

TOUT EN FICHES

LE COURS DE

BIOSTATISTIQUE

LICENCE 3, MASTER, ÉCOLES D'INGÉNIEURS

2^e
ÉDITION

Xavier Noguès, André Garenne et Virgil Fiévet

Enseignants-chercheurs à l'Université de Bordeaux

Xavier Bouteiller

Biostatisticien au CHU de Bordeaux-IHU Liryc

DUNOD

Illustration de couverture : © Sonja Calovini / fotolia.com

© Dunod, 2018, 2022

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

ISBN 978-2-10-084995-6

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	VII
Comment utiliser cet ouvrage ?	X
Remerciements	XII

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	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	Méthodologie et éthique lors le l'évaluation de médicaments	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 comparaison de moyennes « <i>t</i> de Student »	54
Fiche 20	L'utilisation des tables pour le test <i>t</i> 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 gérer des valeurs suspectes ?	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	Manipulation des données	115
Fiche 38	Principes de mise en œuvre des tests statistiques dans R	118
Fiche 39	Principe d'utilisation des bibliothèques (<i>packages</i>)	121
Fiche 40	Fonctions graphiques de base	122
Fiche 41	Comment tracer des courbes avec R	124
Fiche 42	Graphiques statistiques avec R	126
Fiche 43	L'écriture de programmes et de scripts	129
Fiche 44	L'utilisation des boucles	132
Fiche 45	Créer ses propres fonctions	134
Focus	Quelques pistes pour accélérer l'exécution d'un code R	136
QCM		141

Chapitre 4 Clés de choix de tests et méthodes pluri-catégorielles

Fiche 46	Clé : étude de l'effet de facteurs sur une seule variable dépendante quantitative	144
Fiche 47	Clé : questions posées sur un échantillon unique	146
Fiche 48	Clé : étude de l'effet d'un facteur unique sur une seule variable dépendante exprimée en rangs	147
Fiche 49	Clé : étude de l'effet d'un facteur unique sur une seule variable dépendante qualitative	148
Fiche 50	Clé : étude des relations entre quelques variables observées ou dépendantes	150
Fiche 51	Clé : plusieurs variables observées, analyses multivariées	151
Fiche 52	La distribution des données suit-elle une loi normale ?	152
Fiche 53	Vérification de normalité en ANOVA et régression	156
Fiche 54	Transformations mathématiques de variables sans perte d'information	158
Fiche 55	Transformations en rangs	160
Fiche 56	Transformation en classes ou en modalités	162
Fiche 57	Normalisation : centrage et réduction	164
Fiche 58	La procédure de comparaisons planifiées et les corrections de Bonferroni et de Sidak	166
Fiche 59	La taille de l'effet	170
Focus	Choisissez les bons tests statistiques, pour sauver le monde ! (Jeu en « <i>Scenario Based Learning</i> »)	173
QCM		177

Chapitre 5 Les tests paramétriques pour analyses univariées

Fiche 60	Comment comparer une moyenne observée à une moyenne théorique ?	180
Fiche 61	Le test <i>t</i> de Student pour échantillons indépendants et la correction de Welch	182
Fiche 62	Le test <i>t</i> de Student pour échantillons appariés	186
Fiche 63	L'analyse de variance à un facteur pour échantillons indépendants et le test de Tukey	188
Fiche 64	Les tests de comparaisons multiples	192
Fiche 65	L'analyse de variance à un facteur en mesures répétées	194
Fiche 66	La condition de sphéricité en ANOVA en mesures répétées	196
Fiche 67	L'ANOVA pour plans factoriels équilibrés	198
Fiche 68	L'ANOVA vue comme un modèle linéaire	202
Fiche 69	L'ANOVA pour plans hiérarchisés	204
Fiche 70	L'ANOVA pour plans mixtes (modèle III)	206

