Avant-propos

Tiens, encore un nouveau livre de statistique ! Mais qu'est-ce que celui-ci pourrait m'apporter de plus que les autres ouvrages de statistique que je possède déjà sur mon étagère ? La réponse brève est que dans cet ouvrage le lecteur va trouver non seulement une présentation des concepts et des méthodes statistiques les plus utilisées dans la recherche clinique et les études épidémiologiques, mais surtout il va apprendre à réaliser des analyses au moyen du logiciel Stata et de nombreux jeux de données. Ce dernier point est très important, car la plupart des ouvrages d'introduction à la statistique se limitent à présenter la théorie et l'illustrer au moyen d'exemples mais n'apprennent pas au lecteur à mettre en pratique ces techniques, ni même ne montrent les limites de leur application. Très souvent, les exercices proposés sont purement théoriques ou nécessitent l'emploi d'une simple calculette, mais bien plus rarement l'auteur montre comment appliquer la théorie et réaliser des exercices complexes, basé sur des vraies données, au moyen d'un logiciel statistique. Cet ouvrage tente, donc, de combler un manque pour l'utilisateur de la statistique qui est le passage de la théorie à la pratique.

Je me souviens de mes premières années d'activité dans la recherche clinique en tant que jeune (bio-) statisticien. Après avoir achevé une thèse en Econométrie et Statistique en 1997 à l'Université de Genève, je rejoignais en 1998 la bio-statistique et la recherche clinique à l'Institut universitaire de médecine sociale et préventive (IUMSP) de Lausanne, ainsi qu'à l'Etude suisse de cohorte HIV (SHCS) du Centre hospitalier universitaire vaudois (CHUV). Il ne suffisait pas d'avoir une bonne connaissance théorique et maitriser l'outil statistique, il fallait surtout appendre à communiquer avec le chercheur clinicien, pour qui la statistique est souvent une science obscure, et à poser les bonnes questions afin de comprendre la question de recherche clinique et la traduire en une question statistique, c'est-à-dire identifier le paramètre d'intérêt et formuler une hypothèse concernant ce paramètre.

Le passage de la théorie à la pratique n'est pas des plus évidents non plus. La plupart des ouvrages de statistiques traitent avant tout de la théorie, mais plus rarement les auteurs montrent comment mettre en œuvre les méthodes statistiques présentées et passer de la théorie à la pratique, ni même comment apprécier si les hypothèses sur lesquelles reposent ces méthodes sont validées dans un contexte donné, afin justifier leur utilisation. C'est dans un souci de partager mon expérience pratique de biostatisticien (ou statisticien de terrain ☺) que j'ai écrit ce livre qui se veut complémentaire aux autres livres qui traite de la statistique d'un point de vue théorique avant tout.

L'objectif principal de cet ouvrage est l'apprentissage de la méthodologie statistique par l'application. A cette fin, une méthode pédagogique originale d'enseignement, par rapport aux ouvrages traditionnels d'introduction à la statistique, a été adoptée dont l'objectif est de faciliter le passage de la théorie à la pratique. Chaque concept théorique nouvellement introduit est d'abord présenté de façon rigoureuse en mettant en évidence les hypothèses sur lesquelles il repose. Ensuite, il est illustré au moyen d'un ou plusieurs exemples concrets, issus pour la plupart du domaine biomédical. Enfin, des exercices avec « fil rouge » à réaliser au moyen du logiciel Stata et impliquant le plus souvent un petit jeu de données sont proposés à la fin de chaque section afin de mettre en œuvre les connaissances nouvellement acquises. Le principe de l'exercice avec « fil rouge » consiste à prendre le lecteur par la main et à le guider tout au long de la réalisation du problème en lui fournissant toutes les instructions et commandes Stata utiles.

Le lecteur trouvera non seulement un exposé rigoureux des concepts et méthodes statistiques les plus utilisées dans la recherche clinique et les études épidémiologiques, mais il apprendra aussi à choisir la méthode statistique la plus simple et adéquate pour traiter sa question de recherche, ainsi qu'à valider les hypothèses sur lesquelles ces méthodes reposent. Ce dernier point est très important puisque dans la pratique ces hypothèses ne sont jamais vérifiées stricto-sensu et l'investigateur doit évaluer si elles sont acceptables (i.e. suffisamment vérifiées).

L'ouvrage s'adresse en premier lieu au chercheur dans le domaine des sciences de la santé (médecin, infirmière et infirmier, épidémiologue, biologiste, biostatisticien, etc.), qu'il soit débutant ou qu'il maitrise déjà les concepts de base de la statistique, mais aussi aux chercheurs d'autres domaines (e.g. économie, psychologie, démographie, etc.) qui désirent acquérir les fondements de la statistique. Le lecteur avertit ou le statisticien appréciera certainement la présentation des résultats des nombreuses simulations réalisées afin d'illustrer ce qui se passe lorsque, par exemple, l'hypothèse de Normalité ou de symétrie de la distribution est violée, ou encore lorsque les événements sont rares dans le cas d'une variable dichotomique (i.e. variable codée 0/1). De nombreuses illustrations et nombreux exemples d'application sont donnés et lorsque les méthodes d'analyse standard ne s'appliquent pas des références à des méthodologies plus générales (mais souvent plus complexes) sont fournies.

