HAL
open science

Le cours de Biostatistiques

Xavier Nogues, André Garenne, Xavier Bouteiller, Virgil Fievet

To cite this version:

Xavier Nogues, André Garenne, Xavier Bouteiller, Virgil Fievet. Le cours de Biostatistiques. Dunod, 2018, 978-2-10-076976-6. hal-01939213

HAL Id: hal-01939213

https://inria.hal.science/hal-01939213

Submitted on 30 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BIOSTATISTIQUE

Tout le catalogue sur
www.dunod.com

DUNOD
ÉDITEUR DE SAVOIRS

Xavier Noguès,
André Garenne,
Xavier Bouteiller,
Virgil Fiévet

TOUT EN FICHES

LICENCE 3/MASTER/ÉCOLES D'INGÉNIEURS

LE COURS DE BIOSTATISTIQUE

110 FICHES DE COURS
120 SCHÉMAS
50 QCM

DUNOD

Illustration de couverture : © Sonja Calovini / fotolia.com

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).

DANGER
LE PHOTOCOPILLAGE TUE LE LIVRE

© Dunod, 2018

11, rue Paul Bert, 92240 Malakoff
www.dunod.com

ISBN 978-2-10-076976-6

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.

Table des matières

Avant-propos IX
Comment utiliser cet ouvrage ? XII
Remerciements XIV

Chapitre 1 Méthodologie de la recherche et vocabulaire de base

Fiche 1 Le déroulement d'une recherche 2
Fiche 2 Trois approches complémentaires : approche observationnelle, expérimentation et simulation 6
Fiche 3 Le statut des variables dans la recherche et le contrôle des facteurs 8
Fiche 4 Les types d'hypothèses au cours d'une recherche 10
Fiche 5 Qu'est-ce qu'une interaction statistique ? 12
Fiche 6 Généralisation du concept d'interaction statistique 14
Fiche 7 Les approches expérimentale et quasi-expérimentale 16
Fiche 8 Comment choisir une variable dépendante ? 18
Fiche 9 La conception d'un plan expérimental 22
Fiche 10 Comment neutraliser l'effet des facteurs secondaires ? 26
Fiche 11 Quel plan expérimental faut-il mettre en œuvre ? 28
Fiche 12 Comment constituer un échantillon représentatif ? 32
Fiche 13 Pourquoi les biologistes doivent-ils faire des statistiques ? 34

Focus Biologistes, devenez célèbres grâce aux statistiques ! 36
QCM 39

Chapitre 2 Comprendre les statistiques

Fiche 14 Paramètres de positions 42
Fiche 15 Indices de dispersion d'une population 44
Fiche 16 Indices de dispersion d'une population estimés à partir d'un échantillon 48
Fiche 17 Logique de raisonnement des statistiques inférentielles et notion de p-value 50
Fiche 18 Les méthodes de rééchantillonnage 52
Fiche 19 Comprendre le test de comparaisons de moyennes « *t* de Student » 54
Fiche 20 L'utilisation des tables pour le test *t* de Student 58
Fiche 21 Hypothèses fortes, hypothèses faibles, tests uni- et bilatéraux 60
Fiche 22 Comprendre la notion d'appariement et de mesures répétées 64
Fiche 23 Le théorème central limite et les principales lois de probabilité 66
Fiche 24 Les risques d'erreurs de première et deuxième espèce 70
Fiche 25 L'intervalle de fluctuation et intervalle de confiance 72
Fiche 26 Puissance d'un test et taille minimale d'échantillons 74
Fiche 27 Comprendre la formule de l'analyse de variance à un facteur 76
Fiche 28 Comprendre la covariance et la corrélation 80
Fiche 29 Régression linéaire, coefficient de détermination et analyse de la variance 84
Fiche 30 Les tests non paramétriques 88
Fiche 31 Principe des tests non paramétriques « par rangs » 90
Fiche 32 Le principe du test du χ^2 92
Fiche 33 Les analyses multivariées : comprendre l'analyse en composantes principales 96

Focus Comment faire un tour de magie avec les statistiques : le théorème central limite 100
QCM 103

Chapitre 3 Notions de base pour utiliser R en statistiques

Fiche 34	Les fondamentaux du logiciel R	106
Fiche 35	Création et manipulation de variables	108
Fiche 36	Les variables à deux dimensions	112
Fiche 37	Principe d'utilisation des bibliothèques (*packages*)	115
Fiche 38	Manipulation des données	116
Fiche 39	Fonctions graphiques de base	120
Fiche 40	Comment tracer des courbes avec R ?	122
Fiche 41	Les graphiques statistiques avec R	124
Fiche 42	Les tests statistiques avec R	127
Fiche 43	L'écriture de scripts	130
Fiche 44	L'utilisation des boucles	133
Fiche 45	Créer ses propres fonctions	135
Focus	Utilisons R pour fabriquer nos propres tests statistiques	137
QCM		139

Chapitre 4 Choisir le test approprié

Fiche 46	Clé : étude de l'effet de facteurs sur une seule variable dépendante quantitative	142
Fiche 47	Clé : questions posées sur un échantillon unique	144
Fiche 48	Clé : étude de l'effet d'un facteur unique sur une seule variable dépendante exprimée en rangs	145
Fiche 49	Clé : étude de l'effet d'un facteur unique sur une seule variable dépendante qualitative	146
Fiche 50	Clé : étude des relations entre quelques variables observées ou dépendantes	148
Fiche 51	Clé : plusieurs variables observées, analyses multivariées	149
Fiche 52	La distribution des données suit-elle une loi normale ?	150
Fiche 53	Vérification de normalité en ANOVA et régression	154
Fiche 54	Transformations mathématiques de variables sans perte d'information	156
Fiche 55	Transformations en rangs	158
Fiche 56	Transformation en classes ou en modalités	160
Fiche 57	Normalisation : centrage et réduction	162
Focus	« Normale » ou « pas normale » ?	164
Focus	Test « paramétrique » ou « non paramétrique » ?	167
QCM		169

