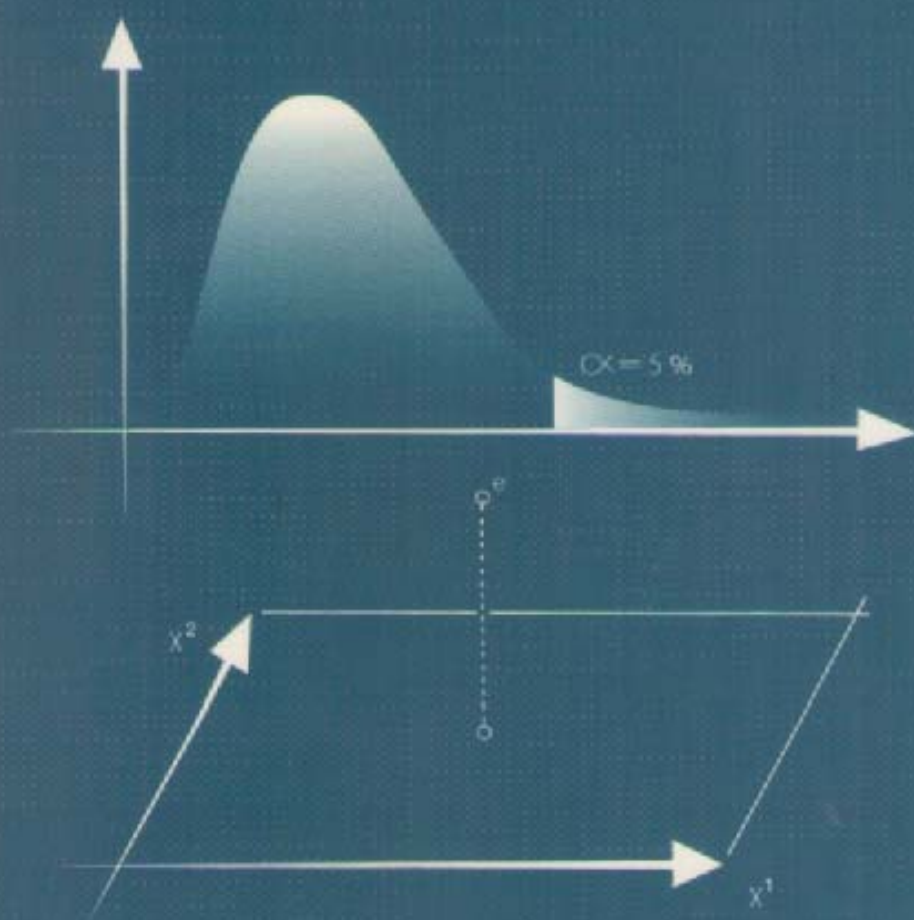


G. SAPORTA

PROBABILITÉS ANALYSE DES DONNÉES ET STATISTIQUE



Editions TECHNIP

2-519-28-1

Gilbert SAPORTAProfesseur au Conservatoire National
des Arts et Métiers

31/05

avant-propos

PROBABILITÉS ANALYSE DES DONNÉES ET STATISTIQUE

1990

t

Editions TECHNIP 27 rue Ginoux, 75737 PARIS Cedex 15, FRANCE

table des matières

Introduction

La statistique, les statistiques et le calcul des probabilités	1
Individus et variables	1
Définitions générales	1
Tableaux de données	2
Statistique et probabilités	3
La démarche statistique	4
La statistique exploratoire	5
La statistique inférentielle	5
Estimation d'une moyenne	5
Vérification d'une hypothèse ou test	6
La modélisation et la prévision statistique	6

Première partie

LES OUTILS PROBABILISTES

Chapitre 1

LE MODÈLE PROBABILISTE

1.1 Espace probabilisable	7
1.1.1 Expérience aléatoire et événements	7
1.1.2 Algèbre des événements	8
1.2 Espace probabilisé	9
1.2.1 L'axiomatique de Kolmogorov	9
1.2.2 Propriétés élémentaires	9
1.3 Lois de probabilités conditionnelles, indépendance	10
1.3.1 Introduction et définitions	10

1.3.2 Indépendance	11
A. Indépendance de deux événements	11
B. Indépendance deux à deux et indépendance mutuelle	12
1.3.3 Formules de Bayes	12
1.4 Réflexions sur le concept de probabilité	14
1.4.1 La conception objectiviste	14
A. La vision classique	14
B. Un paradoxe célèbre	14
C. La vision fréquentiste	16
1.4.2 La conception subjectiviste	16
A. Mesure d'incertitude	17
B. Les bayésiens	17