Fiche 71	L'ANOVA à plusieurs facteurs pour plans déséquilibrés	210
Fiche 72	La régression linéaire simple	214
Fiche 73	La régression linéaire multiple (RLM)	216
Fiche 74	Comment gérer de nombreux facteurs en RLM : les régressions par pas	220
Fiche 75	La régression par les moindres carrés partiels	224
Fiche 76	Les modèles linéaires généralisés (GLM)	228
Fiche 77	Modèles linéaires généralisés sur données binaires : la régression logistique	232
Fiche 78	L'ANCOVA	236
Fiche 79	GLM binomial avec variable concomitante	238
Fiche 80	Comment comparer deux variances : le test de Snedecor	240
Fiche 81	Les tests d'hétérogénéité de variances	242
Focus	L'esprit des lois	246
QCM		249

Chapitre 6 Les tests non paramétriques pour analyses univariées

Fiche 82	Le test U de Mann-Whitney	252
Fiche 83	Le test de Kruskal-Wallis	256
Fiche 84	Le test T de Wilcoxon	258
Fiche 85	Le test de Friedman	260
Fiche 86	Quels tests <i>post hoc</i> utiliser après un test sur les rangs	262
Fiche 87	Les PERMANOVA	264
Fiche 88	Le test du χ^2 sur table de contingence	266
Fiche 89	Le calcul de probabilité exacte (CPE) de Fisher	268
Fiche 90	Comment comparer une proportion observée à une proportion théorique	270
Fiche 91	Comment comparer plusieurs proportions indépendantes	272
Fiche 92	Comment comparer des proportions en échantillons appariés : le test Q de Cochran	276
Fiche 93	Comment comparer deux distributions empiriques : le test de Kolmogorov-Smirnov	278
Fiche 94	Comment comparer une distribution empirique à une distribution théorique	280
Fiche 95	Les tests d'asymétrie et d'aplatissement	284
Focus	La librairie {dplyr} pour vous faciliter le travail	286
QCM		289

Chapitre 7 Les analyses multivariées

Fiche 96	Le coefficient de corrélation de Pearson et le coefficient de détermination	292
Fiche 97	Les corrélations de rangs	294
Fiche 98	Les coefficients de corrélations partielles	298
Fiche 99	Comparaison de deux coefficients de corrélations de Pearson	300
Fiche 100	Analyse de variance multivariée (MANOVA)	302
Fiche 101	L'algorithme des k-moyens	304
Fiche 102	Le positionnement multidimensionnel non métrique	308
Fiche 103	La classification ascendante hiérarchique (CAH)	312
Fiche 104	L'analyse en composantes principales (ACP) : la préparation des données	316
Fiche 105	L'ACP : choix du nombre d'axes à conserver	320
Fiche 106	L'ACP : interprétation de l'espace factoriel	322
Fiche 107	L'ACP : l'analyse des individus	324
Fiche 108	L'ACP : variables supplémentaires	326
Fiche 109	L'ACP : individus supplémentaires	328
Fiche 110	L'analyse factorielle des correspondances (AFC)	330

Fiche 111	L'analyse des correspondances multiples (ACM)	334
Fiche 112	Les variables supplémentaires en ACM	338
Fiche 113	La CAH sur résultats d'analyse factorielle	340
Fiche 114	L'ACP non paramétrique	344
Fiche 115	L'analyse de Hill-Smith	346
Fiche 116	L'analyse factorielle discriminante (AFD)	350
Focus	{ggplot2} et autres trucs pour faire de beaux graphiques	355
<i>QCM</i>		359
Exercices		361
Corrigés		373
Index		379

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'écologie**, 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 hésitation, 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 différenciable. 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 Dunod, à l'adresse 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 étudiants à passer leur examen avec succès, mais également leur fournir les compétences qui leurs seront utiles lorsqu'ils intégreront 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**.

4. Le site d'auteur, pour quoi faire ?

À l'adresse internet suivante, <https://biostatistique.blog4ever.com/> vous trouverez le « site d'auteur » qui vous permettra de :

- poser des questions (anonymement pour les timides) si vous n'avez pas compris certains points. Nous répondrons à vos questions qui, en retour, nous aident à améliorer nos explications ;
- discuter les points que vous souhaitez. En effet, les avis en biostatistiques ne sont pas universels. Les enseignants n'indiquent pas tous les mêmes réponses pour un même problème. Cette partie est là pour comprendre la diversité des avis et aussi pour identifier les « idées fausses » ;
- retrouver les bonus Web (également présents sur le site Dunod).



