LE COURS DE

BIOSTATISTIQUE

TOUT EN FICHES

LE COURS DE

BIOSTATISTIQUE

LICENCE 3, MASTER, ÉCOLES D'INGENIEURS

2e ÉDITION

Xavier Nogués, **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

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

ISBN 978-2-10-084286-5

Table des matières

Chapitre 3 Notions de base pour utiliser R en statistiques

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

Chapitre 5 Les tests paramétriques pour analyses univariées

Chapitre 6 Les tests non paramétriques pour analyses univariées

Chapitre 7 Les analyses multivariées

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 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 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 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 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**.

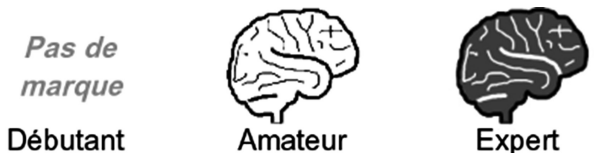
4. Le site d'auteur, pour quoi faire ?

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).

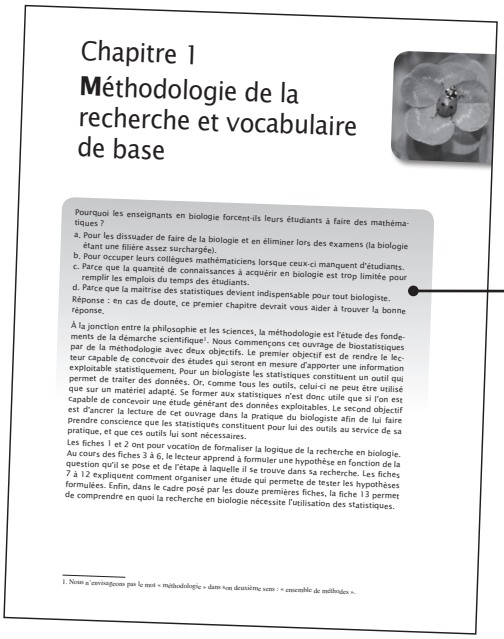
Les stats, ça peut être amusant. Vous trouverez sur ce site des jeux vous permettant de réviser et de mettre en application ce que vous avez appris dans l'ouvrage.

Et parce que tout le monde ne connait 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.



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 ».

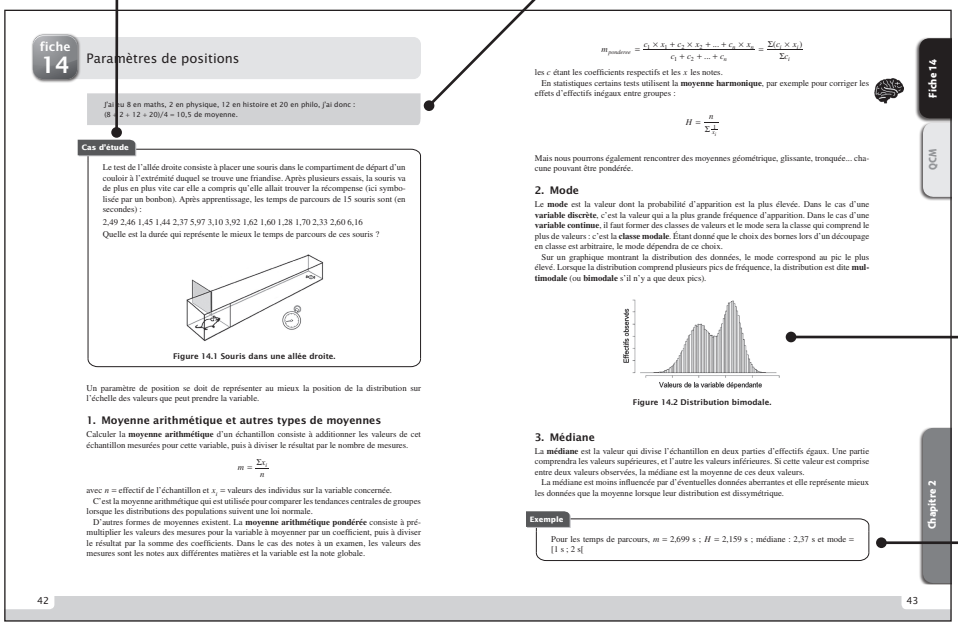
7 chapitres

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

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

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

FOCUS Comment gérer des valeurs suspectes ?