Chapitre 2

VARIABLES ALÉATOIRES

2.1 Loi de probabilité et moments d'une variable aléatoire réelle	19
2.1.1 Définition et fonction de répartition	19
A. Généralités	19
B. Fonction de répartition	21
C. Variables continues	22
2.1.2 Loi d'une fonction d'une variable aléatoire $Y = \varphi(X)$	23
A. φ bijective	23
B. φ quelconque	24
2.1.3 Indépendance de deux variables aléatoires	25
2.1.4 Moments d'une variable aléatoire	25
A. L'espérance mathématique	25
B. La variance	27
C. Autres moments	29
2.2 Loïs de probabilité d'usage courant	30
2.2.1 Loïs discrètes	30
A. Loi discrète uniforme	30
B. Loi de Bernoulli de paramètre p	31
C. Loi binomiale $\mathcal{B}(n; p)$	31
D. Loi de Poisson $\mathcal{P}(\lambda)$	33
E. Loi hypergéométrique $\mathcal{H}(N, n, p)$ ou du tirage exhaustif	36
F. Loïs géométriques, de Pascal, loi binomiale négative	38
2.2.2 Distributions continues usuelles	39
A. Loi uniforme sur $[0, a]$	39
B. Loi exponentielle de paramètre λ	39
C. Loïs gamma	40
D. Loïs bêta	41
E. La loi de Laplace-Gauss	43

F. La loi log-normale	45
G. Loi de Cauchy	46
H. Loi de Weibull	47
I. Loi de Gumbel	48
2.3 Convolution	49
2.3.1 Cas discret	49
2.3.2 Cas général	50
2.3.3 Applications	51
A. Somme de lois γ	51
B. Somme de lois uniformes sur $[0, 1]$	52
2.4 Fonctions caractéristiques	52
2.4.1 Définitions et principales propriétés	52
A. Définition	52
B. Fonction caractéristique d'une forme linéaire	53
C. Convolution	53
D. Cas d'une distribution symétrique	53
E. Dérivées à l'origine et moments non centrés	53
F. Unicité et inversion de la fonction caractéristique	54
2.4.2 Fonctions caractéristiques des lois usuelles	55
A. Lois discrètes	55
B. Lois continues	55
2.5 Convergences des suites de variables aléatoires	57
2.5.1 Les différents types de convergence	57
A. La convergence en probabilité	57
B. La convergence presque sûre ou convergence forte	57
C. La convergence en moyenne d'ordre p	58
D. La convergence en loi	58
2.5.2 Applications	59
A. Convergence en loi de la binomiale vers la loi de Laplace-Gauss (théorème de De Moivre-Laplace)	59
B. Convergence de la loi de Poisson vers la loi de Gauss	61
C. Somme de variables aléatoires et théorème central-limite	62

Chapitre 3

COUPLES DE VARIABLES ALÉATOIRES CONDITIONNEMENT

3.1 Étude élémentaire d'un couple de variables discrètes	65
3.1.1 Lois associées à un couple (X, Y)	65
A. Loi jointe	65
B. Lois marginales	66
C. Lois conditionnelles	66

3.1.2 Moments conditionnels	67
A. L'espérance conditionnelle	67
B. La variance conditionnelle	68
C. Exemple d'utilisation de l'espérance et de la variance conditionnelle	69
3.1.3 Extension au conditionnement d'une variable continue Y par une variable discrète X ...	71
3.2 Extension à des variables quelconques	72
3.2.1 Lois conjointes et lois marginales d'un couple de variables aléatoires réelles	72
3.2.2 Conditionnement	72
A. Présentation naïve	73
B. Aperçus théoriques	73
C. Ce qu'il faut retenir	74
3.2.3 Application à l'estimation bayésienne	75
3.3 Synthèse géométrique	76
3.3.1 Espace de Hilbert des classes de variables aléatoires de carré intégrables	76
3.3.2 Espérance conditionnelle et projection	78
3.3.3 Corrélation de Y en X	79

Chapitre 4

VECTEURS ALÉATOIRES FORMES QUADRATIQUES ET LOIS ASSOCIÉES

4.1 Généralités sur les vecteurs aléatoires réels	81
4.1.1 Fonction de répartition et densité	81
A. Fonction de répartition	81
B. Densité	81
C. Changement de variables dans une densité	82
4.1.2 Fonction caractéristique	82
4.1.3 Espérance et matrice de variance-covariance	83
A. Définitions	83
B. Transformations linéaires	84
4.2 Vecteurs aléatoires gaussiens : la loi multinormale	85
4.2.1 Définitions et fonction caractéristique	85
4.2.2 Densité de la loi normale à p dimensions	86
4.2.3 Cas particulier de la loi normale à deux dimensions	86
4.2.4 Lois conditionnelles	88
4.2.5 Théorème central-limite multidimensionnel	88
4.3 Formes quadratiques définies sur un vecteur gaussien et lois dérivées	88
4.3.1 Lois du χ^2 (Chi-deux)	88

4.3.2 Formes quadratiques	90
4.3.3 Lois de Fisher-Snedecor et de Student	93
A. Loi du F de Fisher-Snedecor	93
B. Variable de Student	93
4.4 La loi multinomiale. Introduction au test du χ^2	95
4.4.1 Le schéma de l'urne à k catégories	95
4.4.2 Espérance et matrice de variance	96
4.4.3 Lois limites lorsque $n \rightarrow \infty$	97
4.5 Lois de Wishart, de Hotelling, de Wilks	98
4.5.1 Loi de Wishart	98
4.5.2 La loi de T^2 de Hotelling	99
4.5.3 La loi du lambda Λ de Wilks	100