Les stats, ça peut être amusant. Vous trouverez sur ce site des jeux vous permettant de réviser et de mettre en application ce que vous avez appris dans l'ouvrage.

Et parce que tout le monde ne connaît pas l'ouvrage vous trouverez la présentation de son fonctionnement et un index détaillé qui aide à savoir si l'ouvrage peut ou pas répondre à la question que vous vous posez.

*Pas de
marque*

Débutant



Amateur



Expert

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 ?

- Pour les dissuader de faire de la biologie et en éliminer lors des examens (la biologie étant une filière assez surchargée).
- Pour occuper leurs collègues mathématiciens lorsque ceux-ci manquent d'étudiants.
- Parce que la quantité de connaissances à acquérir en biologie est trop limitée pour remplir les emplois du temps des étudiants.
- 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 scientifique¹. 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 fermer 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. Note : il conviendrait peut-être de dire « 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 :

$$(8 + 2 + 12 + 20) / 4 = 10,5 \text{ de moyenne.}$$

Cas d'étude

Le test de l'allée droite consiste à placer une souris dans le compartiment de départ d'un labyrinthe 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 ?

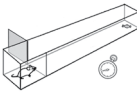


Figure 14.1 Souris dans une allée droite.

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.

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

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.

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

les c_i étant les coefficients respectifs et les x_i les notes.

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

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

Mais nous pourrions 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).

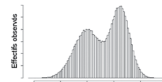


Figure 14.2 Distribution bimodale.

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.

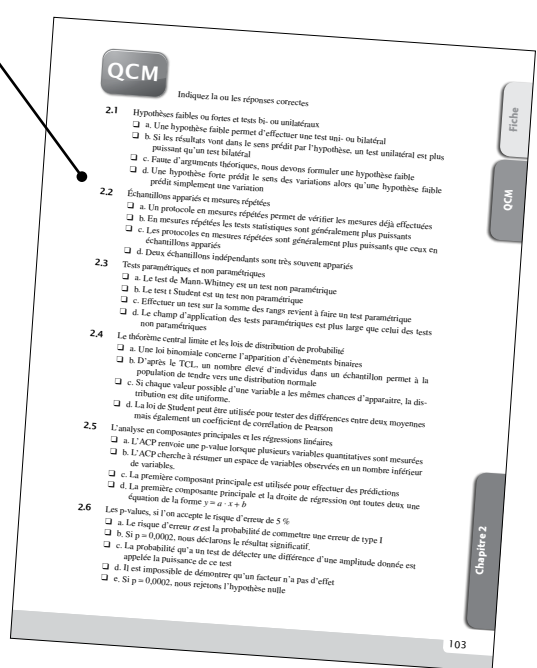
De nombreux
schémas

Des exemples

cet ouvrage ?

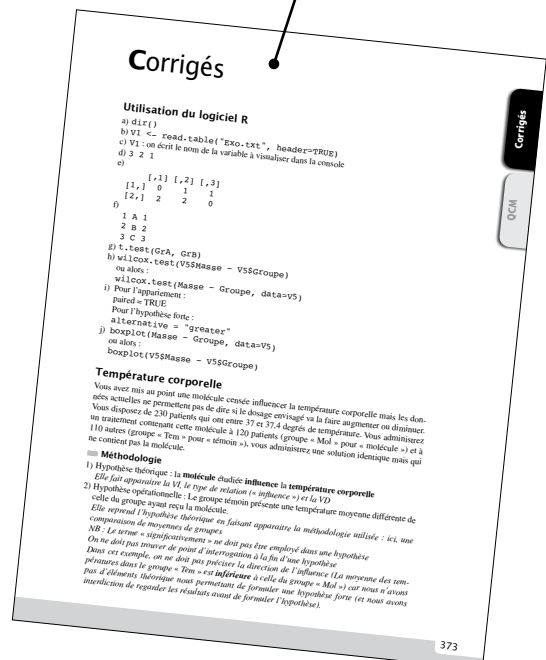
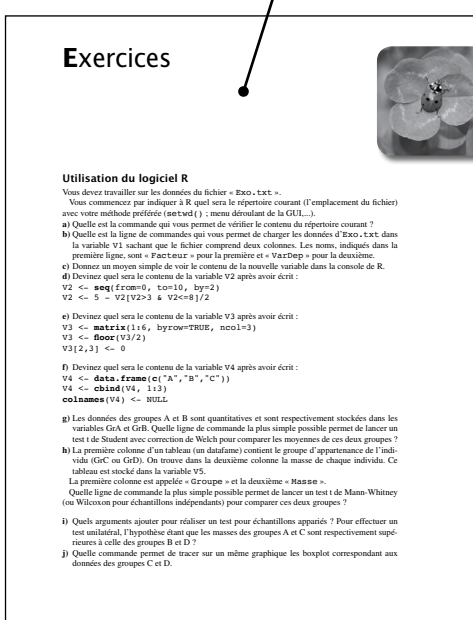
Des focus à la fin de chaque chapitre