Comment définir une « valeur suspecte » ?

Vous avez demandé à des individus d'inscrire leur taille sur un questionnaire numérisé et de l'exprimer en centimètres. Des valeurs de 165 ou 188 n'éveillent aucune inquiétude. Une valeur de 1,78 (exprimée en centimètres comme demandé) est impossible et révèle certainement que l'individu a répondu en mètres. Une valeur de 1655 également impossible incite à penser que la touche « 5 » a été doublée mais le chercheur hésitera un peu plus entre « corriger » la valeur et l'éliminer. Enfin, une valeur de 246 le mettra probablement dans l'embarras. Cette valeur est suspecte, elle est physiologiquement plausible en cas de gigantisme, elle est aberrante si l'on accepte la définition de « valeur qui s'écarte du commun », mais on peut aussi se demander s'il ne s'agit pas d'une erreur de saisie pour 146. Faut-il éliminer, conserver ou corriger cette valeur ?

Plus généralement, aux deux extrémités d'une série de données, nous trouvons la valeur la plus grande et la valeur la plus petite : ce sont les valeurs limites. Si les données forment une sorte de « peloton », le chercheur est rassuré, tout va bien. Si au contraire, une ou deux valeurs s'écartent de ce peloton, ces valeurs deviennent suspectes. Parmi ces valeurs suspectes, certaines peuvent être impossible et seront considérées comme aberrantes.

Admettons que nous disposions d'une série de valeurs. Si ces valeurs sont classées par ordre croissant, chaque valeur peut être caractérisée par les distances entre elle et les deux autres valeurs qui l'encadrent (une seule distance pour les valeurs limites et plus de deux s'il y a des *ex aequo*). Intuitivement, nous avons tendance à considérer qu'une valeur limite est suspecte si sa distance à sa plus proche voisine est anormalement élevée et nous suivrons le même raisonnement pour des valeurs qui ne sont pas à la limite. Enfin, les valeurs suspectes seront toujours en nombre extrêmement faible.

Figure F2.1. a) pas de valeurs suspectes ; b) deux valeurs suspectes ; c) pas de valeurs suspectes

Quand les valeurs aberrantes apparaissent-elles ?

En règle générale, les valeurs aberrantes n'existent pas... dans les exercices de statistiques. Les valeurs aberrantes n'existent que dans la réalité. Vous entraînez 50 souris et mesurez les temps de résolution d'une tâche dans un labyrinthe. Ces rongeurs mettent environ 1 minute pour cette épreuve. Hélas, la souris n°37 a eu envie de dormir un peu et y a passé 20 minutes. La normalité de vos données qui s'annonçait exemplaire s'effondre.

Vous effectuez des coupes de tissus, un marquage immunohistochimique puis un comptage cellulaire. Sur vos coupes, le nombre de cellules marquées est toujours compris entre 400 et 600 sur la surface choisie, excepté sur deux coupes : l'une n'en compte que 23 et l'autre 325. Si pour la première vous êtes intimement convaincu qu'un mauvais pipetage ou un problème de réactif a eu lieu, que dire de la deuxième qui a un nombre de cellules marquées anormalement bas mais à la limite du plausible ?

Quelques méthodes de décision et exemples de traitement pratiqués

Méthodes non paramétriques

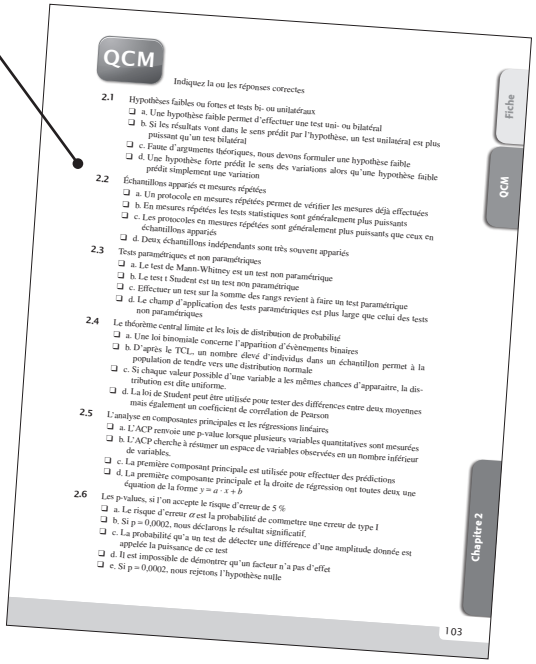
Pour traiter les données comportant des valeurs suspectes mais que l'on hésite à exclure, une première solution consiste à utiliser une méthode non paramétrique basée sur les sommes de rangs (Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman) qui sont plus robustes à ce genre de cas que les méthodes paramétriques.