Chapitre 5

NOTIONS ÉLÉMENTAIRES SUR LES PROCESSUS ALÉATOIRES

5.1 Généralités	103
5.1.1 Définitions	103
5.1.2 Moments	104
5.1.3 Continuité, dérivabilité, intégrabilité	105
5.1.4 Stationnarité	105
5.2 Le processus ponctuel de Poisson	107
5.2.1 Flux poissonnien d'événements	107
5.2.2 Étude de la durée T séparant deux événements consécutifs E_i et E_{i+1}	108
5.2.3 Étude de la durée Y séparant $n + 1$ événements	108
5.2.4 Étude du nombre d'événements se produisant pendant une période de durée T fixée	108
5.2.5 Étude de la répartition des dates $E_1, E_2, \dots, E_n$ dans l'intervalle AB	109
5.2.6 Le processus (N_t)	110
5.3 Application aux phénomènes d'attente	110
5.3.1 Les probabilités élémentaires	111
5.3.2 La matrice de transition	111
5.3.3 Les équations d'état	112
5.3.4 Caractéristiques moyennes du phénomène	113

Deuxième partie

LA STATISTIQUE EXPLORATOIRE

Chapitre 6

DESCRIPTION UNIDIMENSIONNELLE DE DONNÉES NUMÉRIQUES

6.1 Les tableaux statistiques	115
6.1.1 Variables discrètes	115
6.1.2 Variables continues ou assimilées	116
6.2 Représentations graphiques	117
6.2.1 Variables discrètes : le diagramme en bâtons	117
6.2.2 Variable continue répartie en classes : l'histogramme et les estimations de densité	118
6.2.3 Le diagramme <i>stem and leaf</i>	121
6.2.4 Courbe de concentration	122
A. Propriétés mathématiques	122
B. Indice de concentration ou indice de Gini	123
6.3 Résumés numériques	125
6.3.1 Caractéristiques de tendance centrale	125
A. La médiane	125
B. La moyenne arithmétique	126
C. Le mode	126
6.3.2 Caractéristiques de dispersion	126
A. L'étendue ou intervalle de variation	126
B. L'intervalle interquartile	126
C. La variance et l'écart-type	127
6.3.3 Cohérence entre tendance centrale et dispersion	128
6.3.4 Caractéristiques de forme	129

Chapitre 7

DESCRIPTION BIDIMENSIONNELLE ET MESURES DE LIAISON ENTRE VARIABLES

7.1 Liaison entre deux variables numériques	131
7.1.1 Étude graphique de la corrélation	131
7.1.2 Le coefficient de corrélation linéaire	132
A. Définition	132
B. Du bon usage du coefficient r	133
C. Matrice de corrélation de p variables	134

7.1.3 Caractère significatif d'un coefficient de corrélation	135
7.1.4 Corrélation partielle	137
A. Le modèle normal à p dimensions	137
B. Corrélation entre résidus	138
C. Signification d'un coefficient de corrélation partielle	138
7.2 Corrélation multiple entre une variable numérique et p autres variables numériques	139
7.2.1 Définition	139
7.2.2 Interprétation géométrique	140
7.2.3 Calcul de R	140
7.2.4 Signification d'un coefficient de corrélation multiple	141
7.3 Liaison entre variables ordinales : la corrélation des rangs	141
7.3.1 Le coefficient de corrélation des rangs de Spearman	141
7.3.2 Le coefficient de corrélation des rangs τ de M. G. Kendall	143
A. Aspect théorique	143
B. Calcul sur un échantillon	143
7.3.3 Coefficients de Daniels et de Guttman	145
7.3.4 Une extension : concordance de p classements; le coefficient W de Kendall	146
7.4 Liaison entre une variable numérique et une variable qualitative	147
7.4.1 Le rapport de corrélation théorique	147
7.4.2 Le rapport de corrélation empirique	148
7.4.3 Interprétation géométrique et lien avec le coefficient de corrélation multiple	149
7.5 Liaison entre deux variables qualitatives	150
7.5.1 Tableau de contingence, marges et profils	150
7.5.2 L'écart à l'indépendance	152
A. Le χ^2 d'écart à l'indépendance et les autres mesures associées	153
B. Cas des tableaux 2×2	154
C. Caractère significatif de l'écart à l'indépendance	155
D. Autres mesures de dépendance	156
7.5.3 Un indice non symétrique de dépendance : le τ_b de Goodman et Kruskal	156

Chapitre 8

DESCRIPTION MULTIDIMENSIONNELLE DE DONNÉES NUMÉRIQUES ANALYSE EN COMPOSANTES PRINCIPALES

8.1 Tableaux de données, résumés numériques et espaces associés	159
8.1.1 Les données et leurs caractéristiques	159
A. Le tableau des données	159
B. La matrice des poids	160