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



Des exercices pour s'entraîner
en fin d'ouvrage

Les corrigés commentés



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 Jammet, Raquel De Macedo e Santos 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.

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 scientifique¹. 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 ».

La recherche scientifique vue comme un voyage.

Vous avez envie de partir en voyage dans un pays qui vous est inconnu. Que faites-vous ?

1. Vous choisissez un pays qui vous intéresse (par exemple la Mongolie).
2. Étant limité en temps et en ressources, des choix s'imposent et vous devez préciser vos motivations : culture nomade ? Équitation ? Pour faire ces choix, vous consultez des guides de voyage.
3. Votre objectif est maintenant plus précis. D'après vos informations, vous supposez que c'est dans l'ouest du pays que vous pourrez au mieux satisfaire vos attentes.
4. Vous planifiez votre voyage et sa logistique pour qu'il réponde à vos attentes.
5. Vous imaginez maintenant comment aller au point qui vous intéresse et comment pratiquer l'activité de votre choix.
6. Vous partez en voyage et revenez avec vos souvenirs, vos notes et vos photos.
7. Vous triez les photos et mettez vos notes au propre.
8. Vous faites de nouveaux projets pour les prochaines vacances.
9. Vous montrez vos photos à vos amis et racontez votre aventure.

En conduisant ce projet, vous venez de reproduire les mêmes étapes que celles d'une recherche scientifique. Les numéros de paragraphes de cette fiche correspondent à ces étapes.

1. La thématique et la question de départ

Tout commence par une thématique et une question de départ. La **thématique** sera par exemple « le cancer », « la mémoire », « les dauphins ». La **question de départ** sera : comment peut-on soigner le cancer ? Comment faire pour avoir une bonne mémoire ? Comment protéger les dauphins ? La thématique et la question de départ dépendent des motivations individuelles. La science ne peut pas, en général, apporter de réponse à la question de départ pendant le temps imparti pour une recherche (un stage, une thèse, voire même la carrière d'un chercheur). Il faut préciser cette question ; c'est le rôle de la constitution de la problématique.

2. La recherche bibliographique et la constitution de la problématique

La même question a déjà probablement été posée antérieurement. Une **recherche bibliographique** permet d'éclairer les voies qui ont déjà été explorées. Construire la **problématique** consiste à agencer les connaissances extraites de la recherche bibliographique pour préciser la question de départ et arriver à une question précise : le **problème**. Dans certains cas, la problématique conduit à la construction d'un **modèle d'analyse**, c'est-à-dire un certain nombre de variables mises en relation. Chacune de ces relations peut être testée et donc constituer un problème.

3. Le problème et l'hypothèse théorique

C'est une question issue de la problématique et à laquelle le chercheur se propose de répondre. Le problème est suffisamment précis lorsqu'il est possible de mettre en place une expérience (expérimentation ou étude observationnelle).

La molécule X intervient-elle dans la mémorisation ?