Méthode du *cut-off*

Dans les laboratoires, nous trouvons la méthode du *cut-off*. Lorsque certains individus risquent de présenter des valeurs qui sortent du commun et que l'on sait que de telles valeurs ne révèlent plus ce que l'on cherche à mesurer (dans l'exemple des souris qui s'endorment, la valeur recueillie n'est plus une mesure de la motivation ou de la qualité de la mémoire), le chercheur décide, avant de commencer les expériences, que toute mesure qui dépassera une valeur déterminée limite sera remplacée par cette valeur limite. Cette méthode présente l'avantage d'être décidée *a priori*. Elle fausse cependant la normalité et contraint généralement à utiliser des tests non paramétriques. Une alternative consiste à supprimer ces valeurs.

100

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



QCM

Indiquez la ou les réponses correctes

2.1 Hypothèses faibles ou fortes et tests bi- ou unilatéraux
- ❑ a. Une hypothèse faible permet d'effectuer une test uni- ou bilatéral
- ❑ b. Si les résultats vont dans le sens prédit par l'hypothèse, un test unilatéral est plus puissant qu'un test bilatéral
- ❑ c. Faute d'arguments théoriques, nous devons formuler une hypothèse faible
- ❑ d. Une hypothèse forte prédit le sens des variations alors qu'une hypothèse faible prédit simplement une variation

2.2 Échantillons appariés et mesures répétées
- ❑ a. Un protocole en mesures répétées permet de vérifier les mesures déjà effectuées
- ❑ b. En mesures répétées les tests statistiques sont généralement plus puissants
- ❑ c. Les protocoles en mesures répétées sont généralement plus puissants que ceux en échantillons appariés
- ❑ d. Deux échantillons indépendants sont très souvent appariés

2.3 Tests paramétriques et non paramétriques
- ❑ a. Le test de Mann-Whitney est un test non paramétrique
- ❑ b. Le test t Student est un test non paramétrique
- ❑ c. Effectuer un test sur la somme des rangs revient à faire un test paramétrique
- ❑ d. Le champ d'application des tests paramétriques est plus large que celui des tests non paramétriques

2.4 Le théorème central limite et les lois de distribution de probabilité
- ❑ a. Une loi binomiale concerne l'apparition d'événements binaires
- ❑ b. D'après le TCL, un nombre élevé d'individus dans un échantillon permet à la population de tendre vers une distribution normale
- ❑ c. Si chaque valeur possible d'une variable a les mêmes chances d'apparaître, la distribution est dite uniforme.
- ❑ d. La loi de Student peut être utilisée pour tester des différences entre deux moyennes mais également un coefficient de corrélation de Pearson

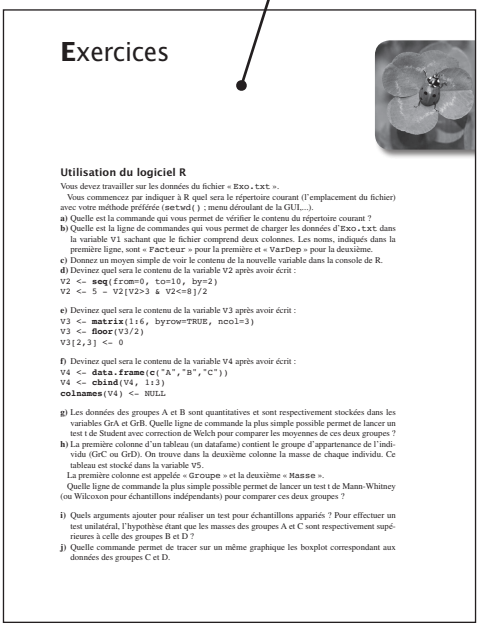
2.5 L'analyse en composantes principales et les régressions linéaires
- ❑ a. L'ACP renvoie une p-value lorsque plusieurs variables quantitatives sont mesurées
- ❑ b. L'ACP cherche à résumer un espace de variables observées en un nombre inférieur de variables.
- ❑ c. La première composant principale est utilisée pour effectuer des prédictions
- ❑ d. La première composante principale et la droite de régression ont toutes deux une équation de la forme $y = a \cdot x + b$

