UNIVERSITE DE LISALA

CENTRE INTERUNIVERSITAIRE DE RECHERCHE PLURIDISCIPLINAIRE (CIREP)
STATUT : UNIVERSITE PUBLIQUE
Web : www.cirep.ac.cd
Email : info@cirep.ac.cd

NOTES DU COURS DE L'EPIDEMIOLOGIE ET BIO STATISTIQUE DE BASE

OBJECTIFS DU COURS

Objectif général

L'objectif général de ce cours est de fournir aux étudiants les connaissances et les compétences nécessaires pour comprendre et appliquer les principes fondamentaux de l'épidémiologie et de la biostatistique dans le domaine de la santé publique.

Les objectifs spécifiques :

- ✓ Comprendre les concepts clés de l'épidémiologie, tels que l'incidence, la prévalence, les facteurs de risque, les études observationnelles et expérimentales, etc.
- ✓ Savoir utiliser des méthodes statistiques de base pour analyser des données épidémiologiques, telles que le calcul de taux, de risques relatifs, de tests d'hypothèses, etc.
- ✓ Être capable d'interpréter et de communiquer les résultats d'études épidémiologiques de manière critique et rigoureuse.
- ✓ Comprendre les principes éthiques liés à la recherche en épidémiologie et en biostatistique.
- ✓ Être en mesure de reconnaître et d'évaluer les biais potentiels dans les études épidémiologiques et d'appliquer des stratégies pour les contrôler.
- ✓ Savoir comment utiliser des logiciels statistiques pour l'analyse des données épidémiologiques.

Ces objectifs visent à fournir aux étudiants une base solide en épidémiologie et en biostatistique qui leur permettra de contribuer efficacement à la promotion de la santé publique et à la prise de décisions basées sur des données probantes.

Partie 1 : EPIDEMIOLOGIE

1. INTRODUCTION

L'épidémiologie est une des disciplines qui con• courent de façon importante aux progrès des connaissances en santé environnementale et en santé au travail. Son atout majeur est d'étudier les relations entre environnement et santé à un niveau global. Cette globalisation est faite selon deux «dimensions». En premier lieu, les résultats des études épidémiologiques portent sur des groupes de sujets, définis par exemple par leur exposition à telle ou telle condition environnementale ou de travail. Ce niveau d'analyse permet de dégager des moyennes et des tendances stables que l'observation individuelle ne rend pas perceptibles en raison de la variabilité importante entre les individus, qu'elle soit d'origine «biologique» ou qu'elle résulte de conditions d'exposition variables. Le second niveau de globalisation se situe au sein de l'individu lui- même. L'épidémiologie ne cherche pas à étudier ni à définir les mécanismes d'action des expositions sur l'organisme humain. Elle mesure leur effet «intégratif» par la survenue de pathologies ou, de façon plus générale, d'événements de santé. C'est le côté «boîte noire» de l'épidémiologie qui a été beaucoup débattu (Savitz, 1994; Skrabanek, 1994), et dont on voit bien les limites, mais aussi les avantages puisqu'à la fois il masque la connaissance des mécanismes biologiques fins et la rend non indispensable à la progression des connaissances des effets de l'environnement sur l'Homme.

L'épidémiologie n'est bien sûr pas seule et, comme dans la plupart des domaines scientifiques, les avancées en santé environnementale ont été et seront le fruit des résultats conjoints de plusieurs disciplines. C'est ainsi que la toxicologie apporte les «preuves» expérimentales de la nocivité de certaines substances que l'épidémiologie ne peut pas donner, mais elle les apporte sur l'animal ou dans des conditions «idéales», souvent

éloignées de la réalité des expositions humaines. À l'inverse, l'épidémiologie est proche des conditions réelles d'exposition, mais a parfois du mal à séparer les effets d'expositions survenant de façon conjointe.

Ce chapitre présente les méthodes épidémiologiques en mettant l'accent sur les mesures utilisées en épidémiologie et les principaux types d'étude. Il permettra aussi d'aborder succincte• ment les principes de l'analyse des données épidémiologiques et de leur interprétation. Il devrait permettre de comprendre les résultats des études épidémiologiques en ayant un regard critique sur leur méthodologie.

2. DOMAINE DE L'ÉPIDÉMIOLOGIE

L'épidémiologie est une discipline plutôt récente. Son champ d'intérêt s'accroît d'années en années, et sa méthodologie est encore en pleine évolution. Même si on a coutume de rappeler que déjà Hippocrate 400 ans avant Jésus-Christ s'intéressait aux déterminants de la maladie, nous devons reconnaître que l'épidémiologie comme science a commencé à voir le jour au XIX^{e} siècle (en particulier en Angleterre et en France) et qu'elle s'est véritablement développée au XX^{e} siècle en particulier, et de plus en plus à l'aide des statisticiens et de la révolution informatique.

L'épidémiologie est classiquement définie comme l'étude de la distribution des maladies et de leurs déterminants dans les populations humaines (Bouyer et coll., 1993; Rothman et coll., 1998). Cependant, son champ s'est rapidement étendu pour couvrir l'étiologie de l'ensemble des problèmes de santé ainsi que leur contrôle (Last, 1983). Les définitions modernes de l'épidémiologie incluent même l'évaluation des interventions et le support aux poli• tiques de santé. Cependant, dans ce chapitre, nous nous concentrerons

sur sa définition primaire, soit l'étude de l'apparition de la maladie dans les populations humaines et son apport à l'évaluation des risques environnementaux.

3. MESURES UTILISÉES EN ÉPIDÉMIOLOGIE

La définition de l'épidémiologie montre qu'il faut s'intéresser à deux types de mesures. D'une part, celles qui permettent de caractériser la distribution des maladies; il s'agit des mesures de risque et d'incidence (nous verrons la prévalence, le taux d'incidence et le risque cumulé). D'autre part, celles qui permettent de quantifier le lien entre une exposition et la maladie; il s'agit des mesures d'association (nous verrons principalement le risque relatif et l'oddsratio).

Les mesures qui caractérisent la distribution des maladies englobent des mesures de risque au sens strict, c'est-à-dire les probabilités d'être ou de devenir malade, et les mesures d'incidence qui indiquent la «vitesse d'apparition» des cas de maladie (pour le taux d'incidence) ou de décès (pour le taux de mortalité). Une probabilité est un nombre sans unité de mesure, compris entre 0 et 1, alors qu'un taux d'incidence est pourvu d'une unité et peut être supérieur à 1.

3.1 Prévalence

La façon la plus «naturelle» de mesurer la fréquence d'une maladie dans une population est de calculer la proportion de malades *présents* dans la population *à un instant donné.* Cette mesure est dénommée la prévalence, notée P, et définie par $P = \frac{M}{N}$ où M est le nombre de malades et N le nombre total de sujets (malades et non malades) de la population.

La prévalence intègre deux dimensions différentes de la fréquence de la maladie. D'une part, la durée de maladie (ou, du moins, la durée de la

présence d'un malade dans la population). D'autre part, la «vitesse d'apparition» de nouveaux cas de maladie au sein de la population (c'est-à-dire le taux d'incidence qui est défini plus bas).

La prévalence est surtout utile en santé publique lorsqu'on s'intéresse à la planification des ressources de santé nécessaires dans une population. En recherche étiologique, cet indiceest rarement utilisé, sauf dans quelques domaines particuliers, comme la périnatalité, ou dans le cas de pathologies fréquentes et sous- diagnostiquées (la dépression, par exemple).

L'estimation de la prévalence sur un échantillon est notée p_0, et l'intervalle de confiance correspondant est, lorsque la taille n de l'échantillon est assez grande, $p_0 \pm z_{\alpha} \sqrt{\frac{p_0(1-p_0)}{n}}$ *

3.2 Taux d'incidence

Par définition, le taux d'incidence (TI) de la maladie est la «vitesse de production» de nouveaux cas au cours d'un intervalle de temps. Il est égal au nombre de nouveaux cas survenus dans cet intervalle de temps divisé par la taille de la population à risque. La «taille» de la population se mesure en «personnes-temps». Il s'agit de la somme des durées, cumulées sur l'ensemble de la population à l'étude et sur l'ensemble de la durée de suivi, pendant laquelle les sujets sont susceptibles d'être enregistrés comme de nouveaux cas. La population peut être ouverte, c'est-à-dire que des sujets peuvent y entrer ou en sortir au cours de la période de suivi. Dans ce cas, les durées cumulées correspondent aux périodes où le sujet est présent dans la population. L'unité de mesure la plus fréquente en épidémiologie est la personne-année, au point que le terme «personnes-années» est souvent employé comme terme générique à la place de

«personnes-temps». Cependant, si l'unité de mesure du temps est le

mois, la semaine ou le jour, on peut être amené à compter en personnes — mois — semaines ou jours.

La définition formelle du taux d'incidence est $TI = \frac{m}{PA}$ où m est le nombre de nouveaux cas

pendant la période [t, t+□t[et PA le nombre de personnes-années cumulé sur la période [t, t+□t[. Le taux d'incidence n'est pas une probabilité. En particulier, on exprime sa valeur en nombre de cas par personnes-temps. Par exemple, si l'unité de temps est l'année, TI = 10^{-4} se lit 1 cas pour 10 000 personnes-années (ou 10 cas pour 100 000 personnes-années).

Lorsqu'on estime le taux d'incidence sur un échantillon, son intervalle de confiance est donné par : $TI \pm z_{\alpha/2}\sqrt{\frac{TI}{PA}}$

La plupart des calculs concernant le taux d'incidence sont sous-tendus, sur le plan mathématique, par l'existence d'une variable T qui mesure la date de survenue de l'événement étudiée (par exemple, la maladie ou le décès), etpar la notion d'incidence instantanée (notée □(t)) qui lui est associée. La variable T a la propriété d'être «censurée» pour certains sujets, ceux pour lesquels l'événement n'est pas survenu pendant la période de suivi et pour lesquels la seule information est que T est supérieure à leur durée de suivi. Ces notions dépassent le cadre de ce livre et peuvent être trouvées ailleurs (par exemple: Hill et coll., 1990; Bouyer et coll., 1993; Estève et coll., 1993) sous le nom de «données de survie» (en référence à l'étude de la mortalité qui a été leur premier champ d'appli• cation) ou de «données censurées».

* $z_{\square/2}$ est le percentile (1-□/2) de la loi normale centrée réduite. Le plus souvent, on prend □ = 5 % et ona alors $Z_{\square/2}$ = 1,96.

Notons enfin que, lorsque la population est stationnaire — c'est-à-dire

lorsque aucune des caractéristiques de la maladie (telle que taux d'incidence ou prévalence) n'évolue au cours du temps —, la prévalence et le taux d'incidence sont liés par la relation $P = \frac{TId}{1 + TId}$ où d est la durée moyenne de la maladie. Dans le cas, fréquent, où TId est petit, cette relation est approchée par P = TId qui illustre bien les deux dimensions de la prévalence indiquée précédemment.