Chapitre 5 Les tests paramétriques pour analyses univariées

Fiche 58	Comment comparer une moyenne observée à une moyenne théorique ?	172
Fiche 59	Le test *t* de Student pour échantillons indépendants et la correction de Welch	174
Fiche 60	Le test *t* de Student pour échantillons appariés	178
Fiche 61	L'analyse de variance à un facteur pour échantillons indépendants et le test de Tukey	180
Fiche 62	Les tests de comparaisons multiples	184
Fiche 63	La procédure de comparaisons planifiées et les corrections de Bonferroni et de Sidak	186
Fiche 64	L'analyse de variance à 1 facteur en mesures répétées	190
Fiche 65	La condition de sphéricité en ANOVA en mesures répétées	192
Fiche 66	L'ANOVA pour plans factoriels équilibrés	194

Fiche 67 L'ANOVA vue comme un modèle linéaire 198
Fiche 68 L'ANOVA pour plans hiérarchisés 200
Fiche 69 L'ANOVA pour plans mixtes (modèle III) 202
Fiche 70 L'ANOVA à plusieurs facteurs pour plans déséquilibrés 206
Fiche 71 La régression linéaire simple 210
Fiche 72 La régression linéaire multiple (RLM) 212
Fiche 73 Comment gérer de nombreux facteurs en RLM : les régressions par pas 216
Fiche 74 La régression par les moindres carrés partiels 221
Fiche 75 L'ANCOVA 224
Fiche 76 Comment comparer deux variances : le test de Snedecor 226
Fiche 77 Les tests d'hétérogénéité de variances 228

Focus Faut-il ajouter un « s » à « statistique » ? 231
QCM 233

Chapitre 6 Les tests non paramétriques pour analyses univariées

Fiche 78 Le test *U* de Mann-Whitney 236
Fiche 79 Le test de Kruskal-Wallis 240
Fiche 80 Le test *T* de Wilcoxon 242
Fiche 81 Le test de Friedman 244
Fiche 82 Quels tests *post hoc* utiliser après un test sur les rangs 246
Fiche 83 Le test du χ^2 sur table de contingence 248
Fiche 84 Le calcul de probabilité exacte (CPE) de Fisher 250
Fiche 85 Comment comparer une proportion observée à une proportion théorique 252
Fiche 86 Comment comparer plusieurs proportions indépendantes 254
Fiche 87 Comment comparer deux proportions en échantillons appariés : le test de McNemar 258
Fiche 88 Comment comparer plus de deux proportions en échantillons appariés : le test *Q* de Cochran 260
Fiche 89 Comment comparer deux distributions empiriques : le test de Kolmogorov-Smirnov 262
Fiche 90 Comment comparer une distribution empirique à une distribution théorique 264
Fiche 91 Les tests d'asymétrie et d'aplatissement 268

Focus Les tests statistiques à l'épreuve des tests statistiques : *crash test* 270
QCM 273

Chapitre 7 Les analyses multivariées

Fiche 92 Le coefficient de corrélation de Pearson et le coefficient de détermination 276
Fiche 93 Les corrélations de rangs 278
Fiche 94 Les coefficients de corrélations partielles 282
Fiche 95 Comparaison de deux coefficients de corrélations de Pearson 284
Fiche 96 Analyse de variance multivariée (MANOVA) 286
Fiche 97 L'algorithme des k-moyens 288
Fiche 98 Le positionnement multidimensionnel non métrique 292
Fiche 99 La classification ascendante hiérarchique (CAH) 296
Fiche 100 L'analyse en composantes principales (ACP) : la préparation des données 300
Fiche 101 L'ACP : choix du nombre d'axes à conserver 304
Fiche 102 L'ACP : interprétation de l'espace factoriel 306
Fiche 103 L'ACP : l'analyse des individus 308
Fiche 104 L'ACP : variables supplémentaires 310

Fiche 105 L'ACP : individus supplémentaires 312
Fiche 106 L'analyse factorielle des correspondances (AFC) 314
Fiche 107 L'analyse des correspondances multiples (ACM) 318
Fiche 108 Les variables supplémentaires en ACM 322
Fiche 109 La CAH couplée à une analyse factorielle 324
Fiche 110 L'analyse factorielle discriminante (AFD) 328

Focus L'ACP, les « véritables » analyses factorielles et les pistes pour la biologie 333
QCM 335

Exercices 337

Corrigés 347

Index 357

Avant-propos

Un grand nombre de personnes aiment remplir des grilles de mots croisés ou de sudokus, nous pensons que le même plaisir peut être pris en apprenant les statistiques.

1. À qui cet ouvrage s'adresse-t-il ?

