

Olivier Marchal

8

Statistiques appliquées

avec introduction au logiciel R



ellipses

Table des matières

I	Statistiques Descriptives	11
1	Terminologie statistique	11
2	Les grands types de caractères statistiques	13
3	Tri à plat des données	14
4	Grandeurs synthétiques	16
5	Représentations graphiques	18
5.1	Diagramme circulaire	18
5.2	Diagramme en bâtons	19
5.3	Histogramme	20
5.4	Diagramme des fréquences cumulées	21
5.5	Diagramme Moyenne / Ecart-type	21
5.6	Box-plot	22
6	Interprétation et analyse des données	23
7	Logiciel R	24
7.1	L'interface	24
7.2	Fichiers de données et interface graphique	25
7.3	Tableaux et extraction de données	26
7.4	Moyenne, Ecart-type, quantile	26
7.5	Commandes graphiques	28
8	Exercices	30
II	Estimation de paramètres	37
1	Aléatoire et dispersion des mesures	37
2	Le modèle Gaussien	38
3	Principe de l'estimation paramétrique	41
4	Estimation de la moyenne et de la variance	42
5	Estimation d'une proportion	44
6	Modèles non Gaussien	45

6.1	Modèle exponentiel	45
6.2	Modèle discret	47
7	Calculs sur la loi normale	49
8	Exercices	51
III	Intervalles de confiance	57
1	Principe général	57
2	Quantile d'une loi absolument continue	59
3	Le modèle Gaussien	61
3.1	Intervalle de confiance sur la moyenne	61
3.2	Intervalle de confiance sur la variance	63
3.3	Intervalle de confiance sur la différence de deux moyennes	64
4	Intervalle de confiance sur une proportion	65
4.1	Méthode de Wald	65
4.2	Méthode du score de Wilson	66
4.3	Correction de continuité de Yates	67
5	Intervalle de confiance : modèles non-Gaussien	68
5.1	Intervalle de confiance asymptotique pour une espérance ou une variance	68
5.2	Cas d'un modèle explicite non Gaussien	68
6	Exercices	70
IV	Tests d'hypothèses	75
1	Principe général des tests d'hypothèses	75
2	Liste de tests paramétriques usuels	78
2.1	Test de comparaison d'une moyenne à une moyenne fixe a : cas rare où σ est connu	78
2.2	Test de comparaison d'une moyenne à une moyenne fixe a : cas usuel où σ est inconnu	80
2.3	Test de comparaison de deux moyennes	81
2.4	Modèle non Gaussien	82
2.5	Test d'une proportion	83
2.6	Test d'indépendance de deux variables discrètes	84
2.7	Test d'adéquation à une loi discrète : test du Chi-2 de Pearson	86
2.8	Test d'adéquation à une loi continue	87
3	Tests non paramétriques	89
3.1	Test de Wilcoxon	91

3.2	Test exact de Fisher	92
3.3	Autres tests	93
4	Exercices	94
V	Régression linéaire simple	99
1	Cadre général	99
2	Principe des moindres carrés	101
2.1	Minimisation de $U_n(b_0, b_1)$	102
3	Le modèle Gaussien sous-jacent	103
3.1	Présentation du modèle	103
3.2	Exemple	105
3.3	Le coefficient R de la régression	107
4	Intervalles de confiance et de prédiction	108
5	Intervalles de confiance et tests sur les paramètres	110
5.1	Intervalles de confiance sur les paramètres	110
5.2	Test d'existence d'une relation linéaire	111
5.3	Test sur l'ordonnée à l'origine	112
5.4	Commandes en R	113
6	Compléments	115
7	Exercices	116
VI	Régression linéaire multiple	121
1	Cadre général	121
2	Exemple pour $p = 2$	122
3	Représentation matricielle et résolution	124
3.1	Représentation matricielle	124
3.2	Estimation des paramètres	125
3.3	Intervalles de confiance sur les paramètres	126
4	Tests sur les régresseurs	127
4.1	Tests sur un sous-ensemble des régresseurs	127
4.2	Test sur un seul régresseur	127
4.3	Test sur l'ensemble des régresseurs	128
5	Elimination de variables	129
6	Résultats sur l'exemple	130
7	Exercices	133
VII	Test ANOVA à un facteur	137
1	Notation et cadre général	138