3.3 Risque cumulé de maladie (ou incidence cumulée)

Par définition, le risque cumulé est la probabilité de devenir malade au cours d'une période fixée. Cela nécessite donc de préciser la durée de la période considérée. Le calcul est facile dans une population fermée (sans entrée ni sortie de sujets) et sans sujets perdus de vue: il suffit de diviser le nombre de nouveaux cas par le nombre de sujets non malades au début de la période. Sinon (population ouverte ou sujets perdus de vue), le risque cumulé de maladie pendant la période □t est donnée par R(□t) = 1- exp {-TI □t}(Bouyer et coll., 1093). Il s'agit alors d'une probabilité conditionnelle (à l'absence de cen• sure) pendant la période à l'étude.

Si le taux d'incidence est petit (ou plus précisément si TI □t est petit), et seulement dans ce cas, cette expression est approchée par R(□t) = TI □t. On voit donc que, si □t = 1 (1 an si l'unité est l'année), on obtient R = TI, ce qui est commode à retenir, mais source de confusion entre les notions de taux d'incidence et d'incidence cumulée.

Lorsqu'on estimé le risque cumulé sur un échantillon, l'intervalle de confiance est donné par [1- exp(-a □t) ; 1- exp(-b □t)], où [a ; b] est l'intervalle de confiance de TI.

La formule permettant de calculer R(□t) en fonction de TI qui vient d'être donnée nécessite que TI soit constant sur l'intervalle de temps At. Lorsqu'on veut calculer le risque cumulé de maladie sur une longue

période, cette hypothèse n'est, en général, plus satisfaite. On doit tenir compte, par exemple, du fait que l'âge augmentant, l'incidence de la maladie augmente aussi. On est alors conduit à découper la période sur laquelle on veut calculer le risque cumulé de maladie en sous-périodes au sein desquelles on peut supposer le taux d'incidence constant. Nous noterons p le nombre de sous-périodes,

Δt_k la durée de la k^e sous-période (les Δt_k ne sont pas nécessairement tous égaux) et Tl_k le taux d'incidence correspondant. On montre alors que le risque de maladie pendant l'ensemble de la période est*

$$R = 1 - \prod_{k=1}^{p} (1 - R_k) = 1 - \prod_{k=1}^{p} (\exp\{-TI_k \Delta_{t_k}\}) = 1 - \exp\left\{-\sum_{k=1}^{p} TI_k \Delta_{t_k}\right\}.$$

Si les taux d'incidence TI_k sont petits (ou plus précisément si les TI_k Δt_k sont petits), cette expression se simplifie en). $R = 1 - \prod_{k=1}^{p} (1 - TI_k \Delta_{t_k}$

Cas particulier de la mortalité

Lorsqu'on s'intéresse à la survie, l'incidence de décès est appelée mortalité, et les mêmes mesures que précédemment peuvent être décrites. Les plus connues sont le *taux de mortalité* qui est le taux d'incidence de décès et *la létalité* qui est l'incidence cumulée ou le risque de décès parmi les personnes atteintes d'une maladie au cours d'une période donnée.

3.4 Mesures d'association

L'étude de l'association entre une exposition E (ou facteur de risque) et la maladie M est une des étapes majeures de la recherche des facteurs étiologiques des maladies. Plusieurs questions complémentaires se posent. D'une part, celle de l'existence même d'un lien statistique entre l'ex• position et la maladie. D'autre part, celle de la mesure de la force du lien entre E et M qui permet de quantifier l'accroissement du risque en fonction de l'exposition au facteur de risque; cela nécessite de

choisir une mesure d'association. Enfin, bien sûr, celle de l'interprétation de l'association lorsque celle-ci a été établie. De façon générale, une mesure d'association est une mesure descriptive: elle permet de mesurer l'association statistique entre deux variables (une exposition et la fréquence d'une maladie), mais ne permet pas directement de savoir s'il y a un lien de cause à effet entre elles. Nous reviendrons sur cette question à la fin du chapitre. Il nous arrivera, cependant, pour alléger la présen-

* Dans les expressions qui suivent, $\prod_{k=1}^{p}(1-R_k)$ indique le produit des $(1-R_k)$ pour k variant de 1 à p, et $\sum_{k=1}^{p}$ indique la somme.

tation, d'employer le terme «effet de E» au lieu de «mesure de l'association entre E et M». Sauf indication explicite du contraire, cela n'aura pas de sens causal.

Dans ce qui suit, le facteur de risque et la maladie sont caractérisés par des variables dichotomiques (ayant uniquement deux valeurs: présence ou absence). Pour le facteur de risque, les deux catégories sont notées E+ pour les sujets exposés et E- pour les sujets non exposés. Pour la maladie, les malades sont notés M+ et les non-malades sont notés M-. Nous nous intéresserons principalement au cas où l'on mesure le *risque* ou probabilité d'être atteint, c'est-à-dire la probabilité d'être malade à un moment donné (ou prévalence) ou la probabilité de le devenir au cours d'une période fixée. Ce risque sera noté R. Nous donnerons quelques éléments concernant le cas où on con•sidère non pas le risque lui-même, mais un taux d'incidence ou une incidence instantanée.

Modèle additif et modèle multiplicatif

Soit R_1 le risque de maladie chez les sujets exposés au facteur de risque et R_0 le risque de maladie chez les sujets non exposés.

Les deux principaux types d'indices pour quantifier l'association entre la maladie et le facteur de risque sont l'excès de risque (modèle additif) : □ = $R_1 - R_0$; et le risque relatif (modèle multiplicatif): $RR = \frac{R_1}{R_0}$

Bien entendu, d'autres indices sont envisageables. Certains, tout en donnant une expression numérique particulière de la relation entre l'exposition et la maladie, se ramènent finalement à l'un des deux précédents. C'est le cas de la différence relative $\frac{R_1 - R_0}{R_0}$ qui apporte la même

information que le risque relatif puisqu'elle est égale à RR-1. Cette dernière est aussi appelée excès de risque relatif; il s'agit d'une mesure du risque fréquemment utilisée en évaluation de risque (Krewski et coll., 1999). D'autres sont de nature plus différente. Le plus utilisé est l'odds ratio défini par OR= $\frac{R_1/(1-R_1)}{R_0/(1-R_0)}$. Il s'agit du rapport de la quantité $\frac{R}{1-R}$ calculée chez les exposés à sa valeur chez les non-exposés. La quantité est appelée «odds» en anglais; $\frac{R}{1-R}$

4. TYPES D'ENQUÊTES ÉPIDÉMIOLOGIQUES

Les enquêtes épidémiologiques se divisent en plusieurs grandes catégories comme le schématise la figure 4.1 . Une première division sépare les enquêtes d'observation des études expérimentales. Les enquêtes d'observation se divisent ensuite en enquêtes descriptives et étiologiques qui ont des objectifs différents comme leur nom l'indique. Enfin, les enquêtes étiologiques se divisent en trois catégories principales selon leur méthodologie: étude de cohorte, étude cas-

témoins et étude transversale. Comme toute classification, celle que nous présentons comporte une part d'arbitraire, et certaines enquêtes ont du mal à y trouver leur place. D'autres types d'enquêtes peuvent être décrits; il s'agit en fait d'enquêtes «hybrides» par rapport aux trois précédentes (Kleinbaum et coll., 1982).

4.1 Études expérimentales

De façon générale, on qualifie d'étude expérimentale toute enquête où l'attribution de l'exposition est contrôlée par l'investigateur, c'est- à-dire que ce dernier a pu choisir quels sujets sont exposés et lesquels ne le sont pas, ainsi que le type d'exposition. De façon générale, l'étude expérimentale s'apparente à une étude de cohorte dans laquelle l'exposition est sous le contrôle du chercheur. La capacité à montrer que les associations mesurées sont de nature causale est meilleure dans les situations expérimentales que dans les enquêtes d'observation. L'avantage, comparativement aux méthodes d'observation, est alors de pouvoir isoler l'exposition étudiée des autres facteurs de risque (facteurs de confusion) afin que tout changement, dans l'effet observé, puisse être attribué unique• ment à l'exposition. Cette attribution causale est d'autant plus aisée qu'il y a tirage au sort et que la comparabilité des groupes a été maintenue tout au long de l'enquête. On parle alors d'essai «randomisé».

Comme la méthode entraîne une exposition choisie par le chercheur et non par le participant, ce type d'étude nécessite que des conditions soient réunies pour garantir la sauvegarde des droits de la personne. En particulier, les participants doivent être clairement informés des avantages et inconvénients résultants de leur participation, ils doivent signer un formulaire de *consentement* confirmant leur volonté à participer et aussi avoir la possibilité en tout temps

d'abandonner l'étude sans aucun préjudice. En fait, compte tenu de sa nature (visant à exposer volontairement des sujets), cette méthode est utilisée en épidémiologie principalement pour étudier l'effet d'une intervention à visée préventive.

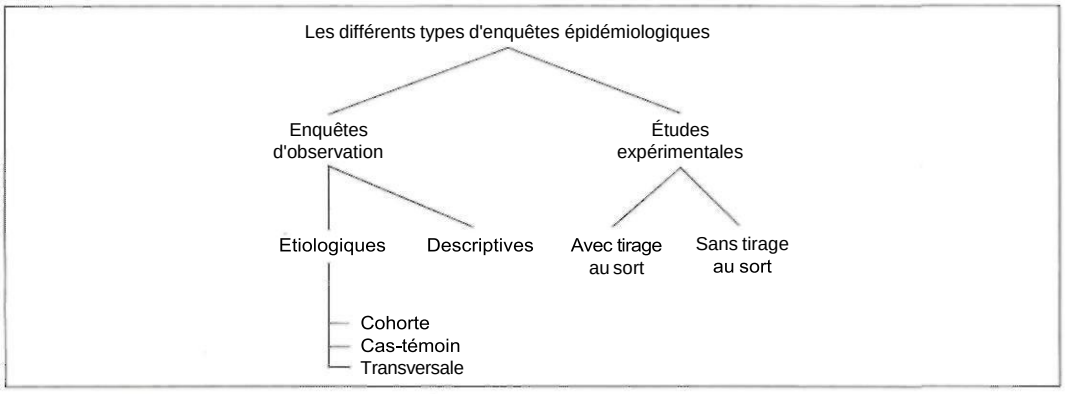


Figure 4.1 Les différents types d'enquêtes épidémiologiques

Essai clinique randomisé

Il s'agit de la méthode de référence pour évaluer l'efficacité d'un traitement, qu'il soit médicamenteux ou non. Afin d'isoler l'effet de la composante active du traitement (par rapport à l'effet psychologique attendu, appelé communément
«effet placebo»), il est courant d'utiliser dans le groupe ne recevant pas le traitement une exposition de type *placebo* ou intervention simulée*. Puisque la connaissance du traitement utilisé peut influencer le diagnostic de la maladie fait par le médecin, mais aussi le respect du protocole par le patient, il est usuel que ni le médecin et ni le patient n'aient connaissance du véritable traite• ment. Cette méthode est ainsi parfois appelée *méthode en double aveugle.* On trouvera plus de détails sur ce type d'études dans des ouvrages spécialisés (Laplanche et coll., 1986; Bouvenot et Vray, 1999).

