MANUELS VISUELS DE LICENCE

Cours | Exercices | Corrigés | 150 photos et schémas | Sites web

Statistiques pour psychologues

Édition : Marie-Laure Davezac-Duhem
Fabrication : Anne Pachiaudi
Composition et mise en pages : Nord Compo
Conception couverture : Pierre-André Gualino
Relecture et correction : Isabelle Chave

Nos équipes ont vérifié le contenu des sites internet mentionnés dans cet ouvrage au moment de sa réalisation et ne pourront pas être tenues pour responsables des changements de contenu intervenant après la parution du livre.

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.
Le Code de la propriété intellectuelle du 1er juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée. Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).

© Dunod, Paris, 2010, 2015
5 rue Laromiguière, 75005 Paris
www.dunod.com
ISBN 978-2-10-073817-5

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).
Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

MANUELS VISUELS DE LICENCE

Cours | Exercices | Corrigés | 150 photos et schémas | Sites web

Statistiques pour psychologues

Analyses descriptives

2e édition revue et augmentée

Jean-Marc Meunier

DUNOD

觀自
菩薩
波羅
弟子
佛光山修持
僧

Table des matières

Préface

Introduction

1 – Les concepts statistiques

I. Individu, échantillon, population 6

II. Notions de variable et de facteur 6

III. La notion de mesure 10

IV. Notion de protocole 11
1. L'emboîtement 12
2. Le croisement 13

V. Les procédures 13
1. Les types d'objets 15
2. Les objectifs des méthodes statistiques 16

2 – Les procédures applicables aux variables

I. Pourquoi recoder une variable ? 24

II. Le recodage par regroupement de modalités 26
1. Recoder une variable nominale 26
2. Recoder une variable ordinale 29
3. Recoder une variable numérique 31

III. Le recodage par transformation 35
1. Les transformations linéaires 36
2. La transformation Z 37
3. Les transformations non linéaires 38
4. L'inverse 40

3 – Les procédures applicables aux protocoles

I. Dériver un protocole d'analyse 46

II. Résumer numériquement un protocole 48

1. Indices de position 49
2. Indices de dispersion 53

III. Comparer des groupes d'observations 59

IV. Faire une distribution 61

1. Faire une distribution sur un protocole univarié 61
2. Faire une distribution sur un protocole bivarié 63
3. Un nouvel objet d'analyse 65

4 – Les procédures applicables aux distributions

I. Résumer une distribution 74

1. Résumer graphiquement une distribution 74
2. Cumuler une distribution 79
3. Résumer numériquement une distribution 81
4. Le choix des résumés numériques 89

II. Comparer les distributions de groupes d'observations 90

1. Cas d'une variable nominale 90
2. Cas d'une variable ordinale 92
3. Cas d'une variable numérique 95

III. Situer un individu dans une distribution 97

1. Situer un individu dans un décilage 98
2. Situer un individu avec un écart réduit 100
3. Situer dans une échelle normalisée 101

IV. Comparer une distribution à une distribution de référence 105

1. Comparaison à une distribution plate 106
2. Comparaison à une distribution normale 107

5 – L'analyse des protocoles bivariés

I. La notion de liaison 120

II. Étudier la liaison entre variables nominales 121

III. Étudier la liaison entre variables ordinales 128

IV. Étudier la liaison entre variables numériques 131

6 – L'utilisation des logiciels de statistiques

I. Principes de fonctionnement d'un tableur 145

1. La feuille de calcul 145
2. Les formules 146
3. L'utilisation de références 146
4. La saisie d'un protocole 147
5. La structuration des données 148

II. Recoder une variable 149

1. Le regroupement de modalités 149
2. Le regroupement en classes 151

III. L'analyse des protocoles 152

1. Dériver un protocole d'analyse 152
2. Faire une distribution 155
3. Calculer les quartiles 155
4. Calculer la moyenne et l'écart-type 157

IV. L'analyse de distribution 158

1. Faire des graphiques 158
2. Cumuler une distribution 160
3. Résumer une distribution 161
4. Normaliser une distribution 163