Le modèle d'analyse montre comment le chercheur conçoit les relations de causalité au sein du système qu'il étudie. La figure 1.1 montre les relations de causalité prises en considérations (trait plein), des exemples de relations négligées (pointillés), et la relation testée dans l'expérience (le problème ; trait double) sachant que cette dernière ne correspond pas forcément à une relation de causalité directe. D'autres relations ont probablement été omises volontairement ou non : un modèle d'analyse est une représentation simplifiée de la réalité, justifiée par les connaissances théoriques.

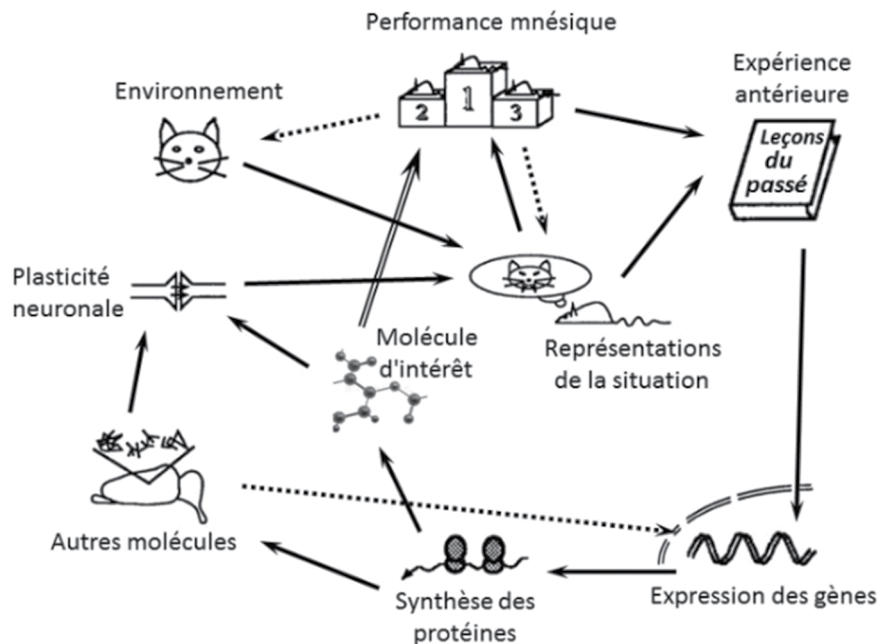


Figure 1.1 Exemple de modèle d'analyse.

Les connaissances apportées par la bibliographie permettent d'avoir une idée de la réponse au problème. Ces réponses potentielles sont appelées **hypothèses théoriques**. Les hypothèses sont des **affirmations hypothétiques**. Chaque hypothèse théorique pourra donner lieu à une étude expérimentale ou observationnelle destinée à la tester.

Cas d'étude

Sachant que telle enzyme est impliquée dans la plasticité neuronale et que la plasticité neuronale semble nécessaire à la mémorisation, je formule l'hypothèse théorique que « le blocage de cette enzyme perturbe la mémoire ».

Une hypothèse est une mise en relation de deux variables (ou plus).

- **Variable.** Lorsqu'une caractéristique différencie les individus statistiques d'une population, elle peut être définie sur chacun par une valeur, une mesure, un qualificatif. Ces qualificatifs

ou valeurs sont appelés **variables**. Étant donné que ces variables décrivent une caractéristique, elles sont parfois appelées **descripteurs**.

- **Variable aléatoire.** Une variable est dite aléatoire lorsqu'il est impossible de prévoir précisément la valeur ou le qualificatif qu'elle prendra chez un individu donné.

4. Le choix des méthodes

Pour tester l'hypothèse théorique, une étude expérimentale ou observationnelle est mise en place. Il faut alors choisir : une population, une méthode d'échantillonnage, du matériel pour collecter les données, un plan expérimental et une procédure d'analyse des données.