C. Point moyen ou centre de gravité	160
D. Matrice de variance-covariance et matrice de corrélation	160
8.1.2 L'espace des individus	161
A. Le rôle de la métrique	161
B. L'inertie	163
8.1.3 L'espace des variables	164
8.1.4 Variables engendrées par un tableau de données	164
8.2 L'analyse en composantes principales	165
8.2.1 Projection des individus sur un sous-espace	165
8.2.2 Axes principaux, composantes principales, facteurs principaux	167
A. Axes principaux	167
B. Facteurs principaux	169
C. Composantes principales	169
D. Formules de reconstitution	170
8.2.3 Cas usuel. La métrique D_{1/n^2} ou l'ACP sur données centrées-réduites	171
8.3 Interprétation et qualité des résultats d'une ACP	172
8.3.1 Interprétation « interne »	173
A. Corrélations entre composantes et variables initiales	173
B. La place et l'importance des individus	174
C. Effet « taille »	175
8.3.2 Interprétation externe : variables et individus supplémentaires	176
8.3.3 Qualité des représentations sur les plans principaux	176
A. Mesure globale	176
B. Mesures locales	177
C. A propos de la représentation simultanée des individus et des variables en ACP ..	178
8.3.4 Nombre d'axes à retenir	178
A. Critères théoriques	178
B. Critères empiriques	179
8.3.5 Un exemple simple	179
8.4 Analyse factorielle sur tableaux de distance et de dissimilarités	183
8.4.1 Analyse d'un tableau de distances euclidiennes	183
A. La solution classique	183
B. Une transformation permettant de passer d'une distance non euclidienne à une distance euclidienne	184
8.4.2 Une méthode ordinale d'analyse des dissimilarités	185
8.4.3 Analyse de plusieurs tableaux de distances	185



Chapitre 9

L'ANALYSE CANONIQUE ET LA COMPARAISON DE DEUX GROUPES DE VARIABLES

9.1 L'analyse canonique	187
9.1.1 Recherche des variables canoniques	188
A. Étude de la solution dans $\mathbb{R}^p$	189
B. Solutions dans $\mathbb{R}^p$ et $\mathbb{R}^q$	190
9.1.2 Représentation des variables et des individus	191
A. Représentation des variables	191
B. Représentation des individus	192
C. Nombre de variables canoniques significatives	192
9.2 L'analyse canonique généralisée	192
9.2.1 Une propriété de l'analyse canonique ordinaire	193
9.2.2 Généralisation	194
9.3 Méthodes diverses	195
9.3.1 Comparaison de deux configurations d'individus : méthodes procustéennes	195
A. Solution sans contrainte	195
B. T transformation orthogonale	195
9.3.2 Méthodes factorielles	196
A. L'analyse en composantes principales de variables instrumentales (ACPVI)	196
B. ACP sous contrainte d'orthogonalité	198
C. ACP des covariances partielles	198

Chapitre 10

L'ANALYSE DES CORRESPONDANCES

10.1 Tableau de contingence et nuages associés	199
10.1.1 Représentations géométriques des profils associés à un tableau de contingence	199
10.1.2 La métrique du χ^2	201
10.2 Analyses en composantes principales des deux nuages de profils	203
10.2.1 ACP non centrées et facteur trivial	203
10.2.2 ACP non centrées des nuages de profils	204
10.2.3 Formules de transition	205

10.2.4 Trace et reconstitution des données	206
A. Décomposition du ϕ^2	206
B. Formule de reconstitution	207
10.2.5 Un exemple	207
10.3 Analyse canonique de deux variables qualitatives. Justification de la représentation simultanée	210
10.3.1 Mise sous forme disjonctive de données qualitatives	210
10.3.2 Quantification de variables qualitatives	211
10.3.3 Analyse canonique de deux groupes d'indicatrices	211
10.3.4 Représentation simultanée optimale des $(m_1 + m_2)$ catégories d'individus	213
10.3.5 La méthode des moyennes réciproques	215
Conclusion	215

Chapitre 11

L'ANALYSE DES CORRESPONDANCES MULTIPLES

11.1 Présentation formelle	217
11.1.1 Données et notations	217
11.1.2 Une propriété remarquable pour $p = 2$	218
A. AFC formelle du tableau disjonctif	218
B. Propriétés particulières des valeurs propres et vecteurs propres	219
11.1.3 Le cas général $p > 2$	220
A. Coordonnées des catégories	220
B. Coordonnées des individus	221
C. Formules de transition et relations barycentriques	222
D. Propriétés des valeurs propres	223
E. AFC du tableau de Burt	224
11.2 Autres présentations	224
11.2.1 Analyse canonique généralisée de p tableaux d'indicatrices	224
11.2.2 Extension du critère de l'ACP	225
11.2.3 Quantification optimale de variables qualitatives	226
A. ACP de variables quantifiées	226
B. Homogénéité maximale	226
11.2.4 Approximation d'ACP non linéaire	228
11.3 Pratique de l'analyse des correspondances multiples	229
11.3.1 Les contributions	229
A. Contributions à un axe factoriel	229
B. Contributions à l'inertie totale	230