1.1	Cadre général	138
1.2	Pourquoi un nouvel outil?	140
2	Le test ANOVA à un facteur	141
2.1	Les différents types de variance	141
2.2	Les degrés de liberté	142
2.3	Test ANOVA	144
3	Etude de l'exemple	144
3.1	Commandes en R	145
4	Exercices	147
VIII Test ANOVA à deux facteurs		151
1	Modèle sous-jacent et notations	151
2	Estimation des paramètres du modèle	154
3	Test ANOVA à deux facteurs	155
3.1	Calcul des SS	155
3.2	Calcul des MS	155
3.3	Test sur l'interaction	156
3.4	Test d'influence du facteur A sur X	157
3.5	Test d'influence du facteur B sur X	157
3.6	Présentation des résultats	158
3.7	Implémentation en R	158
4	Interprétation d'un test ANOVA à deux facteurs	160
5	Exercices	162
IX Rappels de probabilités		165
1	Notions de probabilités	165
2	Probabilités conditionnelles	168
3	Variables aléatoires	170
4	Variables aléatoires discrètes	171
4.1	La loi de Bernoulli	173
4.2	La loi géométrique	173
4.3	La loi binomiale	174
4.4	La loi de Poisson	176
4.5	La loi uniforme discrète	177
4.6	Autres lois discrètes	177
5	Variables aléatoires absolument continues	178
5.1	La loi uniforme continue	179
5.2	La loi exponentielle	180

5.3	La loi normale	181
5.4	Autres lois continues usuelles	184
X	Solutions des exercices	187
1	Statistiques descriptives	187
2	Estimation de paramètres	200
3	Intervalles de confiance	211
4	Tests d'hypothèses	218
5	Régression linéaire simple	231
6	Régression linéaire multiple	242
7	Test ANOVA à un facteur	246
8	Test ANOVA à deux facteurs	250
	Bibliographie	259
	Index	261

1 Terminologie statistique

Comme toute discipline, la statistique possède un langage spécifique qui possède parfois certaines nuances par rapport au sens commun. Pour éviter toute ambiguïté, voici les définitions qui vont être utilisées dans cet ouvrage :

Définition 1.1 Individu, population et échantillon

- *Individu statistique* : On appelle individu statistique ou unité d'observation, "l'objet" observé lors d'une étude (enquête statistique ou expérimentation). Le terme historique d'individu fait référence aux sondages pour lesquels la théorie s'est initialement développée. Cela étant dit, par abus de langage un "individu statistique" n'est pas forcément une personne physique tel que cela est usagé par le sens commun du mot "individu".

Statistiques appliquées avec introduction au logiciel R

Dans cet ouvrage inspiré de son expérience d'enseignant, l'auteur présente, pour un public non spécialiste des mathématiques, les bases des statistiques appliquées modernes ainsi que leur implémentation dans le logiciel libre R. L'ouvrage aborde ainsi l'aspect descriptif des statistiques (représentations graphiques, moyenne, écarts-types empiriques, etc.) puis les intervalles de confiance, les tests d'hypothèses paramétriques et non-paramétriques, les régressions linéaires et les tests ANOVA, qui constituent les outils indispensables de l'analyse de données expérimentales.

Illustré d'exemples types et de nombreux exercices intégralement corrigés, il permet une approche pratique sans négliger le cadre d'emploi ainsi que les limites propres à chaque outil.

Destiné à des étudiants de Licence ou à toute personne souhaitant acquérir rapidement les techniques modernes de statistiques nécessaires afin de réaliser des analyses de données, il propose également, sans technicité excessive, les commandes en R permettant de réaliser chacun des outils abordés et de les adapter rapidement à des cas similaires. Ainsi, il apparaît particulièrement indiqué dans le cadre de la réalisation d'analyses de données pour un stage, un mémoire ou une thèse dans les domaines de la biologie, de la médecine, de la sociologie ou de tout autre domaine des sciences appliquées.

Polytechnicien, Olivier Marchal a obtenu sa thèse de mathématiques appliquées à l'université de Montréal en 2010. Il enseigne depuis 2012 à l'université Jean Monnet. Chercheur à l'institut Camille Jordan de Lyon, il est également auteur de nombreux articles internationaux de recherche et du livre Fondements des probabilités avec exercices corrigés chez le même éditeur.

Du même auteur :