- **Population.** La population est l'ensemble des éléments ou des sujets sur lesquels le scientifique souhaite généraliser les conclusions.
- **Échantillon et échantillonnage.** Il est en général impossible d'avoir accès à toute la population sur laquelle porte l'étude. Le scientifique a donc recours à un échantillon qui est un sous-ensemble de cette population. Pour que l'étude soit valide, il faut que l'échantillon soit représentatif de la population. Un échantillon est représentatif de la population si ses caractéristiques (distribution, moyenne, dispersion...), à l'exception de la taille, sont les mêmes que celles de la population. L'action qui consiste à constituer l'échantillon le plus représentatif possible de la population est appelé **échantillonnage**.
- **Unité statistique.** Également appelée **individu statistique** (ou plus simplement « individu »), c'est l'entité de base de la population sur laquelle vont porter les mesures. Lorsque cette entité de base est un organisme vivant, on parle de **sujet** (êtres humains, animaux...), et dans le cas contraire (station d'observation, prélèvement...), d'**élément**. Dans la suite de l'ouvrage, nous utiliserons ces termes selon la situation. D'autres termes peuvent être rencontrés, certains pouvant prêter à confusion (« échantillon biologique »), ou d'autres étant plus spécifiques (« prélèvement »).
- **Plan expérimental.** Construire le plan expérimental consiste à choisir l'organisation des expériences qui vont permettre de tester l'hypothèse théorique. C'est à ce stade que l'on choisit les observations ou les mesures qui vont être effectuées et que l'on planifie les différents groupes à comparer (par exemple, pour vérifier l'effet d'une substance pharmacologique, il faut au minimum comparer un groupe qui reçoit la substance à un groupe qui ne la reçoit pas).

5. L'hypothèse opérationnelle

L'hypothèse opérationnelle est une reformulation de l'hypothèse théorique qui prend en compte les méthodes choisies (plan expérimental, procédure d'analyse des données...). Elle fait apparaître les résultats des comparaisons qui devront être effectuées. Si cette hypothèse est réfutée, il en sera de même pour l'hypothèse théorique.

Cas d'étude

« La performance mnésique du groupe qui n'a pas reçu la substance bloquante est supérieure à celle du groupe qui a reçu la substance. »

Cette hypothèse opérationnelle sous-entend que si cela est vrai, c'est que la substance a bloqué l'enzyme, que ce blocage a entraîné l'amnésie et donc que l'enzyme est impliquée dans la mémoire.

6. La confrontation à la réalité

Il s'agit de l'étape d'acquisition des données, qui va permettre de confronter les prédictions de l'hypothèse opérationnelle à la réalité. Nous vivons la partie de la recherche qui a le plus motivé nos études : observer les oiseaux, pipeter une solution, filmer le comportement de souris dans un

labyrinthe, interroger des gens sur leur régime alimentaire... Durant cette étape, il est préférable d'oublier les hypothèses formulées pour sa recherche. En effet, il est fréquent que le chercheur ait envie que ses hypothèses se vérifient, ce qui peut provoquer une anxiété au moment de l'étude et influencer les résultats.

Il est important, à chaque fois que c'est possible, d'effectuer cette étape en **double aveugle** : les sujets (humains) ne sont pas informés s'ils reçoivent le principe actif ou non, et le chercheur lui-même ne sait pas quelle substance il administre à ses sujets. Si le chercheur a « envie » que son hypothèse se vérifie ou s'il a une idée préconçue du résultat, la connaissance de la substance administrée pourrait influencer sa relation avec le sujet, ce qui pourrait influencer le résultat. Elle pourrait également influencer sa façon de manipuler les souris, le rendre plus sensible pour certaines observations ambiguës (comptage de cellules sur tranche de tissus, comptage de migrants dans un vol...). Enfin, depuis que les outils statistiques se diversifient et se complexifient, des approches statistiques différentes peuvent mener à des conclusions apparemment contradictoires et des interprétations différentes peuvent émerger de résultats identiques. Il devient donc préférable que les chercheurs effectuent le traitement des données également « en aveugles », c'est-à-dire sans connaître les caractéristiques des groupes ou des individus.

7. Le traitement des données

Les statistiques sont utilisées pour traiter les données. Il faut par exemple quantifier la différence entre les groupes puis se demander si la différence observée est due à l'effet du facteur étudié (ici la substance) ou aux aléas de l'échantillonnage.

