Collection
Statistique et probabilités appliquées

Valentin Rousson

Statistique appliquée aux sciences de la vie

Lavoisier
Hermes

Statistique appliquée aux sciences de la vie

Valentin Rousson

Statistique appliquée aux sciences de la vie

Lavoisier
Hermes

editions.lavoisier.fr

Valentin Rousson
Institut Universitaire de Médecine Sociale et Préventive (IUMSP)
Centre Hospitalier Universitaire Vaudois et Université de Lausanne
Route de la Corniche 10
1010 Lausanne
Suisse

ISBN 978-2-7462-4799-4

Maquette de couverture : Jean-François Montmarché

Collection
Statistique et probabilités appliquées

dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
Suisse
yadolah.dodge@unine.ch

Comité éditorial :

Dans la même collection :

– *Statistique. La théorie et ses applications*
Michel Lejeune, avril 2004

– *Optimisation appliquée*
Yadolah Dodge, octobre 2004

– *Le choix bayésien. Principes et pratique*
Christian P. Robert, novembre 2005

– *Régression. Théorie et applications*
Pierre-André Cornillon, Éric Matzner-Løber, janvier 2007

– *Le raisonnement bayésien. Modélisation et inférence*
Éric Parent, Jacques Bernier, juillet 2007

– *Premiers pas en simulation*
Yadolah Dodge, Giuseppe Melfi, juin 2008

– *Génétique statistique*
Stephan Morgenthaler, juillet 2008

– *Maîtriser l'aléatoire. Exercices résolus de probabilités et statistique, 2e édition*
Eva Cantoni, Philippe Huber, Elvezio Ronchetti, septembre 2009

– *Pratique du calcul bayésien*
Jean-Jacques Boreux, Éric Parent, décembre 2009

– *Statistique. La théorie et ses applications, 2e édition*
Michel Lejeune, septembre 2010

– *Le logiciel R*
Pierre Lafaye de Micheaux, Rémy Drouilhet, Benoît Liquet, novembre 2010

– *Probabilités et processus stochastiques*
Yves Caumel, avril 2011

– *Analyse statistique des risques agro-environnementaux*
David Makowski, Hervé Monod, septembre 2011

Préface

Ce texte d'introduction à la statistique a été initialement écrit en tant que support d'un cours intitulé « Statistique pour biologistes » dispensé aux étudiants de deuxième année du Bachelor ès Sciences en biologie à l'Université de Lausanne (UNIL). Comme les méthodes et la philosophie de la statistique ne diffèrent pas vraiment d'un domaine d'application à un autre, nous avons l'intention de réutiliser tout ou partie de ce support pour d'autres cours d'introduction à la statistique destinés à des étudiants en médecine, pharmacie, neuroscience ou méthodologie clinique. C'est pourquoi ce texte est intitulé « Statistique appliquée aux sciences de la vie ».

Nous avons essayé d'écrire un texte qui soit le plus possible autonome et qui ne repose pas sur une littérature statistique abondante. Nous ne faisons également que peu de référence à l'utilisation d'un logiciel statistique. Les quelques fois où nous l'avons fait, il s'agit du logiciel gratuit R (`www.r-project.org`). Il ne s'agit donc pas d'un manuel d'utilisation de la statistique, mais d'un ouvrage qui doit nous aider à comprendre les principes importants de la statistique.

Les exemples présentés proviennent de différents domaines des sciences de la vie. Certaines des données utilisées sont des données réelles (dont la source est alors indiquée dans le texte), d'autres sont des données fictives. Précisons toutefois que la plupart des analyses présentées sont en grande partie sorties de leur contexte, de sorte que les résultats qui en découlent n'ont ici aucune valeur scientifique sérieuse. Ces exemples servent avant tout à illustrer l'application des méthodes statistiques.

Les étudiants des diverses sciences de la vie ont au moins deux points en commun. Le premier est que les sciences qu'ils étudient sont loin d'être exactes et les méthodes élémentaires de la statistique leur sont particulièrement utiles. Le second est qu'ils sont peu habitués au formalisme mathématique, très présent dans la science statistique et qui a malheureusement mauvaise réputation. Ceci ne peut être cependant qu'un malentendu car le formalisme mathématique ne devrait pas venir compliquer un exposé, mais le préciser et le clarifier. Aussi, bien que nous sommes restés un peu informels sur certains points, notamment en ce qui concerne la définition des variables aléatoires, et plus généralement sur les concepts de probabilités, nous n'avons pas renoncé aux formules mathématiques, notre devise étant qu'« une belle formule vaut mieux que mille mots ». De nombreuses notes de bas de page contiennent des développements

et commentaires qui intéresseront peut-être les statisticiens plus expérimentés et qui permettront de faire le lien avec d'autres ouvrages plus avancés ou mathématiquement plus rigoureux.