En contrepartie de ses avantages pour évaluer l'efficacité d'un

traitement, la grande standardisation des essais randomisés diffère des situations généralement rencontrées dans la réalité où les expositions ne sont pas distribuées au hasard ni de façon constante au cours du temps. Le recours à des expériences randomisées est en fait assez peu fréquent en épidémiologie environnementale. Elles se rencontrent principalement dans les essais thérapeutiques destinés à étudier l'efficacité des traitements. À titre d'exemple, l'essai cli• nique a été utilisé pour évaluer l'efficacité des traitements (médicamenteux ou non) de l'intoxication au plomb (Lanphear et coll., 1999; O'Connor et Rich, 1999). Ce type d'étude est aussi parfois utilisé pour évaluer l'effet possible d'une exposition environnementale à des niveaux inférieurs aux normes en vigueur, particulièrement dans le cas de courte exposition. Ainsi, on peut citer le cas de l'effet des expositions aux champs magnétiques d'extrêmes basses fréquences (Selmaoui et coll., 1996).

Essai préventif

Il s'agit d'une étude expérimentale s'adressant généralement à des personnes bien portantes, dans le but de prévenir l'apparition de maladies éventuelles. Ce type de méthode est utilisé fréquemment en médecine préventive afin d'évaluer certaines interventions préventives. C'est le cas en particulier de l'évaluation de l'efficacité des vaccins. Peu d'essais préventifs à notre con• naissance ont été réalisés en santé environ• nementale. On peut cependant citer l'exemple de l'étude de l'efficacité de suppléments de vita• mines pour la prévention du cancer du poumon (Omenn et coll., 1996).

Essai communautaire

Le traitement ou l'exposition se fait alors au niveau d'une

communauté (ville, village, école, etc.) plutôt qu'au plan individuel. L'intervention est habituellement de nature préventive, et l'é• tude vise à évaluer son efficacité pour réduire l'apparition de problèmes de santé. L'essai peut être véritablement de nature expérimentale (l'ex• position étant choisie par le chercheur et son allo• cation étant faite de façon aléatoire). L'essai peut aussi être de nature *quasi expérimentale.* Dans ce dernier cas, le choix de l'exposition peut être décidé selon différents critères de faisabilité. Cependant, on essaie toujours d'avoir un groupe sans traitement le plus comparable possible au groupe avec traitement. Le consentement de participation est habituellement donné par les responsables politiques ou administratifs de la communauté sous étude. Des procédures doivent cependant être établies pour que les individus participants soient informés des objectifs de l'étude. On retrouve dans la littérature plusieurs exemples de l'application de cette méthode en santé environnementale, dont celui historique de la fluoruration des eaux de consommation (Arnold et Dean, 1956).

4.2 Enquêtes descriptives

La distinction généralement faite entre études descriptives et étiologiques comporte une large part d'arbitraire. Les secondes doivent fournir des arguments essentiels en faveur ou à l'encontre de l'hypothèse d'un rôle étiologique des facteurs de risque étudiés alors que les premières ont pour objectif principal de fournir des statistiques permettant de connaître l'état sanitaire de la population (fréquence de la maladie, tendances temporelles ou géographiques) sans le mettre explicitement en rapport avec des facteurs de risque.

Partie 2 : ÉLOGE DE « BIOSTATISTIQUE, UNE APPROCHE INTUITIVE »

Biostatistique, une approche intuitive est un beau livre qui a beaucoup à apprendre aux biologistes expérimentaux de toutes sortes. Contrairement à d'autres ouvrages de statis- tique que j'ai vus, ce livre inclut des analyses critiques étendues et soigneusement étayées sur les dangers des comparaisons multiples, des mises en garde concernant les erreurs fréquentes et évitables dans l'analyse des données, un examen des hypothèses sous- jacentes aux divers tests, une priorité aux intervalles de confiance par rapport aux P-valeurs, des éclaircissements concernant les raisons qui font que la signification statis- tique est rarement nécessaire dans un travail scientifique et une explication claire de la régression non linéaire (fréquemment utilisée en laboratoire mais rarement exposée dans les livres).

PARTIE A

Introduction à la statistique

CHAPITRE 1

Statistique et probabilité ne sont pas intuitives

> Si un événement quelconque a 50 % de chances de se produire, alors 9 fois sur 10, il se produira.

L'adjectif *intuitif* a deux significations. La première est *facile à utiliser et à comprendre*. C'est l'objectif de ce livre, d'où son titre. L'autre est *instinctif*, c'est-à-dire *agir en fonction de ce que l'on sent être vrai, même sans raison*. Ce chapitre amusant (vraiment amusant!) montre combien nos instincts nous induisent souvent en erreur en matière de probabilités.

NOUS AVONS TENDANCE À PASSER DIRECTEMENT AUX CONCLUSIONS

Une petite fille de trois ans disait à son copain: « tu ne peux pas devenir médecin, seules les filles peuvent devenir médecins ». Pour elle, c'était évident puisque les trois médecins qu'elle connaissait étaient toutes des femmes.

Lorsque ma fille aînée avait 4 ans, elle avait « compris » qu'elle venait de Chine et avait été adoptée tandis que son frère « venait du ventre de sa maman». Lorsque nous lui avons lu un livre à propos d'une femme tombant enceinte et donnant naissance à une petite fille, sa réaction a été : « C'est stupide. Les filles ne viennent pas du ventre de leur maman. Les filles viennent de Chine». Avec n = 1 dans chaque groupe, elle avait tiré une conclusion générale. À l'instar de beaucoup de scientifiques, dès qu'une nouvelle donnée est venue contredire sa conclusion, elle a spontanément remis en question cette nouvelle donnée plutôt que la validité de sa conclusion.

La tendance à extrapoler d'un échantillon à toute une population est ancrée dans nos cerveaux et a même été observée chez des bébés de 8 mois (Xu & Garcia, 2008). C'est pour résister à la tentation de tirer hâtivement des conclusions erronées à partir de données limitées que les scientifiques ont besoin des statistiques.

NOUS AVONS TENDANCE À ÊTRE TROP CONFIANTS

Les gens peuvent-ils vraiment juger dans quelle mesure ils sont confiants ? Vous pouvez tester votre propre capacité à quantifier l'incertitude en utilisant un test mis au point par Russo et Schoemaker (1989). Répondez aux questions de ce test par un intervalle dont vous êtes certain à 90 % qu'il contient la réponse correcte. N'utilisez pas Google pour trouver la réponse. N'abandonnez pas sous prétexte que vous ne savez pas. Bien sûr, vous ne connaissez pas les réponses exactement ! L'objectif n'est pas de fournir des réponses

exactes, mais plutôt de quantifier correctement votre incertitude et de proposer des intervalles qui selon vous auraient 90 % de chance d'inclure la réponse correcte. Si vous n'avez pas d'idée, répondez avec un intervalle super large. Par exemple, si vous n'avez vraiment pas d'idée de la réponse à la première question, choisissez l'intervalle de 0 à 120 ans de manière à être sûr à 100 % qu'il contiendra la réponse correcte. Mais essayez de répondre par des intervalles les plus étroits possibles dont vous êtes sûr à 90 % qu'ils contiennent la réponse correcte :

- L'âge de Martin Luther King Jr. à sa mort.
- La longueur du Nil, en miles ou kilomètres.
- Le nombre de pays de l'OPEP.
- Le nombre de livres de l'Ancien Testament.
- Le diamètre de la lune, en miles ou en kilomètres.
- La masse d'un Boeing 747 vide, en livres ou en kilos.
- L'année de naissance de Mozart.
- La durée de gestation d'un éléphant d'Asie, en jours.
- La distance de Londres à Tokyo, en miles ou en kilomètres.
- La plus grande profondeur océanique connue, en miles ou en kilomètres.

Comparez vos réponses avec les réponses correctes qui se trouvent à la fin de ce chapitre. Si vous atteignez l'objectif de 90 % de degré de confiance, vous devriez avoir créé 9 intervalles qui incluent la réponse correcte et un qui ne l'inclut pas.

Russo et Schoemaker (1989) ont proposé ce test à plus de 1 000 personnes et ont conclu que 99 % d'entre elles étaient trop confiantes. Le but était de déterminer des intervalles qui incluaient la réponse correcte dans 90 % des cas, mais la plupart des gens créaient des intervalles trop étroits qui incluaient seulement 30 à 60 % de réponses correctes. Des tests semblables ont été effectués auprès d'experts interrogés sur des sujets de leurs domaines et les résultats ont été similaires.

Puisque la tendance est d'être trop sûr de soi, les scientifiques doivent faire appel aux méthodes statistiques pour évaluer correctement la confiance.

NOUS VOYONS DES STRUCTURES DANS DES DONNÉES ALÉATOIRES

Le tableau 1.1 présente les données simulées de 10 joueurs de basket (un par ligne) qui tirent 32 paniers. Un X indique un panier réussi et un –, un manqué. Voyez-vous une structure aléatoire ? Ou au contraire, y voyez-vous des séries non aléatoires ? Observez le tableau 1.1 avant de continuer.

La plupart des gens y voient des séries de tirs réussis et concluent au caractère non aléatoire. Alors que le tableau a été construit de façon aléatoire : chaque tir a 50 % de chance d'être réussi (X) et 50 % de chance d'être manqué (–) indépendamment des tirs précédents. Nous voyons des regroupements peut-être parce que nos cerveaux en se développant sont devenus capables d'identifier des structures et font cela très bien. Même si cette compétence peut avoir été très utile à nos ancêtres pour éviter des prédateurs ou des plantes vénéneuses, il est important d'avoir conscience de ce handicap mental inné. Il faut de la rigueur statistique pour éviter d'être dupé en observant des structures apparentes parmi des données aléatoires.