En premier lieu, cet ouvrage s'adresse aux étudiants en **licence de biologie**, des filières **de la santé à l'écologi**e, mais le programme traité est également assez proche de celui dispensé en **sciences humaines**. Il s'adresse également aux étudiants de **master** dans ces disciplines, même si la couverture de l'ensemble des programmes aurait conduit à la rédaction d'un traité plutôt que d'un manuel. L'étudiant en biologie classique (biochimie, neurosciences, physiologie animale et végétale, biologie cellulaire, génétique...) y retrouvera la quasi-totalité de son programme. L'épidémiologiste ou l'écologue devront approfondir les analyses multivariées pour lesquelles cet ouvrage ne propose qu'une sensibilisation. Nous espérons que cet ouvrage apportera des solutions aux **chercheurs (doctorants et statutaires)**, tout en les incitant et les aidant à réactualiser leurs connaissances. Enfin, nous serions pleinement satisfaits si cet ouvrage pouvait aussi apporter, un réel plaisir aux **autodidactes** qui souhaitent se former à la pratique des statistiques.

2. Pourquoi un manuel supplémentaire en biostatistiques ?

Sans hésitations, nous répondons :

- parce que la pratique des statistiques par les biologistes a fortement évolué,
- parce que notre pratique de l'enseignement des biostatistiques nous incite à rénover et à repenser la didactique de cette discipline lorsque la formation s'adresse à des biologistes.

La pratique des biostatistiques évolue

En quarante ans, la pratique des statistiques a subi une révolution dans les laboratoires de biologie. Durant les années 1980, les tests de Student étaient effectués à la calculatrice et étaient employés comme tests de comparaisons multiples. Des analyses de variances étaient réalisées grâce à des ordinateurs, mais les données devaient être saisies à nouveaux en cas d'erreur. Au début des années 2000 les ordinateurs commencent à envahir les laboratoires, les experts de revues formulent des critiques sur les méthodes statistiques, et les logiciels de statistiques se développent.

Aujourd'hui, l'informatique a mis une très grande variété de méthodes statistiques à disposition de tous. La compétence du biologiste a donc dû évoluer. Il devient inutile de savoir calculer une statistique pour la comparer aux valeurs des tables. En revanche, il est nécessaire de connaître un grand nombre de procédures et de pouvoir justifier ses choix au moment de la présentation de résultats. Il faut également savoir se servir d'un logiciel de statistiques et interpréter correctement les résultats.

C'est vers l'acquisition de ces compétences que cet ouvrage est orienté.

Pédagogie et didactique des statistiques enseignées à des biologistes

Nos étudiants ne se sont pas engagés dans des études de biologie parce qu'ils espéraient y faire des statistiques. De plus, ils sont habitués à raisonner à partir de situations concrètes plus que sur des abstractions mathématiques. Pour ces deux raisons, nous expliquons ici les statistiques en nous basant sur des exemples concrets issus de la biologie et par une approche la plus intuitive possible.

Le langage. Pour la majorité des biologistes, les statistiques sont une activité intermittante. Les unités d'enseignement de statistiques sont souvent espacées de plusieurs mois, et le chercheur ne se plonge dans les statistiques qu'au moment du traitement de ses résultats. Dans cette perspective, nous avons tenté de respecter le langage des biologistes plus que celui des mathématiciens : nous avons considéré ici, que le lecteur doit faire le moins d'efforts possibles pour s'adapter à un langage qui n'est pas le sien, lorsqu'il ouvre son manuel après plusieurs semaines passées en cours de biologie, à la paillasse ou sur le terrain.

Pédagogie active. Sur le plan pédagogique, de nombreux enseignants sont à la recherche d'une approche permettant de faciliter l'enseignement des statistiques aux biologistes. Dans cet ouvrage, nous avons opté pour une pédagogie active sur plusieurs plans, sachant qu'un des grands principes de cette approche réside dans le fait que l'apprenant doit être acteur dans la construction de son savoir. Nous inspirant de l'« approche problème », chaque fiche s'ouvre sur une mise en situation. C'est également dans cette perspective que nous conduisons le lecteur à reconstruire les formules de plusieurs statistiques plutôt que de les lui expliquer. Nous l'incitons, par ailleurs, à mettre en application l'usage des tests au fur et à mesure avec le logiciel R.

Mini-apprentissage. Nous avons été séduits par les potentialités qu'offraient le format de la collection « Tout le cours en fiches ». Ce concept présente au moins deux atouts. Le premier est qu'il permet une forme de « mini-apprentissage », situé entre le micro-apprentissage (qui est une méthode d'apprentissage par séquences très courtes, de quelques secondes à trois minutes) et l'apprentissage plus approfondi. L'apprentissage d'une fiche de cet ouvrage nécessite quelques minutes de concentration. Lors de la lecture d'une entité d'un ouvrage classique, le plus souvent un chapitre, il est très difficile de reprendre là où l'on s'est arrêté. Ce format proposant de ne traiter qu'un seul concept par fiche permet l'acquisition d'entités cohérentes en une seule séance de lecture.

Pédagogie différentiable. Enfin, le format des fiches se prête à une certaine différenciation pédagogique, puisque le lecteur peut personnaliser sa lecture de l'ouvrage en fonction de ses motivations, de son rythme et de son style cognitif. Un apprenant classique pourra lire les fiches dans l'ordre et avec la logique qui lui est proposée. Un autre, plus impatient ou plus original, pourra lire l'ouvrage dans l'ordre qu'il souhaite, n'abordant certaines fiches de début d'ouvrage que lorsqu'il en ressent le besoin. Ainsi, les connaissances fondamentales (rébarbatives pour certains), pourront n'être abordées qu'au moment où elles apparaissent comme un besoin et perdront, par la même occasion, leur côté ennuyeux.