Le champ couvert par cet ouvrage correspond typiquement à la matière enseignée durant les deux premières années de cours d'introduction à la statistique dans les sciences sociales (économie, démographie, géographie, psychologie, etc.), où l'utilisation de l'outil statistique est probablement plus ancrée dans la tradition des chercheurs de ces domaines qu'il ne l'est actuellement dans les sciences de la santé. Une introduction au logiciel statistique Stata est fournie dans un chapitre en annexe, mais l'apprentissage de l'utilisation du logiciel se fait progressivement au fil des chapitres en réalisant les exercices. Pour la plupart des concepts de base introduits nous avons donné entre parenthèses sa traduction anglaise, pour faciliter le passage à la lecture des ouvrages en langue anglaise, ainsi que celle de l'output de Stata (qui bien entendu est en anglais). Pour les lecteurs qui souhaitent approfondir certaines notions, les sections plus techniques sont indiquées par un astérisque « * ».

Les objectifs d'apprentissage correspondent largement aux titres des différents chapitres et sections qui apparaissent en détail dans le sommaire. Les principaux thèmes abordés dans ce cours sont les notions de :

- variable et variable aléatoire
- résumé quantitatif des données (moyenne, médiane, variance, intervalle interquartile, etc.)
- présentations graphiques (scatter plot, box plot, histogramme, etc.)
- probabilité et dénombrement d'événements
- théorie de l'échantillonnage
- test diagnostique, sensibilité, spécificité, valeurs prédictive positive et négative, rapports de vraisemblance
- espérance mathématique et variance de variables aléatoires
- distribution de probabilité, fonction cumulative, fonction de densité

- théorie de l'estimation
- intervalle de confiance
- intervalle de prédiction
- calcul de puissance et de taille d'échantillon
- tests d'hypothèses
- p-valeur
- tests multiples
- tests de Chi2
- tests non paramétriques
- corrélation linéaire, corrélation de rang, corrélation partielle
- mesures d'association entre variables discrètes
- analyse de corrélation
- analyse de régression linéaire simple
- analyse de régression linéaire multiple
- analyse de variance

En conclusion, cet ouvrage de statistique présente aussi simplement et intuitivement que possible, tout en maintenant un haut niveau de rigueur, les fondements de cette discipline à un public peu habitué à la présentation formelle de la statistique que l'on trouve dans bien des ouvrages d'introduction (e.g. Mood, Graybill, and Boes (1963) ; Introduction to the theory of statistics). Les mathématiques sont le langage de la statistique et c'est volontairement que j'ai décidé de ne pas complètement occulter celles-ci de cette présentation. Bien au contraire, les équations et les formules présentées servent avant tout à « démystifier » l'utilisation du logiciel statistique (montrer ce que Stata a dans le ventre…), qui prend en charge tous les aspects formels et calculatoires, et à éclairer l'utilisateur du programme sur la méthodologie implémentée, les hypothèses sur lesquelles cette méthodologie repose, ainsi que les limites de son application.

Préface par Béatrice Ducot et Faroudy Boufassa

De par le monde il existe de nombreux ouvrages théoriques traitant des Probabilités et/ou de la Statistique pour les sciences de la santé. Le professeur Daniel Schwartz, qui nous a quitté en 2009, avait introduit et développé en France, il y a déjà de nombreuses années, les méthodes de la statistique appliquée à l'usage des médecins et des biologistes. Il est à l'origine de l'engouement de nombreux étudiants pour ces méthodes et le « père » de nombreux chercheurs ayant poursuivi, au-delà de son enseignement, le développement des méthodes d'analyse de données d'origine biomédicales. Patrick Taffé est un biostatisticien Suisse renommé, d'une part dans le milieu du VIH pour les nombreux articles dont il est co-auteur au sein de la cohorte Suisse (Swiss HIV Cohort Study), mais aussi pour les nombreux ouvrages didactiques mélangeant agréablement l'austérité de la théorie au plaisir de la mise en pratique de ces mêmes théories à partir d'exemples très détaillés et surtout très pédagogiques.

Après un bref rappel de l'histoire de la statistique, l'auteur présente les différents types de variables que l'on peut trouver dans le domaine de la santé et les paramètres dont on dispose pour les décrire. Les notions fondamentales de probabilité et de distributions sont ensuite présentées suivies d'éléments d'estimation, toutes ces notions amenant progressivement le lecteur aux tests d'hypothèses. Les fondements de la théorie des tests d'hypothèses sont indispensables à connaître si on veut comprendre le principe des tests statistiques utilisés dans la plupart des articles scientifiques qui paraissent dans les revues biomédicales. Enfin les modèles de régression linéaire simple et multiple sont présentés ainsi que l'analyse de variance. La dernière partie de l'ouvrage propose une introduction à l'utilisation du logiciel d'analyse de données « Stata ».