V. L'analyse des protocoles bivariés 164

1. Réaliser une distribution bivariée 164
2. Calculer un Φ^2 166
3. Calculer un ρ de Spearman 168
4. Calculer un *r* de Bravais-Pearson 169

Annexes

Corrigés

Crédits photographiques

Préface

Les statistiques ont envahi notre univers quotidien, et comme elles sont très largement répandues, on peut avoir l'impression qu'elles reposent sur des notions qui vont de soi. C'est pourtant loin d'être le cas. Les statistiques sont une branche des mathématiques qui est apparue assez récemment et dont le développement a été grandement favorisé par l'informatique, qui fournit à la fois des facilités de recueil et des capacités d'enregistrement et de traitement de grandes masses de données. Mais la culture statistique suit-elle le développement technique ?

Une première difficulté est que les observations recueillies et analysées ne portent pas d'ordinaire sur l'ensemble des individus et plus généralement des objets concernés par l'étude, mais seulement sur une partie d'entre eux, ce que l'on appelle un échantillon. Il apparaît paradoxal que l'on puisse, à partir d'un échantillon d'un millier de personnes, évaluer l'opinion des millions de personnes que compte un pays, sur une question ou leur jugement sur un personnage politique. La justification est difficile à comprendre, car elle repose sur le calcul des probabilités et sur un certain nombre de conditions que doivent respecter les techniques de sondage et qui permettent d'appliquer les méthodes de calcul statistique à ce domaine.

Une seconde difficulté est que les techniques statistiques sont centrées sur l'analyse de la variabilité et que la variabilité est mal appréhendée par notre jugement naturel, qui retient ce qui est vu ou senti dans le présent et ce que nous livre notre mémoire, qui est par ailleurs très sélective. De la sorte, nous ne tirons guère d'informations de la variabilité et nous ne pensons guère qu'on peut le faire. La statistique oblige donc à construire de nouveaux concepts qui peuvent paraître abstraits, dans la mesure où ils ne sont pas dérivés de l'expérience quotidienne. Tel est le cas de la notion de distribution qui est à la base du raisonnement statistique et qui exprime la répartition des observations sur une échelle qui permet de les situer. Cette notion est complexe, car il n'en existe guère d'équivalent dans la vie quotidienne ; il est pourtant nécessaire que l'étudiant en acquière une bonne compréhension.

Une difficulté complémentaire vient de ce que les résultats de l'analyse statistique peuvent heurter les conceptions quotidiennes. Par exemple, si on fait une statistique sur le nombre d'enfants d'une famille, on peut trouver que la moyenne pour un pays est 2,2 ou 1,6. Cette valeur n'a pas de sens quand on parle d'une famille : comment se représenter la signification de cette valeur décimale ? Elle n'a de sens que relativement à l'ensemble des familles d'un pays puisqu'elle se réfère à la distribution du nombre

d'enfants par famille dans cette population et en cela n'a certainement rien d'intuitif. Faute d'avoir assimilé un certain nombre de concepts de base, des représentations erronées de la statistique peuvent se développer dans le public, comme l'idée répandue que l'on peut faire dire n'importe quoi aux statistiques. Les techniques statistiques permettent certes différentes analyses, mais la pertinence de ces analyses dépend du type de données auxquelles on a affaire. Le choix n'est pas aveugle et l'on peut en confrontant les conditions de validité de chaque analyse et les propriétés de la situation étudiée déterminer quel type d'analyse est optimale.

L'ouvrage de Jean-Marc Meunier se propose de donner en même temps que les techniques d'analyse une base conceptuelle qui permet des choix judicieux et une interprétation appropriée. Il est exigeant, mais il est à la portée d'un étudiant en sciences humaines, notamment d'un étudiant en psychologie, dans la mesure où l'introduction des notions est faite de façon très progressive et à partir de beaucoup d'exemples.