Que ce soit avec des mots ou avec des formules, notre but principal est resté toutefois de motiver au mieux l'introduction de chaque nouveau concept statistique, de discuter en détail de son interprétation, de son utilité, de sa valeur ajoutée et de sa relation avec les autres concepts. Nous pensons que la clé de la compréhension de la statistique se trouve dans ce que nous avons appelé le « paradigme de la statistique », à savoir le fait qu'un estimateur peut être vu comme une variable, ce qui permet de faire le lien entre la statistique descriptive et la statistique inférentielle et de donner ainsi à un cours de statistique une certaine unité de doctrine.

Je tiens à remercier Yadolah Dodge qui m'a encouragé à écrire cet ouvrage, Alfio Marazzi qui m'a aidé à simplifier certains passages (dont le titre du livre), Patrick Taffé qui m'a suggéré de nombreuses lectures statistiques intéressantes dont quelques-unes sont citées ici, ainsi que Philippe Vuistiner qui a relu et commenté différentes versions de ce texte. Je remercie également Dieter Häring, Oskar Jenni, Remo Largo et Peter Vollenweider qui m'ont mis à disposition des ensembles de données utilisés dans cet ouvrage.

Valentin Rousson
Décembre 2012

Sommaire

Préface **vii**

Table des matières **ix**

1 Premiers concepts **1**
1.1 Population, variable et échantillon 2
1.2 Échantillonnage et indépendance 3
1.3 Principaux types de variables 4

2 Distribution d'une variable **7**
2.1 Distribution d'une variable qualitative 7
2.2 Distribution d'une variable continue 9
2.3 Densité de probabilité 11
2.4 Boxplot et quantiles 14
2.5 Mesures de tendance centrale 18
2.6 Mesures de variabilité 20
2.7 Changement d'unités 22
2.8 Distribution normale 23
2.9 Distribution normale standardisée 26
2.10 Variable standardisée et qq-plot 29
2.11 Mesures de non-normalité 31
2.12 Transformation logarithmique 33
2.13 Distribution d'une variable binaire 35

3 Estimation **37**
3.1 Distribution d'un estimateur 38
3.2 Variable aléatoire 38
3.3 Distribution de la moyenne d'un échantillon 40
3.4 Distribution de la variance d'un échantillon 44
3.5 Distribution d'une proportion calculée dans un échantillon . . 46
3.6 Distribution exacte d'une proportion calculée dans un échantillon 47

4 Intervalle de confiance **49**
4.1 Méthode de Wald 49
4.2 Intervalle de confiance de Wald pour une moyenne 52
4.3 Intervalle de confiance de Student pour une moyenne 53
4.4 Niveau nominal et niveau réel d'un intervalle de confiance . . . 56
4.5 Intervalle de confiance et intervalle de prédiction 59
4.6 Transformation logarithmique 61
4.7 Intervalle de confiance pour une variance 62
4.8 Intervalle de confiance de Wald pour une proportion 63
4.9 Intervalle de confiance de Wilson pour une proportion 64

5 Comparaison de deux distributions **67**
5.1 Différence de moyenne 68
5.2 Intervalle de confiance de Wald pour une différence de moyenne 70
5.3 Intervalle de confiance de Student pour une différence de moyenne 71
5.4 Intervalle de confiance de Welch pour une différence de moyenne 73
5.5 Validité des intervalles de confiance pour différence de moyenne 73
5.6 Transformation logarithmique 74
5.7 Différence de moyenne standardisée 77
5.8 Quotient de variance 79
5.9 Différence de proportion 81

6 Principe d'un test statistique **83**
6.1 L'hypothèse nulle et l'hypothèse alternative 83
6.2 Erreurs de première et de seconde espèce 84
6.3 Concept de valeur p 85
6.4 Tests multiples 88
6.5 Statistique de test 89

7 Tests du khi-deux pour tables de contingence **91**
7.1 Comparaison de distributions de variables qualitatives 91
7.2 Comparaison d'une distribution qualitative avec distribution de référence 96
7.3 Comparaison de deux proportions 97
7.4 Comparaison d'une proportion avec valeur de référence 100

8 Test statistique sur la valeur d'un paramètre **103**
8.1 Test unilatéral et test bilatéral 103
8.2 Test statistique *versus* intervalle de confiance 108
8.3 Test d'équivalence 111

9 Tests de Wald et de Student **115**
9.1 Méthode de Wald 115
9.2 Test de Wald pour une proportion 116
9.3 Test de Wald pour une différence de proportion 117
9.4 Test de Student pour une moyenne 119

9.5 Test de Welch pour une différence de moyenne 120
9.6 Test de Student pour une différence de moyenne 122
9.7 Test de Student pour données pairées 123