11.3.2 L'usage de variables supplémentaires	231
A. Les deux groupes de variables	231
B. Les tests associés	231
11.3.3 Exemple	232

Chapitre 12

MÉTHODES DE CLASSIFICATION

12.1 Généralités	241
12.1.1 Distances et dissimilarités	241
A. Définitions	241
B. Quelques distances pour données numériques	242
C. Similarités entre objets décrits par des variables binaires	242
D. Accord entre distances, entre partitions	243
12.1.2 Aspects combinatoires de la classification	244
A. Nombre de partitions en k classes $P_{n,k}$ d'un ensemble de n éléments (nombre de Stirling de deuxième espèce)	244
B. Nombre total de partitions P_n (nombres de Bell)	244
12.1.3 Sur l'existence et la caractérisation des classes d'un ensemble	245
12.2 Les méthodes de partitionnement	246
12.2.1 Les méthodes du type «nuées dynamiques»	246
A. Inertie interclasse et inertie intraclasse	246
B. La méthode des centres mobiles	246
12.2.2 Méthodes issues de l'agrégation de relations binaires	249
A. Tableau des comparaisons par paires associées à une relation binaire	249
B. Partition centrale par la règle de la majorité sous contraintes de Condorcet	249
C. Les critères d'association maximale	250
12.3 La classification hiérarchique	251
12.3.1 Aspect formel	252
A. Hiérarchie de parties d'un ensemble E	252
B. Distances ultramétriques	253
12.3.2 Stratégies d'agrégation sur dissimilarités	254
A. Le saut minimum	254
B. Le diamètre	255
C. Stratégies diverses	255
12.3.3 La méthode de Ward pour distances euclidiennes	256
12.3.4 Remarques sur la classification de données qualitatives	257
12.3.5 A propos des algorithmes de classification hiérarchique	258
12.3.6 Classification de variables	258
A. Similarité entre variables de même nature	260
B. L'approche de Lerman et l'algorithme de la vraisemblance du lien	260

Troisième partie

LA STATISTIQUE INFÉRENTIELLE

Chapitre 13

DISTRIBUTIONS DES CARACTÉRISTIQUES USUELLES D'UN ÉCHANTILLON

13.1 Fonction de répartition d'un échantillon, statistiques d'ordre	262
13.1.1 Fonction de répartition empirique d'un échantillon	262
13.1.2 Convergence de $F_n^*(x)$ vers $F(x)$	263
13.1.3 Échantillons ordonnés et lois des valeurs extrêmes	263
A. Loi de $Y_1 = \inf X_i$	264
B. Loi de $Y_n = \sup X_i$	264
C. Loi de l'étendue W	264
D. Loi de Y_k	265
E. Résultats asymptotiques pour les extrêmes	265
F. Distributions asymptotiques des quantiles	266
13.2 Distributions d'échantillonnage de certains moments	266
13.2.1 Étude de la statistique $\bar{X}$	266
A. Propriétés élémentaires	267
B. Lois des grands nombres	267
C. Application : loi d'un pourcentage	268
13.2.2 Étude de la statistique S^2	269
A. Propriétés	269
B. Théorème limite pour S^2	270
C. Corrélation entre $\bar{X}$ et S^2	271
13.2.3 Un résultat utile pour les statistiques asymptotiquement normales : la méthode « delta »	271
A. Stabilisation de la variance d'un pourcentage	272
B. Stabilisation de la variance d'une loi de Poisson	272
C. Valeurs propres d'une matrice de variance	272
13.2.4 Cas des échantillons gaussiens	273
13.3 Distribution du centre de gravité et de la matrice de variance d'un échantillon gaussien p-dimensionnel	276
13.4 Échantillons artificiels, simulation	276
13.4.1 Méthodes d'obtention de nombres pseudo-aléatoires	277
13.4.2 Méthodes générales de tirage d'un échantillon artificiel de n valeurs d'une variable aléatoire X continue	278
A. Inversion de la fonction de répartition	278
B. Méthode du rejet	278
13.4.3 Méthodes particulières de simulation de variables aléatoires	279

A. Variable de Bernoulli X de paramètre p	279
B. Loi γ_p entier	279
C. Loi de Poisson $\mathcal{P}(\lambda)$	280
D. Variable de Laplace-Gauss	280
13.4.4 Calcul d'une intégrale par la méthode de Monte Carlo	281
13.5 Étude empirique des fluctuations d'échantillonnage	282
13.5.1 Population de distribution connue	282
13.5.2 Population de distribution inconnue. La méthode de rééchantillonnage <i>bootstrap</i>	283