–	–	X	–	X	–	X	X	X	–	–	–	–	X	X	X	–	X	X	–	X	X	–	–	–	X	X	–	–	–	X	X
X	–	–	X	–	X	X	–	–	X	X	–	–	X	–	X	–	X	–	–	–	X	X	X	X	–	–	X	X	–	–	–
X	X	X	X	–	X	X	–	X	–	X	–	X	X	X	–	–	–	–	–	X	–	X	–	X	X	X	–	–	–	–	X
–	X	–	X	–	–	X	X	–	X	X	–	X	X	–	–	X	X	X	X	–	–	–	–	X	X	–	X	–	X	–	–
–	X	–	X	–	X	X	–	–	–	X	X	–	–	–	–	–	–	X	–	X	–	X	–	–	X	–	–	X	–	X	X
–	–	X	X	X	–	X	–	X	–	–	–	X	X	X	X	–	X	X	X	X	–	–	–	–	–	–	X	X	–	X	X	X
X	–	–	X	X	–	–	X	X	X	X	–	X	X	X	X	–	–	X	–	–	X	X	X	X	X	–	X	X	X	–	–	–
X	–	X	–	–	–	X	X	X	X	X	–	–	X	X	X	–	X	X	–	X	X	X	–	X	X	–	X	–	–	X	–	X
X	X	X	–	–	X	X	X	X	X	–	X	–	X	–	X	X	–	X	–	X	X	X	X	–	X	X	–	X	X	X	X	X
–	–	–	X	X	X	–	–	X	X	X	–	X	X	X	X	–	–	X	–	–	X	–	X	X	X	X	X	–	–	–	X	–

Tableau 1.1. Des structures aléatoires qui ne semblent pas aléatoires

Le tableau 1.1 présente les résultats de 10 joueurs de basket (un par ligne) qui tirent 32 paniers. Un X indique un lancer réussi et un –, un lancer manqué.

NOUS NE NOUS RENDONS PAS COMPTE QUE LES COÏNCIDENCES SONT FRÉQUENTES

En novembre 2008, je participais à un dîner pour l'association « Conservation International ». L'acteur Harrison Ford faisait partie de leur conseil d'administration et j'avais remarqué qu'il portait un clou d'oreille. Le lendemain, j'ai regardé un épisode de la série télévisée *Private Practice* et un personnage faisait remarquer qu'un autre personnage avait un clou d'oreille qui ressemblait à celui d'Harrison Ford. Le jour suivant, je lisais (dans un livre sur les heureux hasards!) que le chercheur Baruch Blumberg, prix Nobel, ressemblait à Indiana Jones, un personnage de film interprété par Harrison Ford (Meyers, 2007).

Quelle est la chance que je tombe sur Harrison Ford trois fois en trois jours? Très faible. Mais cela ne veut pas dire grand-chose. Bien qu'il soit hautement improbable qu'une certaine coïncidence se produise, il est presque certain qu'un ensemble apparemment étonnant d'événements non spécifiés se produise vu que nous observons tellement de choses quotidiennement. Les coïncidences étonnantes sont toujours remarquées a posteriori et jamais annoncées à l'avance.

NOUS NE NOUS ATTENDONS PAS À CE QUE LA VARIABILITÉ DÉPENDE DE LA TAILLE DE L'ÉCHANTILLON

Gelman (1998) s'est penché sur la relation entre les populations de comtés et l'incidence ajustée à l'âge du cancer des reins (un cancer assez rare, qui touche environ 15 adultes sur 100 000 aux États-Unis). Il s'est d'abord intéressé aux comtés à faible prévalence. La plupart de ces comtés étaient peu peuplés. Pourquoi? On peut penser que la qualité de l'environnement dans ces régions rurales a pour effet de faire baisser le taux de cancer du rein. Ensuite, il s'est concentré sur les comtés à prévalence élevée. Là aussi, il s'agit de comtés peu peuplés. Pourquoi? On peut penser que le manque de ressources médicales de ces petits comtés soit la cause de ce taux élevé de cancer du rein. Il est néanmoins assez

étrange de constater que les prévalences les plus élevées et les moins élevées concernent les comtés à faible population.

En y réfléchissant, la raison est simple. La variation autour du taux moyen est faible dans les grands comtés alors que cette variabilité est beaucoup plus forte parmi les petits comtés. Prenons l'exemple extrême d'un petit comté de seulement 1 000 habitants. Si aucun de ces habitants n'a de cancer du rein, ce comté fera partie de ceux dont la prévalence du cancer du rein est la plus basse (zéro). Mais, dès qu'un de ces 1 000 habitants a un cancer du rein, le comté passera parmi ceux qui ont le taux le plus élevé. Dans un comté vraiment petit, il suffit d'un cas pour basculer d'un des taux les plus bas vers un des taux les plus élevés. De façon générale, la chance seule va faire que les taux d'incidence vont varier beaucoup plus dans les comtés à faible population que dans les comtés à population plus nombreuse. Par conséquent, les comtés à prévalence extrême sont moins peuplés que ceux à prévalence plus proche de la moyenne.

Une variation aléatoire peut avoir un plus grand effet sur les moyennes dans des petits groupes que dans des groupes plus nombreux. Bien que logique, ce principe simple n'est pas intuitif chez beaucoup.

NOS INTUITIONS EN PROBABILITÉS SONT FAUSSES

Imaginez que vous pouvez choisir entre deux bols de bonbons (des gommes). Le petit bol contient 9 gommes blanches et une rouge. Le grand bol contient 93 gommes blanches et 7 rouges. Les gommes des deux bols sont bien mélangées et il n'est pas possible de les voir. Vous devez en prendre une et vous gagnez si elle est rouge. Que choisissez-vous : tirer du petit ou du grand bol ?

Si vous choisissez une gomme du petit bol, vous avez 10 % de chance de tirer une gomme rouge. Si vous en prenez une dans le grand bol, la chance d'en avoir une rouge n'est que de 7 %. Clairement, vos chances de gagner sont plus grandes si vous choisissez d'extraire une gomme du petit bol. Cependant, environ 2/3 des gens choisissent le grand bol (Denes-Raj & Epstein, 1994). La plupart de ces personnes font le calcul et savent que la chance de gagner est plus élevée avec le petit bol, mais penchent pour le grand parce qu'il contient plus de gommes rouges et offre donc plus de chances de gagner. Bien sûr, il contient également plus de gommes blanches et donc aussi plus de chances de perdre. Nos cerveaux ne sont tout simplement pas programmés pour une manipulation judicieuse des probabilités et la plupart des gens font un choix illogique.

IL NOUS EST DIFFICILE DE CONCATÉNER DES PROBABILITÉS

Il s'agit cette fois d'un jeu de réflexion classique appelé le problème de Monty Hall librement inspiré du jeu télévisé américain *Let's Make a Deal*. Vous êtes un des concurrents du jeu et vous vous trouvez devant trois portes. Derrière l'une d'elles, il y a une voiture de luxe toute neuve. Vous devez choisir une porte et vous pourrez emporter ce qu'il y a derrière. Vous choisissez une porte. À ce stade, l'animateur choisit une des deux autres portes, il l'ouvre et montre qu'il n'y a pas de voiture derrière elle. Il vous offre la possibilité de changer d'avis et de choisir l'autre porte (celle qui reste).

Allez-vous modifier votre choix ?

Avant de poursuivre votre lecture, vous devriez réfléchir au problème et décider si, à la place du concurrent, vous auriez modifié votre choix. Il n'y a pas de trucs ou attrapes. Une seule porte cache un prix ; toutes les portes sont identiques; l'animateur (qui sait derrière quelle porte se trouve la voiture neuve) a le visage d'un parfait joueur de poker et ne vous donne aucun d'indice. La voiture ne se trouve jamais derrière la porte que l'animateur choisit d'ouvrir. Ne trichez pas. Réfléchissez avant de continuer.

Lors de votre premier choix, il y a trois portes et chacune a la même chance de cacher la voiture. Donc, vous avez une chance sur trois de sélectionner la porte gagnante. Réfléchissons séparément aux deux situations : avoir choisi au départ une porte gagnante ou une porte perdante.

Si vous avez choisi la porte gagnante, c'est que la voiture ne se trouve derrière aucune des deux autres. L'animateur en ouvre une des deux. Si vous modifiez votre choix, vous allez choisir la porte perdante restante.

Et si vous avez choisi une porte perdante? La voiture est derrière une des deux portes restantes. L'animateur ouvre celle derrière laquelle il n'y a pas de voiture. Par conséquent, la dernière porte fermée est forcément la porte gagnante et le fait de modifier votre choix vous amènera alors certainement à gagner.

Récapitulons. Si vous choisissez au départ la porte gagnante (une chance sur trois), alors changer de décision vous fera perdre. Si, vous choisissez au départ une des deux portes perdantes (deux chances sur trois), alors changer de décision vous fera gagner à coup sûr. Passer d'une porte perdante à l'autre porte perdante est impossible puisque l'animateur aura ouvert l'autre porte perdante.

Votre meilleur choix est donc de changer de décision! Bien entendu, vous ne pouvez pas être absolument certain que modifier votre choix va être fructueux. Une fois sur trois, changer d'avis vous fera rater le prix. Mais deux fois sur trois, changer d'avis vous fera gagner le prix. Si vous répétez le jeu de nombreuses fois, vous gagnerez deux fois plus souvent en changeant chaque fois de porte. Si vous ne jouez qu'une fois, alors votre chance est double en changeant de porte.

Presque tout le monde (y compris mathématiciens et statisticiens) arrive intuitivement à la mauvaise conclusion: changer d'avis n'est pas bon (Vos Savant, 1997).

NOUS NE FAISONS PAS DE CALCULS BAYÉSIENS INTUITIVEMENT

Imaginez ce scénario : vous testez des échantillons sanguins pour détecter la présence du virus d'immunodéficience humaine (VIH). La prévalence du VIH est très basse (0,1 %) parmi les donneurs de sang. Le test de détection des anticorps est de très bonne qualité, mais pas parfait. Il identifie correctement 99 % des échantillons sanguins infectés, mais conclut aussi de façon erronée que 1 % des échantillons sanguins non infectés sont porteurs du VIH. Lorsque ce test identifie un échantillon sanguin infecté au VIH, quelle est la probabilité qu'en réalité le donneur ait le VIH et quelle est la probabilité que le résultat du test soit une erreur (faux positif) ?

Essayez de trouver la réponse avant de poursuivre votre lecture.

Imaginons que 100 000 personnes soient testées. Parmi celles-ci, 100 (0,1 %) seront porteurs du VIH et le test sera positif chez 99 (99 %) d'entre elles. Les 99 900 personnes restantes ne sont pas atteintes par le VIH, mais parmi celles-ci, le test fournira de façon

incorrecte, 1 % de résultats positifs. Au total, il y aura donc 99 + 999 = 1 098 tests positifs et seulement 99/1 098 = 9 % seront de vrais positifs. Les 91 % restants de ces tests positifs seront de faux positifs. Donc, si un test est positif, il y a seulement 9 % de chance que le VIH soit présent dans cet échantillon.