10 Calcul de taille d'échantillon **127**
10.1 Valeur p *versus* taille de l'échantillon 128
10.2 Puissance d'un test statistique 129
10.3 Exemples de calculs de taille d'échantillon 133

11 Tests exacts avec statistique de test discrète **139**
11.1 Test binomial pour une proportion 139
11.2 Comparaison des tests binomial et du khi-deux 142
11.3 Test exact de Fisher 147
11.4 Test de McNemar 150
11.5 Test du signe 152
11.6 Test de Mann-Whitney 153
11.7 Comparaison des tests de Welch, Student et Mann-Whitney . 158
11.8 Test de Wilcoxon 161
11.9 Récapitulatif des tests statistiques 162

12 Analyse de corrélation **163**
12.1 Diagramme de dispersion 163
12.2 Covariance 166
12.3 Corrélation de Pearson 169
12.4 Corrélation *versus* causalité 173
12.5 Corrélation et choix de la population 174
12.6 Distribution normale bivariée 178
12.7 Corrélation de Spearman 180
12.8 Inférence sur la corrélation 184

13 Régression linéaire simple **187**
13.1 Droite de régression 187
13.2 Droite de régression sur la population 192
13.3 Variance prédite et variance résiduelle 193
13.4 Hypothèse de linéarité 196
13.5 Interprétation des paramètres de la droite de régression 200
13.6 Modèle de régression linéaire simple 203
13.7 Inférence sur la droite de régression 206
13.8 Intervalle de prédiction 212
13.9 Régression vers la moyenne 215

14 Régression linéaire multiple **219**
14.1 Hyperplan de régression 220
14.2 Hypothèse de linéarité 222
14.3 Interprétation des paramètres 224
14.4 Ajustement pour les variables confondantes 228

14.5 Corrélation partielle 231
14.6 Modèle de régression linéaire multiple 232
14.7 Analyse des résidus 234
14.8 Inférence sur l'hyperplan de régression 240
14.9 Estimation du pourcentage de la variance prédite 242
14.10 Tests sur la nullité de plusieurs paramètres 243
14.11 Multicolinéarité 246
14.12 Intervalle de prédiction 249
14.13 Choix du modèle 252
14.14 Valeurs aberrantes et points leviers 255

15 Régression avec prédicteurs binaires 259
15.1 Comparaison de deux groupes 259
15.2 Comparaison de deux groupes dans une étude observationnelle 262
15.3 Comparaison de deux groupes dans un essai clinique 264
15.4 Planification d'une expérience 267
15.5 Analyse de variance 270
15.6 Analyse de covariance 273

16 Régression logistique 279
16.1 Odds et odds-ratio 279
16.2 Étude cas-témoins 282
16.3 Inférence sur l'odds-ratio 284
16.4 Régression logistique simple 286
16.5 Régression logistique multiple 291
16.6 Ajustement pour les variables confondantes 296
16.7 Comparaison de deux groupes dans un essai clinique 300
16.8 Sensibilité, spécificité et courbe ROC 302
16.9 Vérification du modèle 307
16.10 Méthode du maximum de vraisemblance 309

A Tableaux 313

Bibliographie 318

Collection
Statistique
et probabilités
appliquées

Cette collection met à la disposition du public intéressé par la statistique (étudiants, enseignants, chercheurs) des ouvrages qui concilient effort pédagogique et travail permanent de mise à jour. Cette démarche implique de prendre en compte de façon sélective et critique les renouvellements des concepts, des champs d'application et des outils de traitement. Seules une compréhension profonde et une appropriation des connaissances permettront de s'adapter aux évolutions qui n'ont pas fini de bouleverser cette discipline.

Valentin Rousson

Statistique appliquée aux sciences de la vie

Cet ouvrage propose une introduction à la statistique sans qu'aucune connaissance préalable ne soit nécessaire. À partir du concept central de « variabilité », l'auteur aborde les notions de distribution, de statistique descriptive, d'estimation, d'intervalle de confiance, de test statistique, de corrélation et de modélisation statistique (régression linéaire et logistique), tout en recherchant un certain équilibre entre une description littérale des concepts et un minimum de formalisme mathématique.

Des problématiques plus techniques comme le calcul de la taille d'un échantillon, la question de la validité d'un intervalle de confiance, le principe d'un test d'équivalence ou le choix d'un modèle de régression sont également présentées.

Ce texte a été écrit à l'intention des étudiants des sciences de la vie (par exemple biologie ou médecine) mais s'adresse aussi aux étudiants et chercheurs d'autres domaines désirant s'initier à la statistique et se préparer dans les meilleures conditions à aborder des ouvrages statistiques plus avancés.

editions.lavoisier.fr

978-2-7462-4799-4