3. Comment utiliser ce manuel ?

L'étudiant en licence de biologie devrait trouver dans cet ouvrage la totalité du programme élaboré par les équipes pédagogiques. La structure « en fiches » lui fournira un soutien à l'enseignement qu'il reçoit, par une approche probablement différente et complémentaire, ce qui constitue un des intérêts de cet ouvrage. L'étudiant en master et le chercheur seront probablement plus intéressés par les clés de choix et les différentes solutions proposées pour résoudre leurs problèmes et exploiteront les chapitres 1 et 2 pour vérifier les formalisations qu'ils en font.

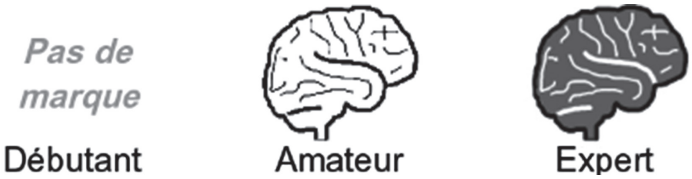
Niveaux de difficulté

En plus des outils très pratiques proposés par la collection, le lecteur trouvera une classification des niveaux des différentes fiches ou paragraphes afin de l'aider à calibrer son attention :

- le niveau « débutant », concerne les parties faciles normalement acquises en licence ;
- le niveau « amateur », concerne des concepts demandant plus de concentration, soit parce que leur acquisition est plus difficile, soit parce que leur maîtrise est incontournable ;
- le niveau « expert », concerne des méthodes acquises généralement en master. Ces méthodes ne sont pas forcément plus difficiles que celles qui ont été apprises en licence, mais demandent souvent un minimum de connaissances en statistiques. Elles peuvent d'ailleurs présenter un côté ludique qui devrait inciter les étudiants de licence à aller plus loin.

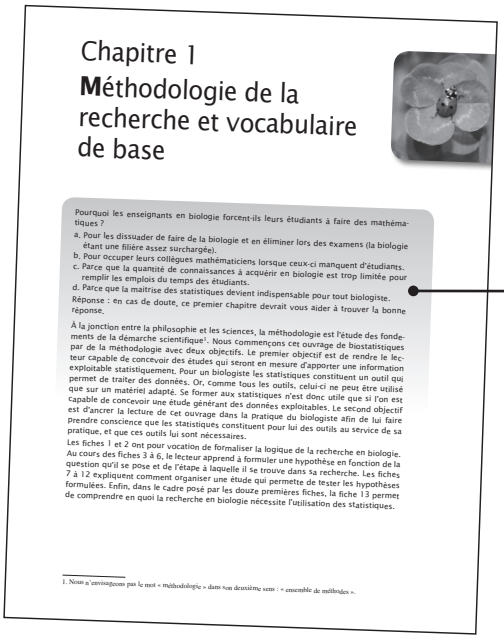
Jeux de données. Les jeux de données sont disponibles sur le site www.dunod.com (sur la page de présentation de l'ouvrage) afin de permettre au lecteur de mettre en pratique ses connaissances au fur et à mesure de la lecture de l'ouvrage. Comme nous l'avons expliqué, notre motivation en écrivant cet ouvrage est avant tout de faciliter l'apprentissage des statistiques. Nous n'avons donc pas hésité à simplifier ou à modifier des jeux de données existants, voire même à créer ces jeux de données de toutes pièces. Nous comptons donc sur le lecteur pour les considérer dans cet unique objectif, et surtout, ne pas citer ces travaux virtuels dans le cadre d'un mémoire !

En résumé, à travers cet ouvrage nous souhaitons aider les étudiant à passer leur examen avec succès, mais également leur fournir les compétences qui leurs seront utiles lorsqu'ils intègreront des équipes de recherches. Nous espérons qu'il aidera les chercheurs en biologie dans le traitement leurs données et qu'il donnera à tous, l'envie de se former aux biostatistiques avec **plaisir et curiosité**.



Les trois niveaux de pratique.

Comment utiliser



Chapitre 1

Méthodologie de la recherche et vocabulaire de base

Pourquoi les enseignants en biologie forcent-ils leurs étudiants à faire des mathématiques ?

a. Pour les dissuader de faire de la biologie et en éliminer lors des examens (la biologie étant une filière assez surchargée).
b. Pour occuper leurs collègues mathématiciens lorsque ceux-ci manquent d'étudiants.
c. Parce que la quantité de connaissances à acquérir en biologie est trop limitée pour remplir les emplois du temps des étudiants.
d. Parce que la maîtrise des statistiques devient indispensable pour tout biologiste.

Réponse : en cas de doute, ce premier chapitre devrait vous aider à trouver la bonne réponse.

À la jonction entre la philosophie et les sciences, la méthodologie est l'étude des fondements de la démarche scientifique[1]. Nous commençons cet ouvrage de biostatistiques par de la méthodologie avec deux objectifs. Le premier objectif est de rendre le lecteur capable de concevoir des études qui seront en mesure d'apporter une information exploitable statistiquement. Pour un biologiste les statistiques constituent un outil qui permet de traiter des données. Or, comme tous les outils, celui-ci ne peut être utilisé que sur un matériel adapté. Se former aux statistiques n'est donc utile que si l'on est capable de concevoir une étude générant des données exploitables. Le second objectif est d'ancrer la lecture de cet ouvrage dans la pratique du biologiste afin de lui faire prendre conscience que les statistiques constituent pour lui des outils au service de sa pratique, et que ces outils lui sont nécessaires.