La plupart des gens, y compris les médecins, pensent intuitivement qu'un résultat positif au test implique presque sûrement que le VIH est présent. Nos cerveaux ne sont pas programmés pour combiner nos connaissances acquises (la prévalence de VIH) avec de nouvelles connaissances (le test est positif).

Imaginons maintenant que le test soit utilisé dans une population de consommateurs de drogues intraveineuses dans laquelle il est supposé que la prévalence du VIH est de 10 %. Imaginons qu'à nouveau 100 000 personnes soient testées. Dans cette population, 10 000 (10 %) personnes sont atteintes par le VIH et le test sera positif chez 9 900 (99 %) d'entre elles. Les 90 000 personnes restantes ne sont pas porteuses du VIH, mais le test va donner, de façon incorrecte, un résultat positif dans 1 % des cas. Il y aura donc 900 faux positifs. Au total, 9 900 + 900 = 10 800 tests seront positifs et 9 900/10 800 = 92 % seront de vrais positifs. Les 8 % de tests positifs restants seront de faux positifs. Si un test est positif, il y a donc 92 % de chance que cet échantillon soit infecté par le VIH.

L'interprétation du résultat du test dépend fortement de la prévalence de la maladie. Pour arriver à la conclusion correcte, il faut combiner une donnée de base comme la fréquence, avec de nouvelles données. Cet exemple vous donne un avant-goût de ce qui est appelé la logique bayésienne (qui sera étudiée à nouveau aux chapitres 2 et 18).

NE SOYONS PAS DUPÉS PAR LES COMPARAISONS MULTIPLES

Austin et ses collègues (2006) ont exploité une base de données constituée de statistiques de santé relatives à 10 millions de résidents de l'Ontario au Canada. Ils ont examiné 233 motifs d'admission à l'hôpital et enregistré le signe astrologique de chaque patient (basé sur la date de naissance). Y a-t-il un rapport entre le signe astrologique d'une personne et la raison de son admission à l'hôpital, telle est la question qu'ils se sont posée.

Les résultats semblent impressionnants. Soixante-douze maladies (raisons d'admission à l'hôpital) apparaissaient significativement plus fréquemment chez les personnes nées sous un signe astrologique particulier que parmi les personnes nées sous les autres signes astrologiques réunis, cette différence étant statistiquement significative. On dit qu'un résultat est *statistiquement significatif* lorsqu'il se produit par hasard dans moins de 5 % des cas (vous en apprendrez plus sur cette notion au chapitre 16).

Impressionnant, n'est-ce pas? Cela fait croire qu'il y a une véritable relation entre l'astrologie et la santé. Mais il y a un problème. Il est trompeur de se focaliser sur les associations fortes entre une maladie et un signe astrologique sans considérer toutes les autres combinaisons. Austin *et al.* (2006) ont examiné l'association entre 223 raisons différentes d'admission à l'hôpital et 12 signes astrologiques. Ils ont donc testé 2 676 hypothèses distinctes (223 x 12 = 2 676). Comme 5 % de 2 676 = 134, on pourrait s'attendre à trouver environ 134 associations significatives juste par hasard. Ce n'est donc guère impressionnant d'en avoir trouvé seulement 72.

Il est à noter que cette étude ne visait pas vraiment l'association entre le signe astrologique et la maladie. Elle servait à démontrer combien il est difficile d'interpréter des résultats statistiques quand plusieurs comparaisons sont réalisées.

Les chapitres 22 et 23 explorent en profondeur le problème des comparaisons multiples.

NOUS AVONS TENDANCE À IGNORER LES EXPLICATIONS ALTERNATIVES

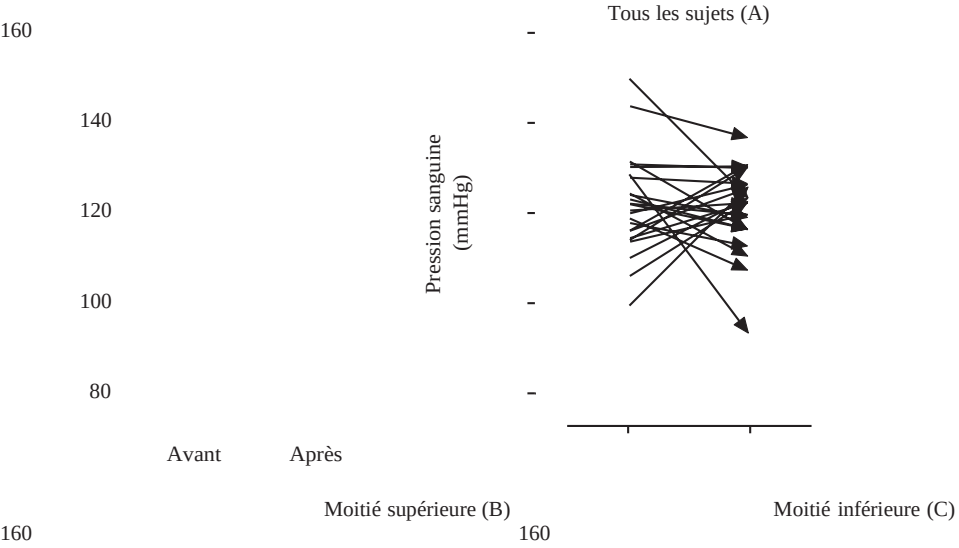
Vous réalisez une étude sur le traitement de l'arthrose par l'acupuncture. Les patients arrivent avec de violentes douleurs aux articulations et vous les traitez par acupuncture. Ils sont priés d'évaluer leur douleur articulaire avant et après le traitement. Comme la douleur diminue chez la plupart des patients et qu'il est fort peu probable que de tels résultats aussi concordants se produisent par hasard, l'acupuncture doit avoir été efficace. Pas vrai ?

Pas nécessairement. La diminution de la douleur peut très bien s'expliquer autrement. Voici cinq explications alternatives (adapté de Bausell, 2007) :

- Le fait d'avoir confiance dans le thérapeute et le traitement peut réduire considérablement la douleur. Le soulagement de la douleur peut être un effet placebo et n'avoir rien à voir avec l'acupuncture.
- Comme les patients veulent se montrer polis, ils peuvent dire au chercheur ce qu'il ou elle veut entendre (que la douleur a diminué). La diminution de douleur est donc due au fait que les patients ne réfèrent pas avec précision l'intensité de la douleur après le traitement.
- Avant, pendant et après le traitement par acupuncture, le thérapeute discute avec les patients. Il conseille peut-être des changements de dosage d'aspirine, la pratique d'exercices ou encore des suppléments nutritionnels. Là, et non dans l'acupuncture, se trouve peut-être la raison de la réduction de douleur.
- L'expérimentateur peut avoir modifié les données. Par exemple, que se passe-t-il si la douleur s'accentue avec l'acupuncture pour trois patients alors qu'elle diminue pour les autres? Après avoir examiné de près les dossiers de ces trois patients, l'expérimentateur décide de les sortir de l'étude parce que l'un des trois a un type d'arthrose différent et deux des trois avaient dû grimper les escaliers pour arriver au rendez-vous en raison d'une panne de l'ascenseur. Les données sont dès lors biaisées et faussées.
- Il est connu que la douleur liée à l'arthrose varie significativement d'un jour à l'autre. Les gens ont donc tendance à consulter lorsque la douleur est la plus aiguë. Si l'on commence à recueillir les données sur la douleur le jour où elle est la plus aiguë, il est fort vraisemblable qu'une amélioration sera observée même sans traitement. Le paragraphe suivant passe en revue ce problème de *régression vers la moyenne*.

NOUS SOMMES DUPÉS PAR LA RÉGRESSION VERS LA MOYENNE

La figure 1.1 présente des pressions obtenues par simulation avant et après un traitement. La figure 1.1A présente 24 paires de valeurs. Les groupes « avant » et « après » sont à peu près les mêmes. Dans certains cas, les valeurs augmentent après traitement, et dans d'autres, elles diminuent. Si ces données étaient réelles, on conclurait qu'il n'y a pas d'évidence du tout que le traitement a eu un effet sur la variable étudiée (pression artérielle).



Moitié supérieure (B)

Moitié inférieure (C)

160

160

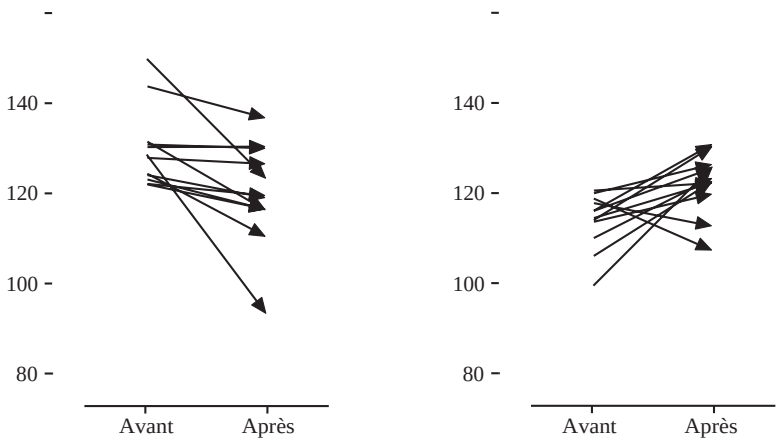


Figure 1.1. Régression vers la moyenne

Toutes les données de la figure (A) proviennent d'une distribution aléatoire gaussienne (moyenne égale à 120, écart-type égal à 15) sans attribuer d'importance à « avant » et « après » ni à un quelconque appariement. (A) montre 48 valeurs aléatoires, divisées arbitrairement en 24 paires avant-après (avec suffisamment de chevauchement pour qu'il soit impossible de les compter). (B) ne montre que les 12 paires des plus hautes valeurs précédentes. Sauf dans un cas, les valeurs « après » sont plus basses que les valeurs « avant ». (C) montre les paires des plus basses valeurs précédentes. Dans 10 des 12 paires, les valeurs après sont plus élevées que les valeurs avant. En présence des graphiques (B) et (C) seulement, vous seriez tenté de conclure que n'importe quel traitement intervenu entre l'avant et l'après a une grange influence sur la pression sanguine. Ces graphiques montrent en fait des valeurs aléatoires, sans changement systématique entre avant et après. L'effet présumé s'appelle *régression vers la moyenne*.

Imaginez maintenant que le protocole de l'étude soit différent. Sur la base des mesures d'avant traitement, on souhaite tester un traitement seulement pour les pressions artérielles élevées. Les personnes dont la pression artérielle n'est pas élevée sont écartées. La figure 1.1B présente uniquement les données des 12 personnes dont les pressions de départ étaient les plus élevées. Dans tous les cas sauf un, les valeurs après traitement sont plus basses. Un test statistique (test t pour échantillons appariés, voir le chapitre 31) montrerait clairement que le traitement fait baisser la pression artérielle. La figure 1.1C illustre la situation des 12 autres paires, celles dont les pressions de départ étaient basses. Pour toutes les paires sauf deux, les valeurs après traitement sont plus élevées. Ce qui met en évidence de façon convaincante que le traitement fait augmenter la pression artérielle.