2.6 Les p-values, si l'on accepte le risque d'erreur de 5 %
- ❑ a. Le risque d'erreur α est la probabilité de commettre une erreur de type I
- ❑ b. Si p = 0,0002, nous déclarons le résultat significatif.
- ❑ c. La probabilité qu'a un test de détecter une différence d'une amplitude donnée est appelée la puissance de ce test
- ❑ d. Il est impossible de démontrer qu'un facteur n'a pas d'effet
- ❑ e. Si p = 0,0002, nous rejetons l'hypothèse nulle

Fiche | QCM | Chapitre 2

103

Des exercices pour s'entraîner en fin d'ouvrage



Exercices

Utilisation du logiciel R

Vous devez travailler sur les données du fichier « `Exo.txt` ».

Vous commencez par indiquer à R quel sera le répertoire courant (l'emplacement du fichier) avec votre méthode préférée (`setwd()` ; menu déroulant de la GUI,...).

a) Quelle est la commande qui vous permet de vérifier le contenu du répertoire courant ?

b) Quelle est la ligne de commandes qui vous permet de charger les données d'`Exo.txt` dans la variable `V1` sachant que le fichier comprend deux colonnes. Les noms, indiqués dans la première ligne, sont « `Facteur` » pour la première et « `VarDep` » pour la deuxième.

c) Donnez un moyen simple de voir le contenu de la nouvelle variable dans la console de R.

d) Devinez quel sera le contenu de la variable `V2` après avoir écrit :

```
V2 <- seq(from=0, to=10, by=2)
V2 <- 5 - V2[V2>3 & V2<=8]/2
```

e) Devinez quel sera le contenu de la variable `V3` après avoir écrit :

```
V3 <- matrix(1:6, byrow=TRUE, ncol=3)
V3 <- floor(V3/2)
V3[2,3] <- 0
```

f) Devinez quel sera le contenu de la variable `V4` après avoir écrit :

```
V4 <- data.frame(c("A","B","C"))
V4 <- cbind(V4, 1:3)
colnames(V4) <- NULL
```

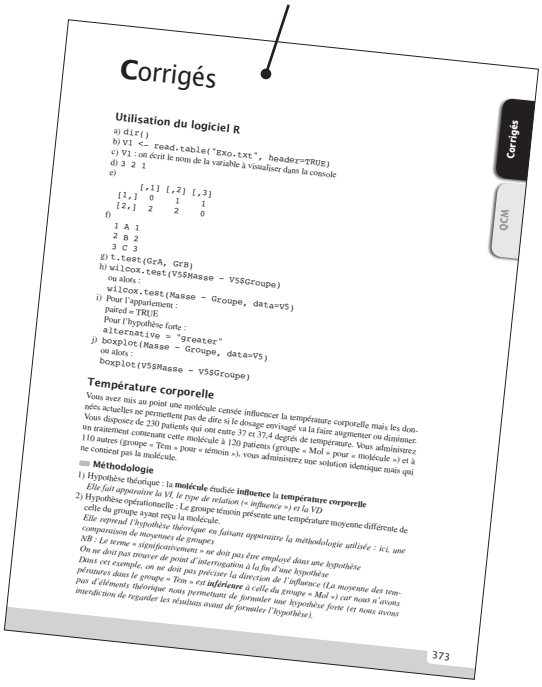
g) Les données des groupes A et B sont quantitatives et sont respectivement stockées dans les variables GrA et GrB. Quelle ligne de commande la plus simple possible permet de lancer un test t de Student avec correction de Welch pour comparer les moyennes de ces deux groupes ?

h) La première colonne d'un tableau (un dataframe) contient le groupe d'appartenance de l'individu (GrC ou GrD). On trouve dans la deuxième colonne la masse de chaque individu. Ce tableau est stocké dans la variable `V5`.
La première colonne est appelée « `Groupe` » et la deuxième « `Masse` ».
Quelle ligne de commande la plus simple possible permet de lancer un test t de Mann-Whitney (ou Wilcoxon pour échantillons indépendants) pour comparer ces deux groupes ?

i) Quels arguments ajouter pour réaliser un test pour échantillons appariés ? Pour effectuer un test unilatéral, l'hypothèse étant que les masses des groupes A et C sont respectivement supérieures à celle des groupes B et D ?

j) Quelle commande permet de tracer sur un même graphique les boxplot correspondant aux données des groupes C et D.