Les fiches 1 et 2 ont pour vocation de formaliser la logique de la recherche en biologie. Au cours des fiches 3 à 6, le lecteur apprend à formuler une hypothèse en fonction de la question qu'il se pose et de l'étape à laquelle il se trouve dans sa recherche. Les fiches 7 à 12 expliquent comment organiser une étude qui permette de tester les hypothèses formulées. Enfin, dans le cadre posé par les douze premières fiches, la fiche 13 permet de comprendre en quoi la recherche en biologie nécessite l'utilisation des statistiques.

1. Nous n'envisageons pas le mot « méthodologie » dans son deuxième sens : « ensemble de méthodes ».

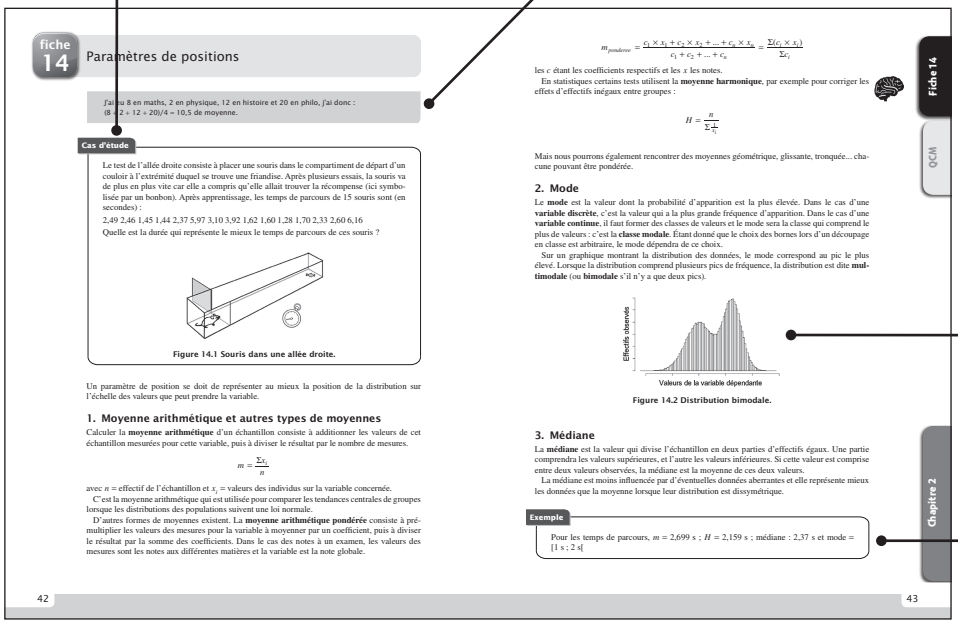
7 chapitres

Retrouvez les tableaux de données des exemples et des exercices sur la page associée à l'ouvrage sur dunod.com @

110 fiches de cours

Des cas d'étude

Les notions essentielles avec des renvois pour naviguer d'une fiche à l'autre



fiche 14 Paramètres de positions

J'ai eu 8 en maths, 2 en physique, 12 en histoire et 20 en philo, j'ai donc : (8 + 2 + 12 + 20)/4 = 10,5 de moyenne.

Cas d'étude

Le test de l'allée droite consiste à placer une souris dans le compartiment de départ d'un couloir à l'extrémité duquel se trouve une friandise. Après plusieurs essais, la souris va de plus en plus vite car elle a compris qu'elle allait trouver la récompense (ici symbolisée par un bonbon). Après apprentissage, les temps de parcours de 15 souris sont (en secondes) :

2,49 2,46 1,45 1,44 2,37 5,97 3,10 3,92 1,62 1,60 1,28 1,70 2,33 2,60 6,16

Quelle est la durée qui représente le mieux le temps de parcours de ces souris ?

Figure 14.1 Souris dans une allée droite.

Un paramètre de position se doit de représenter au mieux la position de la distribution sur l'échelle des valeurs que peut prendre la variable.

1. Moyenne arithmétique et autres types de moyennes

Calculer la **moyenne arithmétique** d'un échantillon consiste à additionner les valeurs de cet échantillon mesurées pour cette variable, puis à diviser le résultat par le nombre de mesures.

$$m = \frac{\Sigma x_i}{n}$$

avec n = effectif de l'échantillon et x_i = valeurs des individus sur la variable concernée.

C'est la moyenne arithmétique qui est utilisée pour comparer les tendances centrales de groupes lorsque les distributions des populations suivent une loi normale.

D'autres formes de moyennes existent. La **moyenne arithmétique pondérée** consiste à pré-multiplier les valeurs des mesures pour la variable à moyenner par un coefficient, puis à diviser le résultat par la somme des coefficients. Dans le cas des notes à un examen, les valeurs des mesures sont les notes aux différentes matières et la variable est la note globale.

42

$$m_{pondérée} = \frac{c_1 \times x_1 + c_2 \times x_2 + ... + c_n \times x_n}{c_1 + c_2 + ... + c_n} = \frac{\Sigma(c_i \times x_i)}{\Sigma c_i}$$

les c étant les coefficients respectifs et les x les notes.

En statistiques certains tests utilisent la **moyenne harmonique**, par exemple pour corriger les effets d'effectifs inégaux entre groupes :

$$H = \frac{n}{\Sigma \frac{1}{x_i}}$$

Mais nous pourrons également rencontrer des moyennes géométrique, glissante, tronquée... chacune pouvant être pondérée.

2. Mode

Le **mode** est la valeur dont la probabilité d'apparition est la plus élevée. Dans le cas d'une **variable discrète**, c'est la valeur qui a la plus grande fréquence d'apparition. Dans le cas d'une **variable continue**, il faut former des classes de valeurs et le mode sera la classe qui comprend le plus de valeurs : c'est la **classe modale**. Étant donné que le choix des bornes lors d'un découpage en classe est arbitraire, le mode dépendra de ce choix.