Toute personne travaillant dans le champ de la santé, que ce soit dans le domaine de la recherche, de la clinique, de la biologie ou de la surveillance est de plus en plus confrontée à la statistique. Le chercheur, pour valider ou rejeter son hypothèse, ne peut se passer d'une démarche statistique tant au moment de rédiger son protocole que lors de l'analyse des données recueillies. Le lecteur d'article scientifique se doit de comprendre le principe des tests et en particulier la signification du "petit p" magique et incontournable lorsqu'il prend connaissance de résultats publiés dans sa revue préférée, son esprit critique peut ainsi s'exercer.

L'auteur a pris le parti de proposer à la fin de chaque chapitre des exercices portant sur les notions présentées. Les corrections des exercices figurent à la fin de l'ouvrage permettant au lecteur de vérifier le niveau de compréhension atteint après chaque étape.

Le choix du logiciel « Stata » pour illustrer l'apprentissage des probabilités et de la statistique pour les sciences de la santé paraît particulièrement judicieux, son utilisation est de plus en plus répandue dans ce domaine et son abord aisé et convivial en font un logiciel extrêmement apprécié par les utilisateurs.

Remercions Patrick Taffé de nous proposer un tel ouvrage et souhaitons bonne lecture à ceux, que nous espérons très nombreux, désirant se former d'une manière agréable et pratique à ces notions.

Béatrice Ducot
Faroudy Boufassa
Ingénieurs de Recherche
INSERM, Le Kremlin-Bicêtre

Comment utiliser cet ouvrage

La meilleure manière de profiter pleinement du contenu de cet ouvrage est de réaliser, après avoir lu la théorie et consulté les exemples d'application, tous les exercices proposés systématiquement à la fin de chaque section. Il s'agit pour la plupart d'exercices à réaliser au moyen de Stata et d'un jeu de données que l'on télécharge simplement dans le logiciel au moyen d'une commande via le web. Pour les premiers chapitres, toutes les commandes Stata nécessaires à la réalisation de l'exercice sont indiquées entre guillemets, par exemple « `use http://www.chuv.ch/cepic/salaires.dta, clear` » pour charger les données, puis graduellement au fil des chapitres certaines commandes (les plus simples et courantes), sont omises de l'énoncé, afin de le raccourcir et tester les connaissances acquises dans l'utilisation

du logiciel. Dans tous les cas, un corrigé (relativement succinct) de l'exercice avec toutes les commandes Stata utiles est proposé dans une annexe.

S'il le désire, le lecteur peut télécharger tous les fichiers de données depuis Stata en exécutant les deux commandes suivantes :

```
« net from http://www.chuv.ch/cepic/ »

« net get Stata »
```

Notations

Signification des divers symboles utilisés :

>>	beaucoup plus grand
$\geq$	plus grand ou égal
>	plus grand
=	égalité
$\equiv$	est défini comme
~	est distribué comme
$\cong$	est distribué approximativement comme
{…}	les accolades permettent de définir un ensemble

Les paramètres sont dénotés au moyen de lettres Grecques :

α (alpha), β (beta), γ (gamma), Δ (delta), ρ (rho), μ (mu), σ (sigma), θ (thêta), κ (kappa), π (pi), etc.

Les variables aléatoires sont dénotées au moyen de lettres majuscules :

X, Y, $\bar{X}$, S_X, etc.

Les observations ou réalisations de variables aléatoires sont dénotées au moyen de lettres minuscules :

x, y, $\bar{x}$, s_X, etc.

Remerciements

Je tiens à remercier avant tout mes collègues Christiane Ruffieux et Alfio Marazi qui ont relu et commenté divers versions de ce manuscrit, ainsi que Arnaud Chiolero qui m'a conseillé pour la mise en page. Mes remerciements vont aussi à Valentin Rousson et à Jean-Luc Bulliard qui ont lu et commenté certaines parties du texte, ainsi qu'à Mohamed Faouzi qui m'a encouragé tout au long de la rédaction de cet ouvrage. Je ne dois pas non plus oublier de mentionner mes élèves des quatre premières volées du Master en Sciences infirmières de

l'Université de Lausanne (soit de 2009 à 2013), ainsi que mon assistant Cédric Mabire, qui par leur volonté de comprendre la matière et leurs nombreuses questions posées durant les cours m'ont incité à inclure de nombreux exemples illustratifs dans cette présentation.

Patrick Taffé
Juin 2014

Merci à ma femme Monika Stuecheli-Taffé, ainsi qu'à mes enfants Juliane et Yanick Taffé, pour leur patience et tolérance pendant toutes ces années où mon ordinateur portable nous a souvent (trop...) accompagné même jusque dans le chalet d'hiver à Adelboden.

P.S. Ceux qui le désirent peuvent me faire part de leurs remarques ou suggestions à l'adresse suivante Patrick.Taffe@chuv.ch pour améliorer/enrichir la présentation dans des éditions ultérieures.