Les corrigés commentés



Corrigés

Utilisation du logiciel *R*

a) `dir()`

b) `V1 <- read.table("Exo.txt", header=TRUE)`

c) `V1` : on écrit le nom de la variable à visualiser dans la console

d) `3 2 1`

e)

```
  [,1] [,2] [,3]
[1,]    0    1    1
[2,]    2    2    0
```

f)

```
  1 A 1
  2 B 2
  3 C 3
```

g) `t.test(GrA, GrB)`

h) `wilcox.test(V5$Masse ~ V5$Groupe)`
ou alors :
`wilcox.test(Masse ~ Groupe, data=V5)`

i) Pour l'appariement :
`paired = TRUE`
Pour l'hypothèse forte :
`alternative = "greater"`

j)
```
boxplot(Masse ~ Groupe, data=V5)
```
ou alors :
```
boxplot(V5$Masse ~ V5$Groupe)
```

Température corporelle

Vous avez mis au point une molécule censée influencer la température corporelle mais les données actuelles ne permettent pas de dire si le dosage envisagé va la faire augmenter ou diminuer. Vous disposez de 230 patients qui ont entre 37 et 37,4 degrés de température. Vous administrez un traitement contenant cette molécule à 120 patients (groupe « Mol » pour « molécule ») et à 110 autres (groupe « Tem » pour « témoin »). vous administrez une solution identique mais qui ne contient pas la molécule.

Méthodologie

1) Hypothèse théorique : la **molécule** étudiée **influence** la **température corporelle**
Elle fait apparaître la VI, le type de relation (« influence ») et la VD

2) Hypothèse opérationnelle : Le groupe témoin présente une température moyenne différente de celle du groupe ayant reçu la molécule.
Elle reprend l'hypothèse théorique en faisant apparaître la méthodologie utilisée : ici, une comparaison de moyennes de groupes

NB : Le terme « significativement » ne doit pas être employé dans une hypothèse
On ne doit pas trouver de point d'interrogation à la fin d'une hypothèse
Dans cet exemple, on ne doit pas préciser la direction de l'influence (La moyenne des températures dans le groupe « Tem » est ***inférieure*** *à celle du groupe « Mol ») car nous n'avons pas d'éléments théorique nous permettant de formuler une hypothèse forte (et nous avons interdiction de regarder les résultats avant de formuler l'hypothèse).*

Corrigés | QCM

373

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 maitrise 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 ».

fiche 1

Le déroulement d'une recherche

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).

Cas d'étude

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.

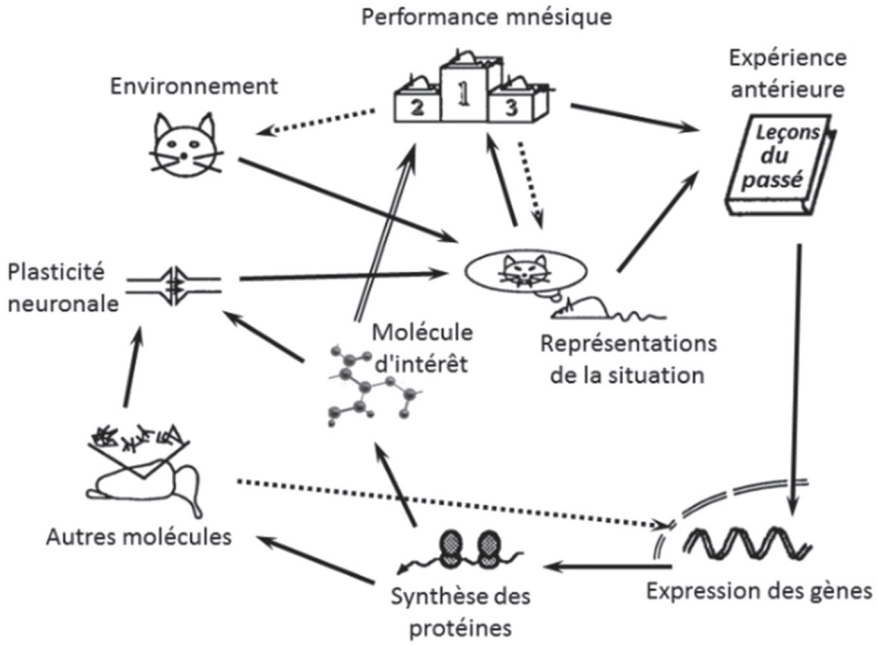


Figure 1.1 Exemple de modèle d'analyse.