- **Statistiques descriptives.** C'est la partie des statistiques qui permet de décrire les ensembles de données (paramètres de centrage comme la moyenne, distributions, paramètres de dispersion comme la variance...).
- **Statistiques inférentielles.** Sachant que le scientifique doit, en général, travailler sur des échantillons pour tirer des conclusions sur une population, les statistiques inférentielles lui permettent de dire s'il est possible de généraliser à la population, les propriétés observées sur les échantillons.
- **Statistiques exploratoires.** C'est la partie des statistiques qui vise à explorer des jeux de données, en général lorsque ceux-ci sont importants. S'il est possible de se représenter des phénomènes en deux, trois, voire quatre dimensions, la tâche devient beaucoup plus difficile au-delà. Les statistiques exploratoires permettent d'explorer et de déceler des phénomènes difficilement visibles sans leur aide. Ce type de statistiques développé dans la première moitié du xx^e siècle pour la psychométrie et le marketing se développe très rapidement en ce début de xxi^e siècle grâce à la puissance informatique. En biologie, il trouve principalement des applications en épidémiologie, en écologie, en génomique et en neurosciences.

8. L'interprétation des résultats

En général, les résultats sont nuancés : ils vont souvent globalement dans le sens prédit par l'hypothèse mais certains aspects de ces résultats semblent la contredire. Il faut donc interpréter les résultats.

9. La communication des résultats

Les résultats sont communiqués dans un rapport de stage, un article scientifique, lors d'un congrès ou sous toute autre forme de rapport d'étape.

Trois approches complémentaires : approche observationnelle, expérimentation et simulation

L'**enquête** PAQUID visait à étudier le vieillissement cérébral chez les personnes âgées afin d'identifier les facteurs de dégradation pathologique. Une cohorte de 4 134 personnes a été suivie de 1988 à 2003. Un des résultats de cette enquête fut d'établir un lien entre une consommation modérée mais régulière de vin rouge et une faible occurrence de maladie d'Alzheimer.

Le vin rouge protège-t-il de la maladie d'Alzheimer, ou sa consommation est-elle associée à un autre facteur (génétique, social...) protégeant de la maladie ? Des **expérimentations** chez l'animal sont venues confirmer l'effet de la consommation de vin rouge sur l'expression de la maladie. Elles ont permis ensuite de mettre en place des études pour préciser les molécules impliquées dans l'effet du vin sur les patients.

Pendant très longtemps, deux types d'approches ont fait progresser la biologie : l'étude observationnelle et l'expérimentation. Depuis l'émergence de l'informatique, la simulation vient constituer une troisième approche très complémentaire.

1. L'approche observationnelle

Le but de l'approche observationnelle est de recueillir des informations en situation la plus naturelle possible, de façon à influencer le moins possible ce qui va être observé. Dans les secteurs de l'épidémiologie et de la santé, l'étude observationnelle est appelée **enquête**. En écologie elle est souvent appelée échantillonnage, mais ce terme est une synecdoque, l'**échantillonnage** étant l'action de constituer les échantillons, il concerne également l'expérimentation. En raison du risque de confusion, nous parlerons d'étude observationnelle dans le cas général et en écologie, et d'enquête lorsque nous aborderons des problématiques liées à la santé.

Exemple

En éthologie, l'observation d'animaux se fera sur leur milieu naturel, mais en tentant de ne pas être vu, senti... par les animaux observés, sans quoi le comportement des animaux (ou leur absence) serait plus la conséquence de la présence du chercheur qu'une attitude naturelle.

Pour des raisons similaires, un écologue qui souhaite analyser l'eau d'un étang évitera de marcher dans l'eau qu'il s'apprête à prélever.

Le chercheur tente en général d'obtenir un grand nombre d'informations, si possible quantitatives, mais celles-ci peuvent également être qualitatives. Ces observations sont ensuite mises en relation et permettent d'émettre des hypothèses sur l'existence de liens de causalité entre les observations qui s'avèrent liées.