Sur un graphique montrant la distribution des données, le mode correspond au pic le plus élevé. Lorsque la distribution comprend plusieurs pics de fréquence, la distribution est dite **multimodale** (ou **bimodale** s'il n'y a que deux pics).

Effectifs observés

Valeurs de la variable dépendante

Figure 14.2 Distribution bimodale.

3. Médiane

La **médiane** est la valeur qui divise l'échantillon en deux parties d'effectifs égaux. Une partie comprendra les valeurs supérieures, et l'autre les valeurs inférieures. Si cette valeur est comprise entre deux valeurs observées, la médiane est la moyenne de ces deux valeurs.

La médiane est moins influencée par d'éventuelles données aberrantes et elle représente mieux les données que la moyenne lorsque leur distribution est dissymétrique.

Exemple

Pour les temps de parcours, m = 2,699 s ; H = 2,159 s ; médiane : 2,37 s et mode = [1 s ; 2 s[

Fiche 14 QCM Chapitre 2

43

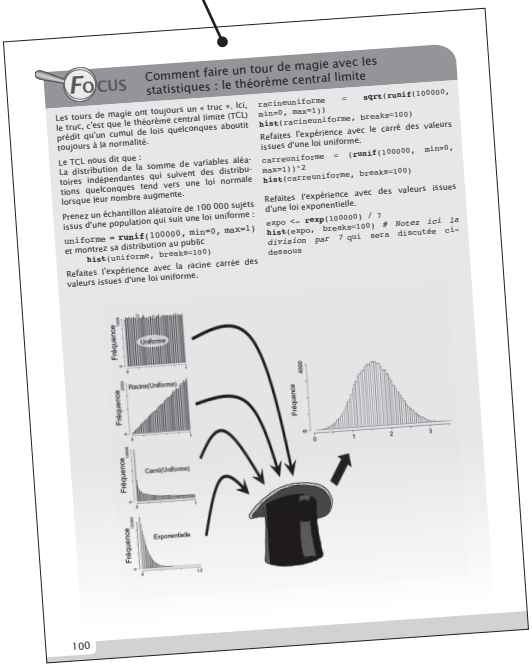
De nombreux schémas

Des exemples

cet ouvrage ?

Des focus à la fin de chaque chapitre

Des QCM et leurs réponses commentées au verso à la fin de chaque chapitre



FOCUS Comment faire un tour de magie avec les statistiques : le théorème central limite

Les tours de magie ont toujours un « truc ». Ici, le truc, c'est que le théorème central limite (TCL) prédit qu'un cumul de lois quelconques aboutit toujours à la normalité.

Le TCL nous dit que :
La distribution de la somme de variables aléatoires indépendantes qui suivent des distributions quelconques tend vers une loi normale lorsque leur nombre augmente.

Prenez un échantillon aléatoire de 100 000 sujets issus d'une population qui suit une loi uniforme :

```
uniforme = runif(100000, min=0, max=1)
hist(uniforme, breaks=100)
```

Refaites l'expérience avec la racine carrée des valeurs issues d'une loi uniforme.

```
racineuniforme = sqrt(runif(100000, min=0, max=1))
hist(racineuniforme, breaks=100)
```

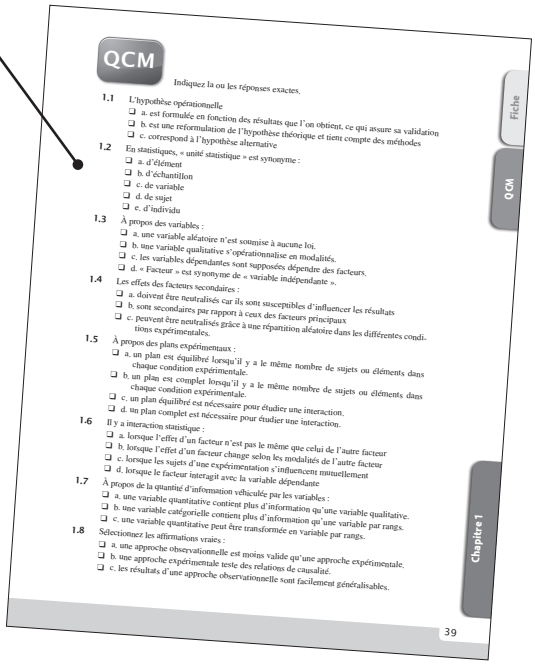
Refaites l'expérience avec le carré des valeurs issues d'une loi uniforme.

```
carreuniforme = (runif(100000, min=0, max=1))^2
hist(carreuniforme, breaks=100)
```

Refaites l'expérience avec des valeurs issues d'une loi exponentielle.

```
expo <- rexp(100000) / 7
hist(expo, breaks=100) # Notez ici la division par 7 qui sera discutée ci-dessous
```


100



QCM

Indiquez la ou les réponses exactes.

1.1 L'hypothèse opérationnelle
- ❑ a. est formulée en fonction des résultats que l'on obtient, ce qui assure sa validation
- ❑ b. est une reformulation de l'hypothèse théorique et tient compte des méthodes
- ❑ c. correspond à l'hypothèse alternative

1.2 En statistiques, « unité statistique » est synonyme :
- ❑ a. d'élément
- ❑ b. d'échantillon
- ❑ c. de variable
- ❑ d. de sujet
- ❑ e. d'individu

1.3 À propos des variables :
- ❑ a. une variable aléatoire n'est soumise à aucune loi.
- ❑ b. une variable qualitative s'opérationnalise en modalités.
- ❑ c. les variables dépendantes sont supposées dépendre des facteurs.
- ❑ d. « Facteur » est synonyme de « variable indépendante ».