Chapitre 14

L'ESTIMATION

14.1 Généralités	285
14.1.1 Exemples élémentaires	285
14.1.2 Qualités d'un estimateur	285
14.1.3 Recherche du meilleur estimateur d'un paramètre θ	287
14.2 L'exhaustivité	287
14.2.1 Définition d'une statistique exhaustive	288
14.2.2 Lois permettant une statistique exhaustive	289
14.2.3 L'information de Fisher	291
14.2.4 Généralisation à plusieurs dimensions, θ paramètre vectoriel $\in \mathbb{R}^k$	293
14.3 L'estimation sans biais de variance minimale	294
14.3.1 Les résultats théoriques	294
14.3.2 Exemple	296
14.3.3 Inégalité de Frechet-Darmois-Cramer-Rao (FDCR)	297
14.4 La méthode du maximum de vraisemblance (MV)	301
14.5 L'estimation par intervalles (les fourchettes d'une estimation)	303
14.5.1 Principe	303
14.5.2 Estimation de la moyenne d'une variable de Laplace-Gauss	304
A. σ connu	304
B. σ inconnu	304
14.5.3 Estimation de σ^2 d'une loi de Laplace-Gauss	305
A. m connu	305
B. m inconnu	305
14.5.4 Intervalle de confiance pour le paramètre d'une loi binomiale quand n est grand	306
14.5.5 Intervalle de confiance pour le paramètre λ d'une loi de Poisson	309
14.5.6 Estimation de la moyenne d'une loi de Gauss multidimensionnelle	309

14.6 Estimation dans une population finie d'effectif N	310
14.6.1 Estimation d'une proportion p	310
A. Tirage avec remise	310
B. Tirage sans remise	310
14.6.2 Estimation de la moyenne m et de l'écart-type σ d'une grandeur X distribuée sur une population finie	310
A. Tirage avec remise	310
B. Tirage sans remise	311
14.7 Remarques sur l'estimation robuste	311
14.7.1 Quelques estimateurs robustes de tendance centrale	311
14.7.2 L'estimateur <i>Jackknife</i>	312
A. Définition	312
B. Réduction du biais	313
C. Intervalle de confiance	313
14.8 Notions élémentaires sur les sondages	313
14.8.1 Méthodes empiriques	314
A. Méthode des quotas ou sondage « raisonné »	314
B. Méthode des unités types	314
14.8.2 Méthodes aléatoires	314
A. Méthode élémentaire	314
B. Le sondage stratifié	314
C. Sondage systématique	315
D. Sondages par grappes	316
E. Sondages à probabilités inégales	316
F. Sondage à plusieurs degrés	316

Chapitre 15

LES TESTS STATISTIQUES

Exemple introductif. Les faiseurs de pluie	317
15.1 Notions générales sur les tests statistiques	319
15.1.1 Risques et probabilités d'erreur	319
15.1.2 Les grandes catégories de tests	321
15.1.3 Choix de la variable de décision et de la région critique optimales : la méthode de Neyman et Pearson	321
15.2 Test entre deux hypothèses simples	322
15.2.1 La méthode de Neyman et Pearson	322
15.2.2 Étude de $1 - \beta$: puissance du test	324
15.2.3 Convergence du test	324
15.2.4 Tests et statistiques exhaustives	324

15.2.5 Exemple	325
15.3 Tests entre hypothèses composites	326
15.3.1 Test d'une hypothèse simple contre une hypothèse composite	326
15.3.2 Test entre deux hypothèses composites	327
15.3.3 Test du rapport des vraisemblances maximales	327
15.3.4 Tests paramétriques usuels	329
A. Valeur moyenne d'une loi de $LG(m, \sigma)$	329
B. Variance d'une loi de $LG(m, \sigma)$	329
C. Test de la valeur théorique p d'un pourcentage pour un grand échantillon	330
15.4 Les tests d'ajustement	331
15.4.1 Les méthodes empiriques	331
A. La forme de l'histogramme	331
B. Vérification sommaire de certaines propriétés mathématiques	331
C. Ajustements graphiques	332
15.4.2 Les tests statistiques généraux	335
A. Le test du χ^2	335
B. Le test d'ajustement de Kolmogorov	336
C. Le test d'ajustement de Cramer-von Mises	338
15.4.3 Exemples d'application en fiabilité et en phénomènes d'attente	339
A. Test du caractère exponentiel d'une loi de fiabilité	339
B. Test du caractère poissonnien des arrivées à une file d'attente	341
15.5 Tests de comparaison d'échantillons	342
15.5.1 Tests de Fisher-Snedecor et de Student pour échantillons indépendants	342
A. Cas de deux échantillons gaussiens $X_1 \in LG(m_1, \sigma_1)$ et $X_2 \in LG(m_2, \sigma_2)$	342
B. Cas d'échantillons non gaussiens	344
15.5.2 Tests non paramétriques de comparaison de deux échantillons indépendants	344
A. Test de Smirnov	344
B. Test de Wilcoxon-Mann-Whitney	345
15.5.3 Test non paramétrique de comparaison de plusieurs échantillons décrits par une variable qualitative : le test du χ^2	347
15.5.4 Test de comparaison de deux pourcentages (grands échantillons)	348
15.5.5 Comparaison de moyennes d'échantillons appariés	349
15.5.6 Comparaison des moyennes de deux échantillons gaussiens indépendants à p dimensions de même matrice de variance	350
A. Notations	350
B. Test de Hotelling	350
C. Distance de Mahalanobis	351
15.6 L'analyse de variance	351
15.6.1 Analyse de variance à un facteur	352
A. Les données et le modèle	352
B. Le test	352
C. L'estimation des effets	354