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 migrateurs 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 connaitre 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.

fiche 2

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 conclue 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.

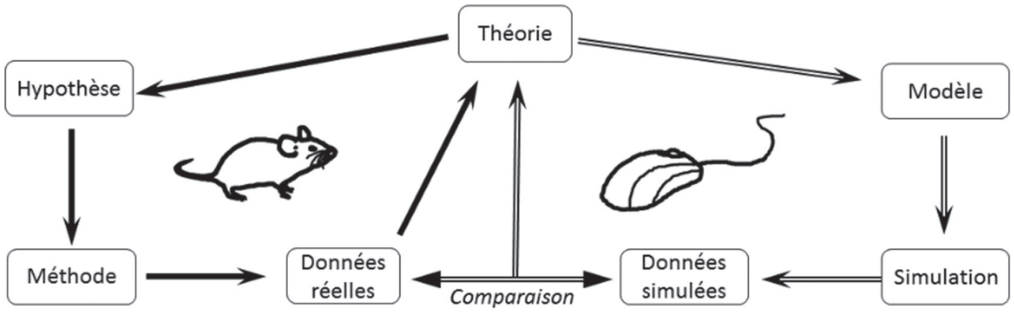


Figure 2.1 Approche observationnelle et expérimentation (flèches noires), et place de la simulation (flèches dédoublées).

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.

fiche 3

Le statut des variables dans la recherche

1. Variables indépendantes et dépendantes

www

Complément fiches 3 et 8

L'ensoleillement détermine, au moins partiellement, l'intensité du bronzage. L'ensoleillement est la variable indépendante, l'intensité du bronzage est la variable dépendante. Le soleil est un facteur de bronzage.

Les valeurs prises par une variable dépendante sont supposées dépendre de celles prise par des variables indépendantes. Le terme de « facteur » est synonyme de « variable indépendante ». Une variable est généralement influencée par un grand nombre de facteurs. Lors d'une expérimentation :

- les **facteurs principaux** sont ceux dont on choisit d'étudier l'effet ;
- les **facteurs secondaires** sont ceux qui sont susceptibles d'influencer la variable dépendante mais dont l'influence ne constitue pas le centre d'intérêt de l'étude actuelle. Il faut neutraliser leur effet afin d'isoler celui des facteurs principaux.

2. Variables manipulées, observées et invoquées

Nous étudions l'effet de l'âge sur la mémoire. Nous constituons un groupe de sujets jeunes et un groupe de sujets âgés. Les sujets sont sélectionnés aléatoirement mais il est impossible de les attribuer aléatoirement à l'un ou à l'autre des deux groupes.

Si l'expérience est menée sur des souris de laboratoire, cette procédure ne pose pas de problème : la population d'appartenance des souris jeunes et des souris âgées peut être considérée comme unique.

Si l'expérience est menée sur des humains, l'âge est associé à des changements de modes de vie qui peuvent avoir provoqué des modifications du style cognitif (démocratisation de l'informatique au début du XXIe siècle par exemple). Il est donc impossible de garantir que les facteurs susceptibles d'influencer ce que l'on cherche à mesurer sont les mêmes dans les deux sous-populations. Les conclusions de ce type d'étude ne pourront donc pas affirmer que les différences de performances mnésiques observées sont dues au vieillissement cérébral ou aux différences d'environnement dans lesquels les sujets se sont développés.

Dans une démarche d'expérimentation, le chercheur manipule le facteur principal. Manipuler le facteur « traitement pharmacologique » peut consister à administrer la substance testée à un groupe de sujets et l'excipient à un autre groupe. L'intérêt de manipuler un facteur est de pouvoir mesurer l'influence de cette manipulation sur la variable dépendante. Le plus souvent, un facteur est opérationnalisé en modalités (également appelées conditions expérimentales), ce qui mène à la constitution de groupes de sujets distincts, ou à la répétition d'une mesure sur les mêmes sujets dans des conditions expérimentales différentes. Lorsque c'est l'expérimentateur qui décide des modalités du facteur principal qui vont être utilisées, le facteur principal est qualifié de facteur principal systématique.