1.4 Les effets des facteurs secondaires :
- ❑ a. doivent être neutralisés car ils sont susceptibles d'influencer les résultats
- ❑ b. sont secondaires par rapport à ceux des facteurs principaux
- ❑ c. peuvent être neutralisés grâce à une répartition aléatoire dans les différentes conditions expérimentales.

1.5 À propos des plans expérimentaux :
- ❑ a. un plan est équilibré lorsqu'il y a le même nombre de sujets ou éléments dans chaque condition expérimentale.
- ❑ b. un plan est complet lorsqu'il y a le même nombre de sujets ou éléments dans chaque condition expérimentale.
- ❑ c. un plan équilibré est nécessaire pour étudier une interaction.
- ❑ d. un plan complet est nécessaire pour étudier une interaction.

1.6 Il y a interaction statistique :
- ❑ a. lorsque l'effet d'un facteur n'est pas le même que celui de l'autre facteur
- ❑ b. lorsque l'effet d'un facteur change selon les modalités de l'autre facteur
- ❑ c. lorsque les sujets d'une expérimentation s'influencent mutuellement
- ❑ d. lorsque le facteur interagit avec la variable dépendante

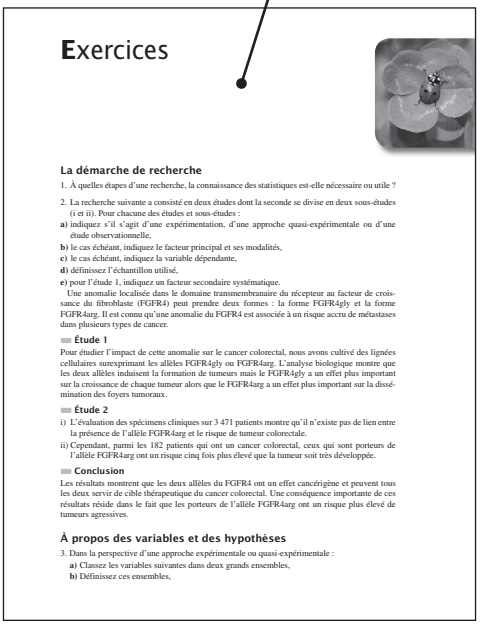
1.7 À propos de la quantité d'information véhiculée par les variables :
- ❑ a. une variable quantitative contient plus d'information qu'une variable qualitative.
- ❑ b. une variable catégorielle contient plus d'information qu'une variable par rangs.
- ❑ c. une variable quantitative peut être transformée en variable par rangs.

1.8 Sélectionnez les affirmations vraies :
- ❑ a. une approche observationnelle est moins valide qu'une approche expérimentale.
- ❑ b. une approche expérimentale teste des relations de causalité.
- ❑ c. les résultats d'une approche observationnelle sont facilement généralisables.

Fiche
QCM
Chapitre 1

39

Des exercices pour s'entraîner en fin d'ouvrage



Exercices

La démarche de recherche

1. À quelles étapes d'une recherche, la connaissance des statistiques est-elle nécessaire ou utile ?

2. La recherche suivante a consisté en deux études dont la seconde se divise en deux sous-études (i et ii). Pour chacune des études et sous-études :
a) indiquez s'il s'agit d'une expérimentation, d'une approche quasi-expérimentale ou d'une étude observationnelle,
b) le cas échéant, indiquez le facteur principal et ses modalités,
c) le cas échéant, indiquez la variable dépendante,
d) définissez l'échantillon utilisé,
e) pour l'étude 1, indiquez un facteur secondaire systématique.

Une anomalie localisée dans le domaine transmembranaire du récepteur au facteur de croissance du fibroblaste (FGFR4) peut prendre deux formes : la forme FGFR4gly et la forme FGFR4arg. Il est connu qu'une anomalie du FGFR4 est associée à un risque accru de métastases dans plusieurs types de cancer.

Étude 1

Pour étudier l'impact de cette anomalie sur le cancer colorectal, nous avons cultivé des lignées cellulaires surexprimant les allèles FGFR4gly ou FGFR4arg. L'analyse biologique montre que les deux allèles induisent la formation de tumeurs mais le FGFR4gly a un effet plus important sur la croissance de chaque tumeur alors que le FGFR4arg a un effet plus important sur la dissémination des foyers tumoraux.

Étude 2

i) L'évaluation des spécimens cliniques sur 3 471 patients montre qu'il n'existe pas de lien entre la présence de l'allèle FGFR4arg et le risque de tumeur colorectale.
ii) Cependant, parmi les 182 patients qui ont un cancer colorectal, ceux qui sont porteurs de l'allèle FGFR4arg ont un risque cinq fois plus élevé que la tumeur soit très développée.