Une étape importante est le passage de l'observation comportementale à la mise en forme de ce qui a été observé en vue du traitement statistique. Cette étape est expliquée de façon très détaillée dans l'ouvrage à partir de l'analyse des différents types d'observations et l'étude des différentes façons de les coder en vue des analyses ultérieures.

Le concept de distribution, si important on l'a dit, est également présenté de façon très détaillée en lien avec les différents types de mesure dont on dispose. En effet, en psychologie, les mesures numériques, qui sont les plus faciles à traiter, sont en effet rares, sauf peut-être en psychologie expérimentale et en psychologie différentielle. On a affaire le plus souvent, que ce soit dans les échelles d'attitude ou d'opinion, ou dans les échelles de personnalité, élaborées à partir de questionnaires, qui sont utilisées en clinique, à des instruments de mesure dont les échelons sont simplement ordonnés et pour lesquels les notions de somme, et donc de moyenne, n'ont pas de sens. Il est d'autres cas où l'on ne dispose que de catégories, qui ne sont même pas ordonnées entre elles ou qui ne peuvent l'être que très partiellement, comme lorsqu'on fait juger des aliments selon un certain nombre de propriétés. Pourtant ces données sont traitables statistiquement ; on peut les résumer à l'aide de distributions, mais cela relève de la nature de l'échelle de mesure qui a servi au recueil de ces données. L'ouvrage de J.-M. Meunier met en relation de façon très précise les échelles de mesure, les procédures qui permettent de faire les distributions et les indices qui peuvent être calculés dans chaque cas pour pouvoir résumer au mieux la distribution. Les relations entre les échelles de mesure, les types de distribution et les types d'indices calculables sont présentées de façon suffisamment précise pour qu'un non-spécialiste des statistiques puisse faire des choix judicieux.

L'ouvrage est présenté de façon agréable avec une grande variété d'exemples et beaucoup d'illustrations qui retiennent l'attention et permettent d'éviter qu'un cours de statistiques soit rébarbatif. L'ouvrage sait aussi utiliser, quand c'est nécessaire, le formalisme des formules mathématiques, qui est indispensable pour décrire de façon non ambigüe les procédures de calcul, qui sont inévitablement complexes. Toutefois, la suite des calculs à effectuer est présentée de façon très progressive en liaison avec la formule, pour que l'étudiant puisse, à l'aide de sa calculette, reproduire la procédure et en comprendre la signification. Un étudiant doit savoir qu'un ouvrage de statistiques se lit calculette à la main, en reproduisant les procédures décrites et en faisant les exercices recommandés. Seul un apprentissage actif de traitement de données, comme en mathématiques la résolution de problèmes, permet une assimilation des procédures et des concepts sur lesquels elles sont basées.

La dernière partie introduit de façon approfondie à l'utilisation des tableurs. Les tableurs sont aujourd'hui l'outil indispensable à l'étudiant qui doit traiter des résultats de travaux pratiques, de travaux d'initiation à la recherche ou de travaux de recherche.

C'est un prolongement très heureux du cours, car la présentation des traitements sur des tableaux de données, qui est faite systématiquement dans le cours, se trouve exactement celle qui est adaptée pour traiter ces données à l'aide d'un tableur. Il devient donc possible de traiter des données conséquentes sans effort de calcul à la main. Les tableurs sont largement suffisants pour le calcul des statistiques descriptives qui font l'objet de cet ouvrage. Leur avantage est qu'ils permettent de suivre le déroulement des calculs et de découvrir éventuellement des incongruités, si la procédure a été mal choisie. L'utilisation de logiciels sophistiqués peut être dangereuse pour un étudiant peu averti, car elle exige une grande maîtrise conceptuelle du domaine pour une utilisation réfléchie et judicieuse. En période d'apprentissage, le tableur est plus transparent et c'est un des mérites de cet ouvrage de permettre de maîtriser cet outil et ainsi d'acquérir progressivement une autonome technique et conceptuelle dans la pratique statistique.