Or, ces données sont aléatoires ! Les valeurs avant et après proviennent de la même distribution. Que s'est-il passé ?

Ce qui précède est un exemple de *régression vers la moyenne :* plus la valeur d'une variable est extrême lors de sa première mesure, plus il est probable qu'elle soit plus proche de la moyenne lors de sa deuxième mesure. Les personnes ayant été particulièrement chanceuses en bourse une année donnée auront vraisemblablement moins de réussite l'année suivante. Les personnes qui ont des notes extrêmement élevées à un examen auront vraisemblablement des notes plus basses une autre fois. Un athlète qui fait d'excellents résultats durant une saison en aura vraisemblablement de moins bons la saison suivante. Ceci explique probablement la réputation de porter malheur de la couverture de *Sports Illustrated* – beaucoup croient que le fait d'apparaître sur la couverture de *Sports Illustrated* portera malchance à un athlète (Wolff, 2002).

NOUS LAISSONS NOS BIAIS DÉTERMINER NOTRE MANIÈRE D'INTERPRÉTER DES DONNÉES

Kahan et ses collègues (2013) ont demandé à une multitude de gens leur avis sur de simples données semblables à celles du chapitre 27. Alors que les tableaux contiennent les mêmes nombres, ils servent tantôt à mesurer l'efficacité d'une crème pour traiter une éruption cutanée (tableau 1.2), tantôt à mesurer l'efficacité d'une loi qui interdit le port des armes de poing dissimulées en public (tableau 1.3).

	L'ÉRUPTION CUTANÉE S'EST AMÉLIORÉE	L'ÉRUPTION CUTANÉE A EMPIRÉ	TOTAL
Patients qui **ont utilisé** l'onguent	223	75	298
Patients qui **n'ont pas** utilisé l'onguent	107	21	128

Tableau 1.2. Un des tableaux utilisé par Kahan et ses collègues. Il a été demandé aux personnes interrogées de dire si, à la vue de ce tableau, l'expérience hypothétique montrait que l'état de la peau des patients traités s'était amélioré ou le contraire, par rapport aux patients non traités.

	DIMINUTION DE LA CRIMINALITÉ	AUGMENTATION DE LA CRIMINALITÉ	TOTAL
Villes qui **ont interdit** de porter des armes de poing en publique	223	75	298
Villes qui **n'ont pas interdit** de porter des armes de poing en public	107	21	128

Tableau 1.3. L'autre tableau utilisé par Kahan et ses collègues. Il a été demandé aux personnes interrogées de dire si, à la vue de ce tableau, ces données montraient que le fait d'interdire ou non de porter des armes de poing en public avait eu un effet positif ou négatif sur le taux de criminalité dans les villes.

Il n'était pas demandé de deviner une subtile interprétation des données, mais seulement si ces données pouvaient soutenir une certaine hypothèse. Mathématiquement, les tableaux 1.2 et 1.3 conduisent assez directement à :

- Selon la première ligne, l'éruption cutanée s'améliore dans 223/298 = 74,8 % des cas traités par l'onguent et la criminalité baisse de 74,8 % dans les villes qui ont interdit le port d'armes dissimulées.
- Selon la deuxième ligne, l'éruption cutanée s'améliore dans 107/128 = 83,6 % des cas non traités par l'onguent et la criminalité baisse de 83,6 % dans les villes qui n'ont pas interdit le port d'armes dissimulées.
- La plupart des gens bénéficient d'une amélioration cutanée et la plupart des villes ont une criminalité en baisse. Mais alors, l'action est-elle pertinente ? Comme 74,8 % est inférieur à 83,6 %, les personnes qui ont utilisé l'onguent étaient moins susceptibles d'obtenir une amélioration de leur peau que celles qui ne l'avaient pas employé. Les villes qui ont instauré la loi sur le port d'armes ont eu une baisse de criminalité moindre que celles qui ne l'ont pas instaurée.

Quand les données se rapportaient à l'efficacité d'une crème pour la peau, les Démocrates et les Conservateurs (les deux principaux partis aux États-Unis) tenaient le même discours. Par contre, quand il s'agissait de l'efficacité d'une mesure de sécurité, les résultats étaient teintés de politique. Les démocrates sont d'avis que les données démontrent une plus grande sécurité. Les conservateurs soutiennent le contraire (c'était au temps où les conservateurs étaient contre une législation sur la sécurité des armes à feu).

Cette étude montre que quand les gens ont une idée préconçue à propos de la conclusion, ils ont tendance à interpréter les données en faveur de cette conclusion.

NOUS AVONS BESOIN DE CERTITUDES, ALORS QUE LES STATISTIQUES OFFRENT DES PROBABILITÉS

Beaucoup de gens attendent que les statistiques conduisent à des conclusions définitives, alors qu'en fait, les conclusions statistiques s'expriment en termes de probabilités. Vous aurez du mal à apprendre les statistiques si vous continuez à vouloir des conclusions définitives. Comme aurait dit le statisticien Myles Hollander: les statistiques signifient ne jamais dire qu'on est certain (cité dans Samaniego, 2008).

RÉSUMÉ

- Nos cerveaux jouent un mauvais rôle dans l'interprétation des données. Nous voyons des structures dans des données aléatoires, nous avons tendance à avoir trop confiance en nos conclusions et à massacrer des interprétations qui impliquent des probabilités combinées.
- Nos intuitions ont tendance à nous égarer quand nous interprétons des probabilités et des comparaisons multiples.
- Une rigueur statistique (et scientifique) est nécessaire pour éviter d'arriver à des conclusions erronées.

CHAPITRE 2

La complexité de la probabilité

> La pensée statistique sera un jour aussi nécessaire pour une citoyenneté efficace que la compétence de lire et d'écrire.

Des livres entiers ont été écrits sur la probabilité. C'est un sujet qui paraît simple de prime abord, mais qui devient plus compliqué à mesure que l'on entre dans les détails. Ce chapitre donne un très large aperçu qui vous aidera à comprendre pourquoi la probabilité est déroutante, à donner du sens aux énoncés concernant une probabilité et à fournir une base solide si vous décidez d'en apprendre plus sur le sujet.

LES BASES DE LA PROBABILITÉ

Les probabilités vont de 0 à 1 (ou 100 %) et servent à mesurer une prédiction au sujet d'un événement futur ou la certitude d'une croyance. Une probabilité nulle signifie que soit l'événement ne se produit pas, soit quelqu'un est absolument sûr que l'énoncé est faux. Une probabilité de 1 (ou 100 %) signifie qu'un événement va arriver avec certitude ou que quelqu'un est certain que l'énoncé est correct. Une probabilité de 0,5 (ou 50 %) signifie qu'il y a autant de chances que l'événement se produise que l'inverse ou qu'il y a autant de personnes qui croient que l'énoncé est correct que de personnes qui croient l'inverse.

Afin de faire la distinction entre deux usages de la probabilité, ce chapitre utilise la terminologie de Kruschke (2011) :

- La probabilité qui est « là-bas » ou « en dehors de la tête ». C'est la probabilité en tant que fréquence à long terme. La probabilité qu'un certain événement se produise a une valeur définie, mais nous n'avons pas assez d'information pour connaître cette valeur avec certitude.
- La probabilité qui est « à l'intérieur de la tête ». C'est la probabilité en tant que force des croyances subjectives qui peut donc varier selon les personnes et même varier entre les différentes évaluations chez une même personne.

PROBABILITÉ EN TANT QUE FRÉQUENCE À LONG TERME

Les probabilités comme prédictions à partir d'un modèle

Un simple exemple pour illustrer : une femme planifie de tomber enceinte et voudrait connaître la chance d'avoir un garçon. Une façon de penser les probabilités est d'y voir les prédictions d'événements futurs qui découlent d'un modèle. Un *modèle* est la description d'un mécanisme. Voici le modèle créé pour notre exemple :

- Chaque ovule contient un chromosome X et aucun n'a de chromosome Y.
- La moitié des spermatozoïdes possède un chromosome X (et pas de Y) et la moitié possède un chromosome Y (et pas de X).
- Un seul spermatozoïde peut féconder un ovule.
- Chaque spermatozoïde a une chance égale de fertiliser un ovule.
- Si le spermatozoïde gagnant a un chromosome Y, le fœtus aura un chromosome X et un chromosome Y, ce sera un garçon. Si le spermatozoïde gagnant a un chromosome X, le fœtus aura deux chromosomes X, ce sera une fille.
- La fausse couche ou l'avortement est susceptible d'arriver de la même façon à un fœtus mâle qu'à un fœtus femelle.

En supposant que ce modèle soit valable, il est facile d'en tirer les prédictions. Puisque tous les spermatozoïdes ont la même chance de fertiliser l'ovule et qu'il y a autant de spermatozoïdes à chromosome X qu'à chromosome Y, la chance que le fœtus ait un chromosome Y est de 50 %. Selon ce modèle, la chance que le fœtus soit mâle est de 50 %. Dans tout groupe de bébés, la fraction des garçons doit être à peu près égale à 1/2. Et à long terme, on peut s'attendre à ce qu'exactement 50 % des bébés soient des garçons.

Les prédictions sur la survenue d'événements futurs peuvent reposer sur n'importe quel modèle, même si celui-ci ne reflète pas la réalité. Certains modèles s'avèrent utiles, d'autres non. Comme le modèle décrit ici est assez correct, ses prédictions sont assez utiles, mais pas parfaites.

Les probabilités basées sur des données

Parmi les bébés nés dans le monde entier en 2011, 51,7 % étaient des garçons (Central Intelligence Agency [CIA], 2012). Il n'est pas nécessaire de savoir *pourquoi* il y a eu plus de garçons que de filles, pas plus que de savoir pourquoi la CIA recueille cette information (bien que je sois curieux). Il y a simplement le fait que cette fraction entre les sexes a été observée de façon assez cohérente pendant plusieurs années dans de nombreux pays. Sur la base de ces données, on peut répondre à la question: quelle est la chance que mon bébé soit un garçon? La réponse est 51,7 %. Dans un groupe de 1 000 femmes enceintes, il faut s'attendre à ce qu'il y ait 517 fœtus mâles et 483 fœtus femelles. Dans un groupe particulier de 1 000 femmes enceintes, le nombre de mâles sera supérieur ou inférieur à 517, mais 517 est le nombre attendu à long terme.