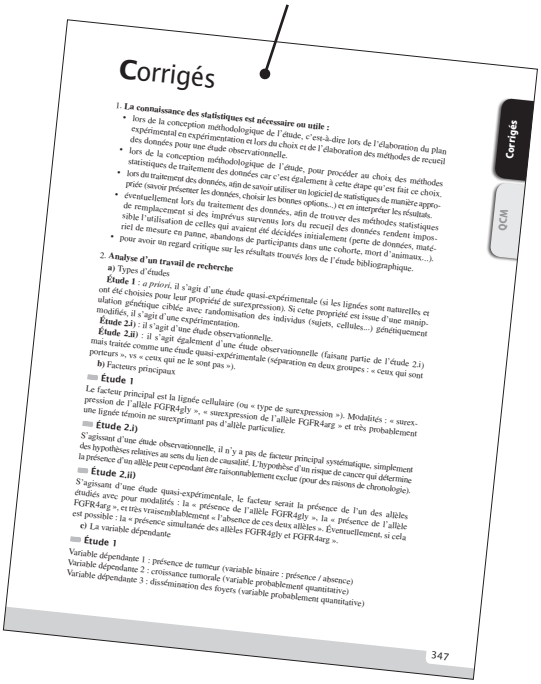
Conclusion

Les résultats montrent que les deux allèles du FGFR4 ont un effet cancérigène et peuvent tous les deux servir de cible thérapeutique du cancer colorectal. Une conséquence importante de ces résultats réside dans le fait que les porteurs de l'allèle FGFR4arg ont un risque plus élevé de tumeurs agressives.

À propos des variables et des hypothèses

3. Dans la perspective d'une approche expérimentale ou quasi-expérimentale :
a) Classez les variables suivantes dans deux grands ensembles,
b) Définissez ces ensembles,

Les corrigés commentés



Corrigés

1. **La connaissance des statistiques est nécessaire ou utile :**
- lors de la conception méthodologique de l'étude, c'est-à-dire lors de l'élaboration du plan expérimental ou expérimentation et lors du choix et de l'élaboration des méthodes de recueil des données pour une étude observationnelle.
- lors de la conception méthodologique de l'étude, pour procéder au choix des méthodes statistiques de traitement des données car c'est également à cette étape qu'est fait ce choix.
- lors du traitement des données, afin de savoir utiliser un logiciel de statistiques de manière appropriée (savoir présenter les données, choisir les bonnes options,...) et en interpréter les résultats.
- éventuellement lors du traitement des données, afin de trouver des méthodes statistiques de remplacement si des imprévus survenus lors du recueil des données rendent impossible l'utilisation de celles qui avaient été décidées initialement (perte de données, matériel de mesure en panne, abandons de participants dans une cohorte, mort d'animaux,...).
- pour avoir un regard critique sur les résultats trouvés lors de l'étude bibliographique.

2. **Analyse d'un travail de recherche**

a) Types d'études

Étude 1 : *a priori*, il s'agit d'une étude quasi-expérimentale (si les lignées sont naturelles et ont été choisies pour leur propriété de surexpression). Si cette propriété est issue d'une manipulation génétique ciblée avec randomisation des individus (sujets, cellules,...) génétiquement modifiés, il s'agit d'une expérimentation.

Étude 2.i) : il s'agit d'une étude observationnelle.

Étude 2.ii) : il s'agit également d'une étude observationnelle (faisant partie de l'étude 2.i) mais traitée comme une étude quasi-expérimentale (séparation en deux groupes : « ceux qui sont porteurs », vs « ceux qui ne le sont pas »).

b) Facteurs principaux

Étude 1

Le facteur principal est la lignée cellulaire (ou « type de surexpression »). Modalités : « surexpression de l'allèle FGFR4gly », « surexpression de l'allèle FGFR4arg » et très probablement une lignée témoin ne surexprimant pas d'allèle particulier.

Étude 2.i)

S'agissant d'une étude observationnelle, il n'y a pas de facteur principal systématique, simplement des hypothèses relatives au sens du lien de causalité. L'hypothèse d'un risque de cancer qui détermine la présence d'un allèle peut cependant être raisonnablement exclue (pour des raisons de chronologie).

Étude 2.ii)

S'agissant d'une étude quasi-expérimentale, le facteur serait la présence de l'un des allèles étudiés avec pour modalités : la « présence de l'allèle FGFR4gly », la « présence de l'allèle FGFR4arg », et très vraisemblablement « l'absence de ces deux allèles ». Éventuellement, si cela est possible : la « présence simultanée des allèles FGFR4gly et FGFR4arg ».

c) La variable dépendante

Étude 1

Variable dépendante 1 : présence de tumeur (variable binaire : présence / absence)
Variable dépendante 2 : croissance tumorale (variable probablement quantitative)
Variable dépendante 3 : dissémination des foyers (variable probablement quantitative)

Corrigés
QCM

347

Remerciements

Nous tenons à remercier chaleureusement nos collègues qui ont accepté de participer au comité de lecture, pour leurs relectures parfois très minutieuses, leur aide, leurs conseils et leurs encouragements. Il a été très enrichissant d'avoir leur avis, tant sur la structure de l'ouvrage que pour la diversité des approches et au sujet de la pédagogie des biostatistiques. Bien sûr, ces personnes qui nous ont apporté leur aide ne sont pas responsables des erreurs qui pourraient persister dans cet ouvrage, ni des avis, choix et arbitrages que nous avons dû faire tout au long de la rédaction.

Nous sommes donc très heureux de pouvoir remercier :

- Leslie Regad, maître de conférences à l'université Paris Diderot,
- Franck Brignolas, professeur à l'université d'Orléans,
- Lionel Denis, professeur à l'université de Lille,
- Léo Gerville-Réache, maître de conférences à l'université de Bordeaux,
- Gilles Hunault, maître de conférences à l'université d'Angers,
- Laurent Pezard, professeur à l'université de Provence.

Enfin, nous remercions Laëtitia Hérin et Vanessa Beunèche des éditions Dunod, avec qui nous avons eu grand plaisir à travailler, et nos familles pour leur patience pendant ces huit mois de rédaction.