Jean-François Richard

Introduction

La nature est caractérisée par une grande diversité. Celle-ci dépend à la fois du climat, des conditions particulières dans lesquelles les espèces ont eu à se développer. La forme des arbres, leur taille, leur couleur, la faune qui les habitent sont autant de détails qui font qu'une forêt ne ressemble à aucune autre, et même qu'une parcelle particulière de celle-ci est différente d'une autre. Ainsi ceux qui se promènent souvent savent que les coins où il est possible de trouver au printemps des jonquilles, du muguet ne sont pas les mêmes que ceux où à l'automne les champignons vont abonder. Dans le même ordre d'idées, certains coins de la forêt vont être peuplés préférentiellement de bouleaux, d'autres de chênes ou de marronniers sans qu'on sache très bien les raisons de cette répartition. Cette diversité que nous décrivons pour les espèces végétales est également vraie pour les animaux. Certaines parties de la forêt constituent un habitat privilégié pour certaines espèces et pas d'autres. La proximité d'un point d'eau, la présence d'espèces végétales particulières, mais aussi la présence ou non de prédateurs peuvent expliquer ces disparités.

Pourtant la première chose qui frappe lorsqu'on regarde une forêt, c'est cette impression d'homogénéité. Elle semble former un tout relativement uniforme. Bien que nous percevions les variations de hauteurs, de formes et de couleurs, l'ensemble est bien appréhendé dans son unité. Cette impression n'est d'ailleurs pas seulement vraie lorsque la forêt est regardée de loin, elle l'est aussi, et peut-être même plus, lorsque nous la regardons de près. Ainsi en pénétrant dans la forêt, rien ne ressemble plus au tronc d'un arbre que le tronc de l'arbre qui est juste à côté ou même quelques centaines de mètres plus loin. À y regarder de plus près pourtant, la forme des branches, la présence d'un nid, quelques champignons au pied, ou une fourmilière installée dans les racines sont autant de caractéristiques qui devraient nous permettre de différencier les arbres que nous croisons dans cette forêt. Mais ces détails nous échappent et si nous n'y prenons pas garde, notre sens de l'orientation peut être sérieusement mis à mal par toutes ces ressemblances.

D'où nous vient cette propension à voir du semblable et à faire abstraction des différences ? Les mécanismes d'adaptation interviennent sans doute. La survie d'un individu passe par sa capacité à reconnaître les lieux où la nourriture est disponible, à identifier des lieux susceptibles de servir d'abri et surtout à identifier les situations de danger de façon à pouvoir y répondre sans délai. Pour cela, il faut identifier ce qu'il y a de commun à un ensemble d'objets ou de situations, c'est-à-dire distinguer ce qu'il y a de régulier

dans l'environnement de ce qui est purement fortuit. On peut ainsi apprendre à reconnaître les objets, les animaux ou les situations, mais aussi les distinguer des autres. Connaître les propriétés qui les caractérisent permet également de ne pas les confondre avec d'autres et de faire des inférences sur les propriétés qui ne sont pas directement perceptibles. Pour continuer notre promenade en forêt, imaginons que nous croisions un serpent. Celui-ci se faufile dans la végétation, sort de manière intermittente une langue fourchue et s'approche de nous. Faut-il être inquiet ? De là où on est, on aperçoit seulement la forme générale et sa couleur gris-brun. Mais on distingue également des rayures plus sombres. Il se pourrait bien que ce soit une couleuvre, sans danger pour l'homme. On arrive à cette conclusion car les rayures noires sont caractéristiques des couleuvres et non des vipères. Cette connaissance provient d'un savoir acquis par l'expérience ou la transmission sur les propriétés des serpents communs en Europe. Cette propriété permet également de faire des inférences sur d'autres caractéristiques de l'animal, notamment sur la forme de ces yeux, le risque ou non de morsure. Nul n'est besoin de s'approcher pour savoir que le serpent rayé de noir a vraisemblablement des yeux ronds et noirs et une tête plutôt ovale

Nous avons pu de loin reconnaître une couleuvre. Serions-nous capables de reconnaître avec la même assurance une vipère ? Outre l'absence de rayures et une robe brune, la vipère est également caractérisée par une tête plus triangulaire, des yeux jaunes et une pupille allongée tout à fait caractéristiques de cette espèce. Mais qui aura le courage d'aller voir suffisamment près le serpent pour voir sa tête afin d'être fixé ? Si nous n'avons pas bien distingué les rayures parce que le serpent se glisse dans des herbes hautes, à quoi avons-nous à faire ?