PROBABILITÉ EN TANT QUE FORCE DE CROYANCES

Les probabilités subjectives

Vous souhaitez un garçon. En cherchant sur Internet, vous découvrez un livre intéressant :

> *Comment choisir le sexe de son enfant* explique la méthode simple de Shettles et expose pas à pas la marche à suivre pour obtenir un enfant d'un certain genre. Cette méthode, bien appliquée, donne aux couples au moins 75 % de chance d'avoir l'enfant du sexe désiré (Shettles, 1996).

Les critiques de ce livre sont élogieuses et vous convainquent que son principe est correct. En conséquence, vous planifiez de suivre les recommandations de manière à accroître vos chances d'avoir un garçon.

Quelle est la chance que vous ayez un garçon? Si vous vous fiez complètement à cette méthode, vous croyez que, comme il est écrit sur la couverture, la probabilité d'avoir un garçon est de 75 %. Bien que vous n'ayez aucune donnée qui vient étayer ce pronostic, vous y croyez fermement. C'est la *probabilité subjective* à laquelle vous êtes fermement attaché.

Et si vous aviez des doutes? Vous pensez que la méthode fonctionne probablement et vous mesurez cela en disant qu'elle fonctionne dans 85 % des cas (c'est-à-dire que votre chance d'avoir un garçon serait de 75 %). Ce qui laisse 15 % de chance que la méthode ne vaut rien et votre chance d'avoir un garçon se réduit à 51,7 %. Quelle est votre chance d'avoir un garçon? Calculons une moyenne pondérée des deux prédictions, basée sur votre évaluation subjective de la chance que chaque théorie est correcte. Le calcul est (0,850 x 0,750) + (0,150 x 0,517) = 0,715, soit 71,5 %.

Il est clair que d'autres personnes ont des croyances différentes au sujet de l'efficacité de la méthode. Je n'ai pas approfondi le sujet, mais j'ai passé un peu de temps à chercher sans trouver d'étude qui prouve l'efficacité de la méthode de Shettles, et j'en ai même trouvé une (dans une revue prestigieuse) qui disait qu'elle ne marchait pas (Wilcox *et al*., 1995). Étant donné cette recherche, je modifie mes croyances en disant qu'il y a seulement 1 % de chance que la méthode de Shettles est efficace (que la chance d'avoir un garçon soit de 75 %). Ce qui laisse 99 % de chance que la méthode ne marche pas (et ainsi la chance d'avoir un garçon est de 51,7 %). La moyenne pondérée est alors (0,010 x 0,750) + (0,990 x 0,517) = 0,519, soit 51,9 %.

Vu que vous et moi n'avons pas le même avis sur la probabilité d'efficacité de la méthode de Shettles, ma réponse sur la chance d'avoir un garçon (51,9 %) ne colle pas avec la vôtre (71,5 %). Cela n'a rien de surprenant.

Les probabilités pour mesurer l'ignorance

Supposons que vous (ou toute autre personne de vos connaissances) êtes enceinte, mais n'avez encore fait aucun examen (par exemple, échographie ou caryotype) pour savoir quel est le sexe de l'enfant ou n'avez subi aucune intervention (voir section précédente) qui prétende modifier les chances d'avoir une fille ou un garçon. Quelle est la probabilité d'avoir un garçon ?

Dans un sens, le concept de probabilité n'est simplement pas concerné. L'événement aléatoire était la course de plusieurs spermatozoïdes pour fertiliser l'ovule. C'est déjà arrivé. Un spermatozoïde a gagné et le fœtus est mâle ou femelle. Comme l'événement aléatoire s'est déjà produit et que le genre du fœtus est maintenant un fait, on pourrait affirmer que cela n'a pas de sens de parler de probabilité, ou hasard, ou chance. La chance n'est plus impliquée. Le problème est l'ignorance, pas le hasard.

Autre perspective : avant de tomber enceinte, il y avait 51,7 % de chance que le fœtus soit mâle. Sur plusieurs centaines de femmes enceintes ignorantes du sexe de leur enfant, il faut s'attendre à ce que 51,7 % portent un fœtus mâle. Il semble dès lors sensé de dire qu'il y a environ 51,7 % de chance que le fœtus soit mâle et 48,3 % qu'il soit femelle. Comme c'était votre chance avant la conception, ce l'est encore maintenant.

Notez le changement de point de vue. Il n'est plus question maintenant de prédire l'issue d'un événement futur aléatoire mais plutôt de mesurer l'ignorance d'un événement qui s'est déjà produit.

Prédictions quantitatives d'événements ponctuels

Au moment où je révisais ce chapitre, en octobre 2016, on était à quatre semaines des élections présidentielles américaines. Les gens exprimaient différentes opinions et interprétaient les sondages diversement. Le *New York Times* du 5 octobre affirmait qu'il y avait 81 % de chance que Clinton soit élue et seulement 19 % que Trump le soit. Le même jour, le site fivethirtyeight.com disait qu'il y avait 76 % de chance que Clinton soit élue et 24 % que Trump le soit.

Ces probabilités sont-elles subjectives ? En partie. Ces probabilités étaient basées sur des sondages. Mais il y avait plusieurs sondages et chaque publication avait fait des choix subjectifs en retenant certains sondages plutôt que d'autres et des pondérations faites au moment d'en faire la moyenne. Voilà qui explique pourquoi les estimations d'institutions honorables diffèrent.

Peut-on interpréter les 81 % comme une fréquence à long terme? Non. Cela ne signifie pas que le *New York Times* pense que Clinton va gagner 81 % de toutes les élections présidentielles auxquelles elle va participer, ni que 81 % des électeurs vont voter pour Clinton. Sur la base des sondages, le *Times* était certain à 81 % que Clinton allait l'emporter. Telle était son opinion le 5 octobre. Le 22 octobre, à deux semaines des élections, le *New York Times* avait relevé son estimation à 93 % et le site fivethirtyeight.com à 86 %. Les probabilités publiées étaient des affirmations chiffrées sur des croyances étayées. Les sondages plus récents leur avaient fait adapter leurs croyances.

En fin de compte, c'est Trump qui a gagné. Les sondeurs avaient-ils tort? Pas vraiment. Ils n'avaient pas dit que Trump n'avait aucune chance de gagner, seulement 24 %, 29 %, 7 % ou 14 % de chance (selon la source consultée et quand).

LES CALCULS DE PROBABILITÉS SONT PLUS FACILES SI VOUS PASSEZ AUX NOMBRES ENTIERS

Gigerenzer (2002) a montré que reformuler les problèmes de probabilité pouvait grandement influencer la chance qu'ils soient résolus correctement. Son exemple concerne l'interprétation des résultats de mammographies effectuées sur des femmes asymptomatiques âgées entre 40 et 50 ans. Il avait demandé à 48 médecins d'interpréter une mammographie positive (anormale). La moitié des médecins ont reçu la question libellée en termes de probabilités :

> La probabilité qu'une de ces femmes ait un cancer du sein est 0,8 %. Si une femme a un cancer du sein, la probabilité qu'elle ait une mammographie positive est de 90 %. Si une femme n'a pas de cancer du sein, la probabilité qu'elle ait une mammographie positive est de 7 %. Supposons qu'une femme présente une mammographie positive. Quelle est la probabilité qu'elle ait réellement un cancer du sein ?

Allez-y. Essayer de trouver la réponse.

L'autre moitié des médecins ont reçu la même question formulée en fréquences plutôt qu'en pourcentages :

Huit femmes sur 1 000 auront le cancer du sein. Parmi ces 8 femmes, 7 vont avoir une mammographie positive. Parmi les 992 restantes qui n'ont pas le cancer du sein, environ 70 vont quand même présenter une mammographie positive. Supposons qu'une femme présente une mammographie positive. Quelle est la probabilité qu'elle ait réellement un cancer du sein ?

Essayez de répondre à la question. Est-ce plus facile ?

Les médecins des deux groupes ont eu le temps de réfléchir au problème. Mais ceux du premier groupe ont mal réussi. Seulement 2 des 24 ont répondu correctement (la probabilité d'avoir le cancer du sein est de 9 %). Quelques autres étaient proches de la réponse. La majorité était tout à fait à côté, la réponse la plus fréquente étant de 90 %. Les médecins du second groupe ont fait nettement mieux avec une majorité de bonnes réponses.

Pourquoi la réponse est-elle 9 % ? Sur les 1 000 femmes soumises au test, il y aura 7 mammographies positives parmi celles qui ont réellement un cancer du sein et 70 parmi celles qui ne l'ont pas. En d'autres mots, compte tenu de la configuration de ce problème, on s'attend à ce que 7 des 77 mammographies positives attestent réellement d'un cancer du sein (environ 9 %).

Si vous êtes bloqués devant des problèmes de probabilités, essayez de changer la formulation en passant à des nombres entiers. Cette astuce rend les choses beaucoup plus faciles.

ERREURS FRÉQUENTES : PROBABILITÉ

Erreur : négliger les hypothèses

Le premier exemple pose la question: quelle chance y a-t-il qu'un fœtus soit mâle? Cette question n'a pourtant pas de sens sans contexte. Pour qu'elle en ait, il faut accepter une série d'hypothèses parmi lesquelles :

- On parle de fœtus humains. La proportion entre les sexes peut être différente pour d'autres espèces.
- Il n'y a qu'un seul fœtus. « Fille ou garçon ? » est une demande ambiguë si on admet des jumeaux ou des triplés.
- Il y a une très faible probabilité (inconnue) que le fœtus ne soit pas complètement mâle ou femelle.
- Le rapport des sexes est le même dans tous les pays et toutes les races.
- Le rapport des sexes ne varie pas d'une année à l'autre, ni d'une saison à l'autre.
- Il n'y a pas de discrimination sur le sexe (il y en a eu) en cas d'avortement ou de fausses couches. Ainsi, le rapport entre les sexes est le même au moment de la conception et de la naissance.

Quand vous entendez des questions ou des énoncés relatifs à une probabilité, rappelez-vous que les probabilités sont *toujours* liées à un ensemble d'hypothèses. Une réflexion éclairée sur une probabilité dans une situation, quelle qu'elle soit, exige de connaître les hypothèses.

Parfois, la question sur les probabilités est formulée comme suit : Si A est vrai, quelle est la chance que B survienne? Cela s'appelle une *probabilité conditionnelle* parce qu'il

s'agit de la probabilité qu'un certain événement B survienne conditionnellement à ce qu'un autre événement A survienne aussi.

Erreur : essayer de comprendre une probabilité sans avoir clairement défini les numérateur et dénominateur

L'exemple traité dans ce chapitre parle de la fraction de bébés mâles. Le numérateur est le nombre de bébés mâles. Le dénominateur est le nombre de bébés nés à un certain endroit pendant une certaine période. Dans ce cas, la fraction est simple et sans ambiguïté.

