

Aurélien Bertrand est docteure en Statistique. Elle enseigne les statistiques à l'Université catholique de Louvain et est consultante en statistique pour la plateforme de Support en Méthodologie et Calcul Statistique.

Conçu pour les étudiants non spécialisés en mathématiques, *Méthodes statistiques en sciences humaines* propose les bases fondamentales pour comprendre l'ensemble des techniques statistiques les plus utilisées en sciences sociales et dans les sciences du comportement.

Ses points forts sont la place importante accordée aux techniques descriptives ainsi que le recours fréquent à des exemples concrets permettant de souligner les liens entre les questions de recherche et les tests statistiques.

Dans cette 3^e édition entièrement mise à jour, on retrouve :

- un nouveau chapitre consacré à la méta-analyse et aux études de cas unique ;
- de nombreux nouveaux exemples avec des données chiffrées actualisées ;
- + de 270 graphiques, schémas et tableaux ainsi que 20 démos SPSS ;
- une actualisation des exercices ainsi que des ressources en ligne.

Par ses qualités, ce manuel est une référence indispensable tant pendant qu'après le cursus universitaire.

RESSOURCES NUMÉRIQUES

Pour les étudiants :

- + de 450 exercices
- + de 50 défis
- 180 flash-cards
- 18 résumés audio
- Sites pour aller plus loin

Pour les professeurs :

- 18 PowerPoints®

DANS LA MÊME COLLECTION



ISBN : 978-2-8073-0552-6



55 €

deboeck
SUPÉRIEUR