Cette petite balade en forêt illustre la nécessité d'identifier dans notre environnement des régularités pour connaître notre environnement, le comprendre et agir dans celui-ci éventuellement sur celui-ci de la façon la plus appropriée. Pour y parvenir, il faut que nous soyons en mesure de différencier ce qui relève du hasard et ce qui est suffisamment fréquent pour qu'on puisse le considérer comme une connaissance utile. Les champignons se rencontrent plus souvent au pied de certains arbres. Presque tous les ans, on trouve dans telle clairière des jonquilles, dans telle autre du muguet. Les couleuvres sont en général de couleur plus sombre que les vipères. Autant de régularités qu'on peut identifier parce qu'elles semblent ne pas relever du hasard.

Quel est le lien avec les statistiques ? L'identification de régularités en est le premier objectif. La mise en évidence des différences aussi. La démarche statistique est le prolongement de mécanismes d'apprentissage et de compréhension élémentaires qui nous permettent de savoir si au sein de la multitude d'événements, d'objets et de situations auxquels nous sommes confrontés, certains éléments sont suffisamment fréquents pour qu'ils puissent devenir une connaissance. Doit-on pour autant en conclure que nous sommes, sans le savoir, des statisticiens ? Pas si simple. Cela suggère quand même que les procédures statistiques ne sont pas si éloignées de notre façon d'appréhender notre environnement.

Henri Rouanet qualifiait de naturelles les procédures de description statistique car elles correspondent à des objectifs familiers (Rouanet, Le Roux et Bert, 1987) : se situer dans un ensemble, comparer des groupes d'individus ou étudier la relation entre deux variables. Rouanet propose une approche ensembliste et développe l'idée que les procédures statistiques reposent sur des structures de données qu'on peut décrire à travers des tableaux structurés. De la compréhension de ces structures de données dépend l'identification de la procédure appropriée. En effet, ces tableaux correspondent à différents types d'objets (variable, protocole ou distribution) qu'il convient de distinguer pour identifier ce qu'il est possible de faire. C'est la raison qui nous a conduits à l'instar de Richard (2004) qui en avait déjà montré l'intérêt pédagogique à en faire l'élément

central de la structuration de ce livre. Celui-ci s'adresse avant tout aux débutants, c'est-à-dire aux étudiants de première année de licence. Dans cet ouvrage, nous nous limiterons donc à leur programme d'études et nous n'y traiterons que les procédures de description statistique, c'est-à-dire les procédures permettant de tirer des conclusions sur un groupe d'observations particulier. Ceux qui poursuivent leurs études en seconde année et au-delà verront que cette description statistique peut être prolongée par l'inférence dont l'objet est de permettre des conclusions sur la population d'où est issu l'échantillon observé.

Une des difficultés majeures des étudiants, outre la mise en œuvre des procédures statistiques, est d'identifier les situations où elles sont pertinentes. Si la première difficulté est aisément surmontable à l'aide d'un logiciel spécialisé, ce dernier ne sera d'aucune aide pour la seconde et suppose une maîtrise conceptuelle de ces outils. Pour surmonter ces difficultés, il faut rendre plus saillantes les caractéristiques des situations d'analyse qui déterminent le choix d'une procédure ou d'une autre et ensuite donner suffisamment d'exemples pour que l'étudiant puisse construire les régularités dont il a besoin pour maîtriser ces connaissances. La structure de l'ouvrage vise ce premier objectif. La diversité des exercices proposés ici répond au second.