Lors d'une étude observationnelle, l'intérêt est d'observer des variables dans des conditions aussi proches que possible des situations naturelles. Comme aucune variable n'est manipulée ces études ne démontrent pas le sens des relations de causalité.

Dans une démarche d'expérimentation, certaines variables choisies comme facteurs principaux sont inhérentes aux individus et ne pourront pas être manipulées. Ces variables sont appelées

variables invoquées. Les études visant à tester les effets de facteurs sur des variables dépendantes et faisant appel à des variables invoquées sont qualifiées de **quasi expérimentales**. Les variables invoquées peuvent véhiculer des biais d'interprétation, en particulier si leur effet est transmis par l'intermédiaire d'une autre variable.

Fiche 7

3. Variables médiatrices, concomitantes et modératrices

Imaginons une étude visant à étudier l'effet de l'âge sur la prédisposition au bronzage. Nous constituons trois groupes : « enfants », « adultes » et « personnes âgées ». À la fin de l'été, la couleur de la peau des sujets de ces trois groupes est mesurée. Il y a de fortes chances de trouver que le bronzage diminue avec l'âge. L'âge a peut-être une influence. Cependant, le temps passé dehors en a très probablement une aussi : les enfants, en vacances jouent dehors, les adultes travaillent la semaine au bureau et les personnes âgées suivent probablement les consignes ministérielles leurs conseillant de rester à l'intérieur pendant les heures chaudes de la journée. Le temps passé dehors est une variable médiatrice dans la relation entre l'âge et le bronzage.

Le lien de causalité qui lie un facteur à une variable dépendante peut être direct ou indirect. S'il est indirect, c'est qu'une autre variable transmet cet effet. Cette nouvelle variable est appelée **variable médiatrice** ou **variable concomitante**.

La variable « crème solaire » (filtre UV) n'a pas d'effet sur le bronzage, qui n'est provoqué que par le soleil. Par contre, elle module l'effet du soleil sur le bronzage. C'est une **variable modératrice**. Notons que si une crème était capable d'amplifier l'effet bronzant du soleil (sans que ce soit une crème auto-bronzante) elle serait également appelée variable modératrice. Le terme « modératrice » doit être pris dans son sens « modulatrice » et non « réductrice » ou « diminutive ».

Une **variable modératrice** modifie l'influence qu'un facteur a sur une variable dépendante, mais n'influence pas directement la variable dépendante.

4. Variables latentes et variables manifestes

Comment mesurer une variable telle que l'intelligence ? Cette mesure passera par l'utilisation de tests d'évaluation. Le chercheur mesurera donc des variables de réponse du sujet ou de comportement de l'animal, mais jamais directement l'intelligence.

Les variables qu'il est possible de mesurer directement sont appelées **variables manifestes** (c'est le cas de la taille d'un individu, du nombre de cellules réactives...).

Une variable est **latente** lorsqu'il est impossible de la mesurer directement. Si elle détermine, au moins partiellement, des variables manifestes, sa valeur peut être estimée à partir des valeurs prises par celles-ci.

fiche 4

Les types d'hypothèses au cours d'une recherche

Comment amoindrir les symptômes de la maladie d'Alzheimer ? La recherche bibliographique suggère que les malades présentent un déficit d'acétylcholine dans le cerveau et que ce déficit semble être la conséquence d'une dégradation plus marquée chez les sujets malades que chez les sujets sains.
Problème : l'inhibition de la dégradation de l'acétylcholine cérébrale permet-elle de diminuer les symptômes de la maladie ?

1. L'hypothèse théorique

C'est une réponse hypothétique au problème, issue d'un raisonnement basé sur les données bibliographiques. Elle est formulée de façon à ce que son énoncé soit compréhensible par tout amateur éclairé. Sa formulation finale pourrait commencer par : « En théorie, si ma vision des choses est exacte, alors je peux formuler l'hypothèse théorique que... ».

Cas d'étude

Hypothèse théorique : « l'inhibition de la dégradation de l'acétylcholine facilite la mémoire chez les patients atteints de la maladie d'Alzheimer ».

2. L'hypothèse opérationnelle

L'hypothèse théorique formulée, il faut choisir une méthode pour la mettre à l'épreuve. La définition de cette méthode va mener à la formulation d'une hypothèse opérationnelle.