Cette approche est intéressante parce que le souci d'influencer le moins possible ce que l'on étudie permet d'approcher des systèmes complexes où de nombreuses variables interagissent. De plus, elle permet de découvrir de nouvelles informations parfois inattendues (comme par exemple l'effet bénéfique de la consommation de vin rouge sur le risque de démence).

Si elle permet d'observer des corrélations ou des co-occurrences entre variables, cette approche présente l'inconvénient de ne pas donner d'indications sur le sens des causalités ou leurs circuits (un lien de causalité pouvant être direct, indirect, uni- ou bidirectionnel...).

2. L'expérimentation

L'expérimentation vise à tester l'existence et le sens des relations de causalité. La stratégie de cette approche consiste à 1) choisir une relation de causalité à tester, 2) manipuler les variations du facteur qui nous intéresse et neutraliser les effets des autres facteurs potentiels, et 3) étudier les effets des manipulations sur une variable que l'on mesure ou que l'on observe. Si ces manipulations provoquent un effet, cela signifie que le facteur étudié influence bien cette variable.

Si cette définition est vraie en théorie, la pratique, en biologie, est parfois moins évidente concernant les conclusions au sujet des voies de causalité.

Je formule l'hypothèse qu'une substance facilite la mémoire grâce à l'action sur un mécanisme neuronal d'apprentissage de l'information. L'expérimentation montre que le groupe qui a reçu la substance a meilleure mémoire que celui qui a reçu l'excipient seul. Je conclus que l'hypothèse est vérifiée. Cette conclusion est légitime, cependant on ne peut pas exclure que l'effet observé soit dû à un effet de la substance sur l'attention, provoquant une amélioration des performances au test de mémoire. La causalité entre le facteur étudié et la mémoire est vérifiée mais la voie de causalité acceptée par la conclusion dépend de l'hypothèse formulée.

Ainsi, un lien de causalité démontré par une expérimentation peut être direct et unique, comme il peut être indirect ou pluriel (plusieurs voies peuvent être sollicitées simultanément). De plus un tel résultat n'exclut pas l'existence de boucles de rétroactions venant encore modifier le résultat observé.

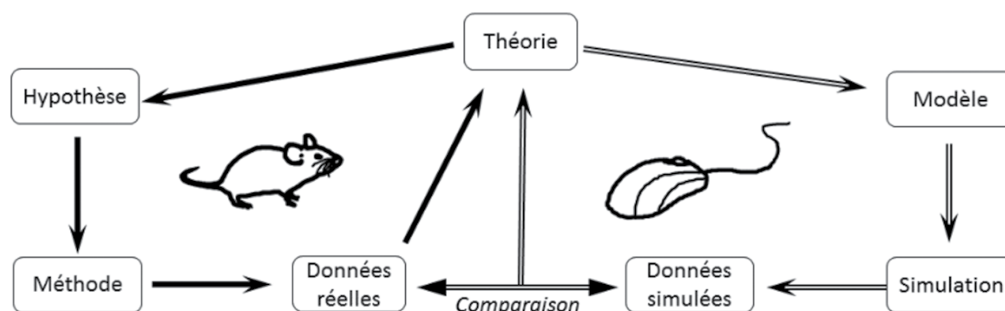


Figure 2.1 Approche observationnelle et expérimentation (flèches noires), et place de la simulation (flèches dédoublées).

3. La simulation

L'approche observationnelle ne donne que peu d'indications sur la nature des liens de causalité, et l'expérimentation prive le système étudié de nombreux phénomènes d'interaction potentiellement en œuvre dans la nature. En combinant ces deux approches, les conclusions risquent donc d'être trop partielles ou non applicables dans la nature.

La simulation consiste à formaliser le réseau de relations de causalité identifiées par la combinaison de l'approche observationnelle et de l'expérimentation et à les implémenter dans un ordinateur afin que celui-ci simule le fonctionnement du système. Si le comportement de ce modèle ressemble fortement à celui du système naturel, le système de causalité décrit est alors plausible.

4. En conclusion

Dans l'état actuel des connaissances et des développements technologiques, c'est par la combinaison de l'approche observationnelle, de l'expérimentation et de la simulation, que les meilleures descriptions du fonctionnement des systèmes biologiques sont possibles.