Passons à un autre exemple à propos de bébés où il y a ambiguïté. En cas d'infertilité, les couples se tournent vers la fécondation in vitro pour accroître leur chance d'avoir un enfant. Après que l'ovule ait été fécondé dans une éprouvette, un ou plusieurs embryons sont implantés dans l'utérus. La chance d'une grossesse est plus grande si on implante plus d'un embryon, mais alors il y a le risque d'avoir des jumeaux ou des triplés (ou plus).

Les couples qui recourent à la fécondation in vitro veulent connaître le taux de succès. Le dénominateur est clairement le nombre de femmes qui ont fait appel à cette technique de procréation assistée. Mais quel est le numérateur? Pour que la réponse soit une fréquence qui puisse être interprétée comme une probabilité, il faut que le numérateur et le dénominateur comptent la même chose. Le numérateur doit donc être le nombre de femmes qui tombent enceintes. Chaque femme, soumise à cette technique, tombe enceinte ou non. Deux issues sont possibles. On peut donc compter le nombre de femmes qui tombent enceintes à la suite du traitement et se servir de ces résultats comme une fréquence qui prédit les résultats pour d'autres femmes dans le futur.

Il y a des décennies, quand mon épouse recourut à la fécondation in vitro, la clinique utilisait une autre méthode pour calculer le taux de réussite. Elle divisait le nombre de bébés nés par le nombre de femmes ayant reçu le traitement. Ce rapport n'est pas une probabilité et n'a pas réellement d'interprétation, car les sujets comptés par le numérateur et le dénominateur sont différents : le numérateur compte des bébés et le dénominateur, des femmes. Ce taux de réussite est supérieur à la probabilité dont il est question dans le paragraphe précédent, car certaines femmes ont plus qu'un bébé.

Il est impossible de comprendre ce que le taux de réussite signifie (ou comparer les taux de deux cliniques) sans savoir exactement comment les numérateur et dénominateur ont été définis.

Erreur : inverser les énoncés de probabilité

Dans l'exemple traité ci-dessus, il n'y a pas de danger d'intervertir accidentellement les énoncés de probabilité. La probabilité qu'un bébé soit un garçon est manifestement très différente de la probabilité qu'un garçon soit un bébé. Mais il y a beaucoup de situations où il est facile de prendre les choses à contresens.

- La probabilité qu'un drogué à l'héroïne ait d'abord consommé de la marijuana n'est pas la même chose que la probabilité qu'un toxicomane à la marijuana passe plus tard à l'héroïne.
- La probabilité que quelqu'un qui a mal au ventre souffre d'une appendicite n'est pas la même chose qu'une personne diagnostiquée appendicite ait eu mal au ventre.

- La probabilité qu'un livre de statistiques soit ennuyeux n'est pas la même chose que la probabilité qu'un livre ennuyeux soit un livre de statistiques.
- La part d'études réalisées sous l'hypothèse nulle dans lesquelles la P-valeur est inférieure à 0,05 n'est pas la même que la part des études avec une P-valeur inférieure à 0,05 pour lesquelles l'hypothèse nulle est vraie. (Ceci vous sera plus compréhensible lorsque vous aurez lu le chapitre 15).

Savoir qu'il est facile d'intervertir malencontreusement un énoncé de probabilité vous aidera à ne pas commettre cette erreur.

Erreur : croire que la probabilité a une mémoire

Un couple a 4 enfants, tous des garçons. Quelle est la chance que l'enfant suivant soit un garçon ? Une erreur fréquente est de penser que la probabilité d'avoir une fille est plus élevée comme si, ayant déjà 4 garçons, il était en quelque sorte obligatoire d'avoir une fille. C'est tout simplement faux.

Cette erreur est fréquente lorsqu'il s'agit de jeux d'argent, comme la roulette ou la loterie. Les gens parient sur un nombre qui n'est pas sorti depuis longtemps croyant qu'il a plus de chance de sortir au prochain tirage. Cela s'appelle la *tromperie des jeux d'argent*. La probabilité n'a pas de vocabulaire.

JARGON

Probabilité versus cote

Jusqu'à présent, on a mesuré les chances comme des probabilités. Mais il est possible d'exprimer ces valeurs comme des *cotes*. Les cotes et les probabilités sont deux façons d'exprimer exactement la même chose. Toute probabilité peut être exprimée par une cote. Toute cote peut être exprimée par une probabilité. Certains domaines scientifiques ont tendance à privilégier probabilité ; d'autres, cote. Il n'y a aucun avantage en faveur de l'un ou de l'autre.

Si vous cherchez des informations démographiques sur la fraction de bébés mâles, vous allez trouver *sex-ratio*. Ce terme signifie dans la bouche des démographes le rapport des sexes à la naissance. À l'échelle mondiale, ce rapport se situe dans beaucoup de pays autour de 1,07. Autrement dit, la cote d'avoir un garçon contre une fille est de 1,07 à 1 ou 107 à 100.

La cote se convertit facilement en probabilité. Si 107 garçons naissent chaque fois que 100 filles naissent, la chance qu'un nouveau-né soit un garçon est 107/(107 + 100) = 0,517, soit 51,7 %.

De même, une probabilité se convertit facilement en cote. Si la probabilité d'avoir un garçon est 51,7 %, c'est qu'il y a 517 garçons sur 1 000 naissances. Donc 483 filles. La cote d'avoir un garçon contre une fille est 517/483 = 1,07 à 1. La cote est définie comme la probabilité que l'événement se produise divisée par la probabilité qu'il ne se produise pas.

Une cote peut être tout nombre positif ou nul, mais pas négatif. Une probabilité doit être un nombre compris entre 0 et 1 quand elle est exprimée sous la forme d'une fraction ou comprise entre 0 et 100 lorsqu'elle est exprimée en pour cent.

Une probabilité de 0,5 équivaut à une cote de 1. La probabilité d'obtenir face en jetant une pièce de monnaie est 50 %. La cote est 50 : 50, ou 1. Quand la probabilité passe de 0,5

à 1, la cote augmente de 1 à l'infini. Par exemple, si une probabilité est égale à 0,75, la cote est 75 : 25, soit 3 contre 1 ou 3.

Probabilité versus statistique

Les termes *probabilité* et *statistique* sont souvent associés dans les cours ou les titres de livres, mais ils désignent des choses différentes.

Ce chapitre a traité de probabilité. Même si les détails peuvent devenir compliqués et faire que la théorie des probabilités est appliquée incorrectement, les concepts sont assez simples. On part du cas général, appelé la population ou le modèle et on fait des prédictions sur ce qui arrivera dans beaucoup d'échantillons de données. Les calculs de probabilité vont du général au particulier, de la population à l'échantillon (ainsi que vous le verrez au chapitre 3) et du modèle aux données (comme vous le verrez au chapitre 34).

Les calculs statistiques fonctionnent dans le sens opposé (tableau 2.1). On part avec un ensemble de données (l'échantillon) et on en tire des déductions sur la totalité de la population ou du modèle. La logique est d'aller du particulier vers le général, de l'échantillon à la population et des données au modèle.

Probabilité versus vraisemblance

Dans le langage courant, les termes *probabilité* et *vraisemblance* sont assez synonymes. En revanche, ils n'ont pas le même sens en statistiques. Ce livre n'utilise pas vraisemblance, mais si vous lisez d'autres livres de mathématiques, vous devez savoir que vraisemblance a une signification technique différente de probabilité. Brièvement, probabilité répond aux questions mentionnées dans la moitié supérieure du tableau 2.1 tandis que vraisemblance répond aux questions mentionnées dans la moitié inférieure.

PROBABILITÉ EN STATISTIQUE

Précédemment dans ce chapitre, j'ai fait remarquer que la probabilité pouvait être « là-bas » ou « à l'intérieur de la tête ». Le reste de ce livre est surtout consacré aux intervalles de confiance et aux P-valeurs qui utilisent les probabilités « là-bas ». Ce type d'analyse des

PROBABILITÉ		
Général	▯	Spécifique
Population	▯	Échantillon
Modèle	▯	Données
STATISTIQUE		
Général	▯	Spécifique
Population	▯	Échantillon
Modèle	▯	Données

Tableau 2.1. Distinction entre probabilité et statistique.

La théorie des probabilités va du général au spécifique, de la population à l'échantillon, du modèle vers les données.

données est appelé s*tatistiques fréquentielles*. Des croyances ou des données antérieures n'entrent jamais dans les calculs fréquentiels. Le calcul des P-valeurs (chapitre 15) et des intervalles de confiance (chapitres 4 et 12) n'est effectué que sur la base de données issues d'un ensemble du moment. Cependant, des scientifiques tiennent souvent compte de données et de théorie antérieures lorsqu'ils interprètent les résultats, aspect sur lequel nous reviendrons aux chapitres 18, 19 et 42.

Beaucoup de statisticiens préfèrent une autre approche, dites *statistiques bayésiennes*, dans lesquelles des croyances antérieures sont quantifiées et partiellement utilisées dans les calculs. Ces probabilités antérieures peuvent être subjectives (basées sur une opinion informée), objectives (basées sur des données solides ou une théorie bien établie) ou peu informatives (basées sur la croyance que toutes les possibilités sont également vraisem- blables). Les calculs bayésiens mélangent ces probabilités antérieures avec les données du moment pour calculer des probabilités et des intervalles de confiance bayésiens, appelés *intervalles crédibles*. Ce livre parle peu des statistiques bayésiennes.

Q & R

Toutes les probabilités peuvent-elles être exprimées en termes de fraction ou de pour-centage?

Oui, il suffit de multiplier la fraction par 100 pour obtenir son expression en pour cent.

Toutes les fractions sont-elles des probabilités?

Non, une fraction n'est une probabilité que s'il n'y a que deux résultats possibles. Les fractions peuvent être utilisées dans beaucoup d'autres situations.

Les valeurs des probabilités sont-elles toujours comprises entre 0 et 1, ou entre 0 % et 100 %?

Oui.

RÉSUMÉ

- Les calculs de probabilité peuvent être déroutants.
- Probabilité a deux significations : fréquence sur le long terme qu'un événement va se réaliser. C'est la probabilité « là-bas ». L'autre, c'est que la probabilité mesure le degré de certitude que l'on a sur le caractère vrai d'une proposition. C'est la probabilité « dans la tête ».
- Tous les énoncés de probabilité sont basés sur un ensemble d'hypothèses.
- Il est impossible de comprendre une probabilité ou une fréquence avant d'avoir clairement défini les numérateur et

dénominateur.

- Les calculs de probabilité vont du général au particulier, de la population à l'échantillon et du modèle aux données. Les calculs statistiques fonctionnent dans le sens opposé. On va du particulier vers le général, de l'échantillon à la population, et des données au modèle.
- Les méthodes de statistiques fréquentielles calculent des probabilités à partir de données (P-valeurs, intervalle de confiance).