Choisir la méthode consiste à identifier la population statistique, les traitements à appliquer, la manière de les administrer, choisir des méthodes de mesure, etc.

L'hypothèse opérationnelle est une paraphrase de l'hypothèse théorique qui intègre les méthodes utilisées. La variable dépendante et les modalités (groupes ou conditions) du facteur principal y apparaissent clairement. Un moyen pratique de savoir si notre hypothèse opérationnelle est correctement formulée consiste à vérifier qu'elle conduit à la comparaison de valeurs.

Cas d'étude

À partir de l'hypothèse théorique formulée précédemment, il faut opérationnaliser :

- **L'acétylcholine et sa dégradation** : faut-il travailler sur des patients présentant des niveaux différents d'expression de la molécule ? Comment mesurer cette dégradation ? Faut-il administrer un inhibiteur de celle-ci ? Lequel ?
- **La facilitation de la mémoire** : par quelle méthode mesurer la mémoire ? Un test psychologique ? Lequel ? Quel indice ?
- **Chez les patients** : utilisera-t-on des sujets humains ? Pourra-t-on utiliser des souris ou un autre modèle biologique ? Quelle classe d'âge ?

Après un ensemble de choix, l'hypothèse opérationnelle devient : « Les scores au test de mémoire des sujets du groupe recevant l'inhibiteur de dégradation, sont en moyenne supérieurs à ceux des sujets du groupe témoin. »

La variable dépendante est le résultat au test de mémoire.

3. Les hypothèses de travail (ou postulats de travail)

Pour réussir le passage de l'hypothèse théorique à l'hypothèse opérationnelle, faut passer par des hypothèses de travail. Les hypothèses de travail sont tous les postulats qui ne font pas l'objet d'une vérification mais qui permettent cette opérationnalisation. Ces hypothèses de travail sont parfois triviales.

Cas d'étude

Si une validation montre qu'un test de mémoire reflète fidèlement les aptitudes mnésiques, le choix de ce test implique que cette validation soit applicable à la population que l'on étudie.

Une hypothèse de travail triviale est que la molécule choisie est en bon état lors de l'administration au sujet.

Toute opérationnalisation d'hypothèse théorique passe par des hypothèses de travail. Si le résultat de l'expérience est négatif, cela peut signifier soit que l'hypothèse théorique est fausse, soit que l'opérationnalisation était basée sur des hypothèses de travail fausses.

Il faut noter, que dans les laboratoires, le terme « hypothèse de travail » est souvent employé pour désigner des hypothèses théoriques, opérationnelles... Certains préfèrent utiliser le terme de **postulat de travail**. L'usage des termes étant arbitraire, l'important est d'avoir conscience que ces hypothèses existent, lors de l'interprétation des résultats.

4. Les deux hypothèses statistiques

Les deux hypothèses statistiques (hypothèse nulle et hypothèse alternative) sont une clé des statistiques inférentielles. La logique de ces statistiques consiste :

1. à formuler une hypothèse plus ou moins contraire aux hypothèses théorique et opérationnelle. Les hypothèses théorique et opérationnelle énonçant toujours que le facteur « a un effet sur », « accroît » ou « diminue » la variable dépendante, la première hypothèse statistique, appelée **hypothèse nulle** (H_0) dit que le facteur n'a pas d'effet. Pour les comparaisons de moyennes, elle est notée : $\mu_1 = \mu_2$ avec μ_1 et μ_2, les moyennes de populations représentées par les deux échantillons.
2. à formuler une **hypothèse alternative**, notée H_{Alt} ou H_1, qui est une formulation mathématique de l'hypothèse opérationnelle : $\mu_1 \neq \mu_2$, $\mu_1 < \mu_2$ ou $\mu_1 > \mu_2$.
3. à calculer la probabilité (exprimée par la p-value) d'obtenir un résultat au moins aussi important que celui observé si cette hypothèse nulle est vraie (le facteur n'a pas d'effet).
4. si cette probabilité est très faible (résultat trop improbable si H_0 est vraie), on rejette H_0.

Cas d'étude

Hypothèses statistiques

$$H_0 : \mu_i = \mu_T$$
$$H_1 : \mu_i > \mu_T$$

avec μ_i, moyenne du score mnésique d'une population traitée avec l'inhibiteur et μ_T, moyenne du score mnésique d'une population témoin (non traitée avec le principe actif).