D. Comparaisons multiples de moyennes	355
E. Test de comparaison de k variances	355
15.6.2 Analyse de variance à deux facteurs	356
A. Le modèle	356
B. L'équation d'analyse de variance et le test	357
C. L'estimation des effets	357
D. Le cas du plan sans répétition	358

Chapitre 16

LA RÉGRESSION SIMPLE

16.1 Le modèle théorique de la régression simple	359
16.1.1 L'approximation conditionnelle	359
16.1.2 Cas où la régression est linéaire	360
16.2 Ajustement du modèle de régression linéaire sur des données expérimentales	361
16.2.1 Estimation de α , β , σ^2 par la méthode des moindres carrés	362
16.2.2 Propriétés des écarts résiduels	365
16.2.3 Cas où le résidu e suit une loi normale	366
16.2.4 Tests dans le modèle linéaire	367
A. Analyse de variance de la régression	367
B. Test d'une équation de régression spécifiée	369
C. Test de linéarité de la régression	369
16.2.5 Contrôle des hypothèses du modèle linéaire	369
A. Le test de Durbin-Watson	370
B. Un cas simple d'hétéroscédasticité	370
16.2.6 Applications	371
A. Exemple	371
B. Prévion d'une valeur ultérieure	373

Chapitre 17

LA RÉGRESSION MULTIPLE ET LE MODÈLE LINÉAIRE GÉNÉRAL

17.1 Régression et modèle linéaire	375
17.1.1 Régression entre variables aléatoires	375
A. Aspect empirique : la recherche d'un ajustement linéaire	375
B. Modèle probabiliste : l'hypothèse de régression linéaire multiple	377

17.1.2 Le modèle linéaire général	378
A. Aspect empirique	378
B. Modèle probabiliste	379
17.1.3 Synthèse	379
17.2 Estimations et tests des paramètres du modèle ($y; X\beta; \sigma^2 I$)	380
17.2.1 Estimation de β et de σ^2	380
A. Propriétés générales	380
B. Propriétés supplémentaires si e est gaussien	382
C. Lois des côtés du triangle rectangle $y, y^*, X\beta$	383
D. Le modèle ($y; X\beta; \Sigma$)	384
17.2.2 Tests dans le modèle linéaire	384
A. Le coefficient de corrélation multiple R et l'analyse de variance de la régression	384
B. Test du caractère significatif d'un des coefficients de régression	386
C. Test de q coefficients de régression, test d'une sous-hypothèse linéaire	386
17.2.3 Intervalle de prévision pour une valeur future	387
17.3 L'analyse des résultats	387
17.3.1 L'étude des résidus et des observations influentes	388
17.3.2 La stabilité des coefficients de régression	389
A. Le facteur d'inflation de la variance (VIF)	390
B. Le rôle des valeurs propres de R	390
17.3.3 Deux techniques pour obtenir des coefficients stables	390
A. Régression sur composantes principales	390
B. La <i>ridge</i> régression	391
17.4 La sélection des variables	391
17.4.1 Les critères de choix	392
17.4.2 Les techniques de sélection	392
A. Recherche exhaustive	392
B. Les méthodes de pas à pas	392
17.5 Un exemple	393
17.5.1 Résultats de la régression complète	393
A. Analyse de variance de la régression	393
B. Estimation des paramètres	394
C. Étude des résidus et de l'influence des observations	395
17.5.2 Recherche d'un modèle restreint	396
17.6 Variables explicatives qualitatives	398
17.6.1 Le principe de la quantification optimale	398
17.6.2 Retour sur l'analyse de la variance	399
17.6.3 Exemple : prix d'une voiture	400
17.6.4 Les méthodes de segmentation	400

Chapitre 18

L'ANALYSE DISCRIMINANTE

18.1 Méthodes géométriques	403
18.1.1 Données et notations	404
18.1.2 L'analyse factorielle discriminante (AFD)	406
A. Les axes et variables discriminantes	406
B. Une analyse en composantes principales (ACP) particulière	408
C. Une analyse canonique particulière	408
D. Analyse de variance et métrique W^{-1}	409
E. Normalisation des vecteurs propres	410
18.1.3 Le cas de deux groupes	410
A. La fonction de Fisher	410
B. Application de l'analyse canonique	411
C. Équivalence avec une régression multiple	412
D. Un exemple	413
18.1.4 Règles géométriques d'affectation	416
A. Règle de Mahalanobis-Fisher	416
B. Insuffisance des règles géométriques	417
18.1.5 Une méthode de discrimination sur variables qualitatives : la méthode Disqual	418
18.2 Méthodes probabilistes	419
18.2.1 La règle bayésienne	419
A. Méthodes « non paramétriques »	419
B. Méthodes paramétriques	420
18.2.2 Le modèle normal multidimensionnel	420
A. Le cas général	420
B. Le cas d'égalité des matrices de variance covariance	420
C. Deux groupes avec égalité des matrices de variance	421
D. Sur certains tests	423
18.2.3 Mesures d'efficacité des règles de classement	424
A. Taux d'erreur théorique pour deux groupes avec $\Sigma_1 = \Sigma_2$ et distribution normale	424
B. La méthode de resubstitution	425
C. Les méthodes de validation croisée	426
18.2.4 La régression logistique	426