1

Les concepts statistiques

Une des principales appréhensions des étudiants en psychologie face aux statistiques provient de leur manque d'aisance avec les mathématiques. Pourtant, même si elle n'est pas triviale, la démarche statistique est le prolongement d'une tendance plutôt naturelle à regrouper, à identifier des régularités dans des phénomènes qui d'un premier abord présentent une grande variabilité. Avant d'être une affaire de procédures mathématiques et de calcul, les statistiques sont avant tout une tentative de compréhension d'un phénomène par l'examen des tendances générales et de la variabilité sur un grand nombre d'observations. Cette démarche repose sur un ensemble de concepts correspondant aux objets sur lesquels portent l'analyse statistique, les propriétés de ces objets. Bien comprendre ces concepts constitue une condition nécessaire pour identifier dans quel cas et pour quel objectif une procédure d'analyse plutôt qu'une autre est pertinente. Dans ce chapitre, nous allons présenter ces concepts, leur structuration et les mettre en relation avec les objectifs d'analyse et les procédures qui leur sont applicables.

I. Individu, échantillon, population

Individu statistique : objet sur lequel on recueille de l'information.

Population : ensemble d'individus étudiés.

Les méthodes statistiques sont applicables aux données relatives à des ensembles d'individus statistiques. Ces individus sont les objets, au sens large du terme, sur lesquels on prélève de l'information. Ces individus sont la plupart du temps, en psychologie, des personnes, mais pas toujours. L'ensemble d'individus sur lesquels on recueille de l'information est appelé « échantillon ». Cet échantillon est un sous-ensemble d'un ensemble d'individus étudiés plus large appelé « population ». Pour bien comprendre ces notions, prenons un exemple. L'encart ci-dessous présente le résumé d'un article de Goigoux (2000).

L'apprentissage de la lecture

« Les performances en lecture de soixante-seize enfants bénéficiant de deux méthodologies didactiques contrastées (une approche idéo-visuelle pure et une approche partiellement phonique) sont comparées au terme d'une étude longitudinale de vingt-huit mois (de la fin de grande section de maternelle au début du cours élémentaire deuxième année). Les élèves bénéficiant d'une didactique idéo-visuelle obtiennent des scores nettement inférieurs à ceux des autres élèves lors des évaluations nationales de CE2 malgré des performances initiales équivalentes en fin de scolarité maternelle. Leurs vitesses d'identification des mots écrits sont plus lentes que celles des élèves bénéficiant d'une didactique phonique rénovée. L'absence d'enseignement du code grapho-phonologique apparaît comme un obstacle à l'apprentissage de la lecture au cycle 2 et elle pénalise les élèves quelque que soit leur appartenance sociale. »
(*Source* : R. Goigoux (2000). « Apprendre à lire à l'école : les limites d'une approche idéo-visuelle », *Psychologie française*, n° 45-3, 2000, 233-243)

Échantillon : ensemble d'individus extraits d'une population.

Dans cette recherche, les individus statistiques sont les 73 enfants de même âge. Ils constituent ensemble l'échantillon étudié. Bien sûr, l'auteur cherche à tirer des conclusions généralisables à l'ensemble des enfants, c'est-à-dire à la population, et pas seulement aux 73 enfants de l'échantillon. Mais avant de pouvoir le faire, il doit mettre en évidence ce qu'il est possible d'observer sur son échantillon.

Les statistiques descriptives visent à caractériser l'échantillon. Dans la suite de vos études, vous verrez avec les statistiques inférentielles comment tirer des conclusions sur la population à partir de ce qui a été observé sur un échantillon.

II. Notions de variable et de facteur

Variable : dimension caractérisant les individus statistiques.

Pour étudier un échantillon, on relève un certain nombre d'informations sur les individus qui le composent. Ces informations concernent des dimensions caractérisant les individus comme l'âge, le sexe, la performance à une tâche. Ces dimensions sont appelées « variables ». Les différentes valeurs que peuvent prendre ces variables sont appelées modalités. Une