ANNEXES

Annexe 1

TABLES USUELLES	431
Table A1.1 : Nombres au hasard	433
Table A1.2 : Loi binomiale : fonction de répartition	434

Table A1.3	: Loi binomiale : probabilités individuelles	439
Table A1.3 bis	: Abaque pour les intervalles de confiance d'une proportion p	440
Table A1.4	: Loi de Poisson	441
Table A1.5	: Loi normale centrée-réduite : fonction de répartition	448
Table A1.5 bis	: Loi normale centrée-réduite : inverse de la fonction de répartition	449
Table A1.6	: Loi du chi-deux	450
Table A1.7	: Loi de Fisher-Snedecor	452
Table A1.8	: Loi de Student	457
Table A1.9	: Valeurs critiques du coefficient de corrélation	458
Table A1.9 bis	: Abaque pour les intervalles de confiance d'un coefficient de corrélation	460
Table A1.10	: Corrélation transformée de Fisher	461
Table A1.11	: Valeurs critiques du coefficient de corrélation des rangs de Spearman	462
Table A1.12	: Test de concordance de p classements (W de Kendall)	464
Table A1.13	: Loi de la statistique de Cramer-von Mises	465
Table A1.14	: Valeurs critiques pour le test de Kolmogorov	466
Table A1.15	: Valeurs critiques du coefficient d'asymétrie	468
Table A1.16	: Valeurs critiques du coefficient d'aplatissement	469
Table A1.17	: Test de Durbin et Watson	470

Annexe 2

FORMULAIRE	471
------------	-----

Annexe 3

CALCUL DES FONCTIONS DE RÉPARTITION DE CERTAINES LOIS CONTINUES	474
---	-----

A3.1 Loi normale centrée-réduite	474
A3.2 Loi du χ_v^2	474
A3.2.1 Formules exactes	474
A. v pair	474
B. v impair	475
A3.2.2 Formules approchées	475
A3.3 Loi de Fisher-Snedecor $F(v_1; v_2)$	475
A3.4 Fonction de répartition de la loi de Student T_n	476
A3.4.1 Formules exactes	476
A. $n = 1$	476
B. $n \geq 2$	476
A3.4.2 Formule approchée	476

Annexe 4

LES FONCTIONS EULÉRIENNES Γ ET B	477
A4.1 La fonction Γ	477
A4.2 La fonction B de deux variables	478

Annexe 5

QUELQUES RÉSULTATS UTILES D'ALGÈBRE LINÉAIRE	480
A5.1 Matrices M-symétriques	480
A5.2 Projecteurs M-orthogonaux	480
A5.3 Pseudo-inverses	481
A5.4 Formules de dérivation vectorielle	482

BIBLIOGRAPHIE	485
----------------------------	-----

INDEX	489
--------------------	-----

Probabilités analyse des données et statistique

G. Saporta

Ce livre présente les méthodes les plus courantes de la statistique moderne sous forme d'un manuel de cours.

Son originalité est de fournir, en un seul volume, l'essentiel des connaissances utiles en calcul des probabilités, analyse des données multidimensionnelles (analyses factorielles, classification) et statistique décisionnelle (tests, estimation, régression, discrimination).

Il est destiné à un vaste public : élèves des grandes écoles, étudiants des maîtrises de mathématiques, d'informatique, de sciences économiques et, plus généralement, à tous ceux qui veulent appliquer en connaissance de cause les méthodes statistiques et possèdent un niveau de connaissances en analyse et en algèbre linéaire équivalant à un premier cycle universitaire.

Sans formalisme excessif et sans tomber dans le recueil de recettes, cet ouvrage est avant tout un guide pour les applications.

Gilbert Saporta est professeur au CNAM. Il enseigne également à l'École nationale de la statistique et de l'administration économique et à l'Institut de statistique de Paris (ISUP). Ingénieur des Arts et Manufactures, il est diplômé de l'ISUP et docteur ès sciences mathématiques de l'Université Pierre et Marie Curie, Paris VI. Au cours des dernières années, il a enseigné à l'IUT de Paris, à l'École du pétrole et des moteurs, à l'École des Mines de Paris et à l'École Centrale. Exerçant des activités de conseil, il est l'auteur de nombreux travaux dans le domaine de l'analyse des données. Ancien président de l'Association pour la statistique et ses utilisations, il est membre de l'Institut international de statistique.



ISBN 2-7108-0565-0