STATISTIQUES POUR LA GESTION

APPLICATIONS AVEC EXCEL ET SPSS

2e édition

Pierre-Charles Pupion

Nouvelle édition revue et augmentée

DUNOD

Statistiques pour la gestion

GESTION SUP

La collection des sciences de gestion

DROIT DE L'ENTREPRISE

Introduction aux droits anglais et américain
Séroussi R.

Introduction au droit comparé
Séroussi R.

STRATÉGIE – MANAGEMENT – RESSOURCES HUMAINES

Entreprises multinationales, *2005*
Meier O., Schier G.,

Gestion des ressources humaines 6e, *2005*
Pilotage social et performances,
Crozet D., Martory B.

Gestion des ressources humaines 3e, *2007*
Pratique et éléments de théorie
Cadin L., Guérin F., Pigeyre F.

L'Entreprise en 20 leçons 4e, *2006*
Conso B., Conso P., Hémici F.

Management de la distribution 2e, *2006*
Cliquet G., Fady A., Basset G.

Management de projet, *2007*
Gray C. F., Langevin Y., Larson E. W.

Management des compétences 2e*, 2005*
Aubret J., Gilbert P., Pigeyre F.

Management du sport 2e, *2006*
Augé B., Tribou G.

Management interculturel 2e, *2006*
Meier O., Schier G.

Méthodes de recherche en management 3e, *2007*
Thiétart R.-A. et coll.

Organisation 3e, *2005*
Livian Y.-F.

Organisations et comportements, *2005*
Gilbert P., Guérin F., Pigeyre F.

Psychologie sociale pour managers, *2006*
Myers D. G.

Statégies, *2005*
Thiétart R.-A., Xuereb J.-M.

Strategor, *2005*
LIPSE

GESTION – FINANCE

Analyse financière 3e, *2006*
de la Bruslerie H.

Finance Corporate, *2005*
Ross S. A., Westerfield R. W., Jaffe J. F.

Fusions acquisitions 2e, *2006*
Meier O., Schier G.

Gestion de la banque 5e, *2007*
de Coussergues S.

Gestion financière 6e, *2005*
Solnik B.

Gestion financière de l'entreprise 11e, *2005*
Conso P., Hemici F.

Introduction à la gestion 2e, *2007*
Calmé I., Ducroux S., Gerbaux F., Hamelin J., Lafontaine J.-P.

Le risque de crédit 3e, *2006*
de Servigny A., Metayer B., Zelenko I.

Pratique des marchés financiers, *2005*
Ogien D.

Statistique, *2007*
Tenenhaus M.

Techniques de gestion 2e, *2007*
Bounab M., Hemici F.

GESTION INDUSTRIELLE

Management de la production 2e, *2005*
Bounab M., Hemici F.

Politique d'achat et gestion des approvisionnements 2e, *2005*
Bruel O.

COMPTABILITÉ – CONTROLE DE GESTION

Comptabilité analytique de gestion 5e, *2007*
Dubrulle L., Jourdain D.

Comptabilité et audit bancaires, *2006*
Ogien D.

Contrôle de gestion et pilotage de l'entreprise 3e, *2006*
Demeestère R., Lorino P., Mottis N..

Instruments financiers et IFRS, *2007*
Barneto P., Gruson P.

Normes IAS/IFRS 2e, *2006*
Barneto P.

Système comptable français et normes IFRS 7e, *2005*
Colette C., Richard J.

MARKETING – COMMUNICATION

Comportements du consommateur, *2003*
Darpy D., Volle P.

Le marketing 3e, *2005*
Jallat F., Lindon D.

Le marketing des nouveaux produits, *2005*
Lenagard E., Manceau D.

Marketing direct 3e, *2005*
Desmet P.

Marketing Research, *2006*
Jolibert A., Jourdan P.

Marketing stratégique et opérationnel 6e, *2005*
Chumpitaz R., de Moerloose C., Lambin J.-J.

Mercator 8e, *2006*
Lendrevie J., Lévy J., Lindon D.

Promotion des ventes 2e, *2007*
Desmet P.

Statistiques pour la gestion

Applications Excel et SPSS

Pierre-Charles Pupion

2e édition

DUNOD

Consultez nos parutions sur dunod.com

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1er juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée. Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).

© Dunod, Paris, 2008
© Dunod, Paris, 2004, pour l'ancienne édition

ISBN 978-2-10-054426-4

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).
Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

AVANT-PROPOS

Cet ouvrage destiné aux étudiants de gestion et d'économie, est le fruit de nombreuses années d'expérience d'enseignement des statistiques en licences et masters d'Université et d'École de commerce.

Il couvre l'intégralité du programme de statistiques que doit acquérir un étudiant de la première année de licence à la cinquième année (Master 1 et 2).

Il s'adresse selon un processus progressif aux étudiants de BTS, d'IUT, de licence d'économie, de gestion, d'AES (première, deuxième et troisième année) ainsi qu'aux étudiants de master (première et deuxième année) d'école de commerce et d'Université.

Les chapitres 1 à 4, accessibles aux étudiants non titulaires d'un baccalauréat scientifique, fournissent les éléments fondamentaux des statistiques descriptives.

Le chapitre 5 permet à l'étudiant d'apprendre la façon d'élaborer un questionnaire et de mesurer sa fiabilité.

Les chapitres 6 à 17 traitent des probabilités, des estimations, des tests d'hypothèses, des tests sur les régressions et présentent les éléments fondamentaux de l'analyse de données. Certains approfondissements comme les outliers ou la régression logistique (chapitre 16) s'adressent plus particulièrement aux étudiants-chercheurs en gestion.

Ce livre est à la fois un manuel et un ouvrage de référence. Il répond aux besoins des professionnels et étudiants en proposant des applications tirées de la vie des affaires et en donnant des exemples de traitement par logiciels de problèmes concrets de gestion. Les exemples sur logiciels dans lesquels sont présentés les menus

© Dunod. La photocopie non autorisée est un délit.

successifs et l'interprétation des résultats guident le profane dans son apprentissage et lui permettent de réaliser facilement des analyses statistiques (analyse financière comparée, études de marché, dépouillement d'enquêtes, prévisions de ventes…).

Chaque chapitre comprend un support théorique (avec les définitions et propriétés essentielles) éclairé immédiatement par des exemples d'utilisation qui permettent à l'étudiant de se familiariser aussitôt avec les notions abordées. L'approche pédagogique de type inductive, caractérisée par la multiplicité des exemples facilite l'appréhension de la matière.

Les exercices en fin de chapitre et leurs corrections figurant en fin d'ouvrage permettent à l'étudiant de vérifier sa compréhension des notions abordées. Les QCM corrigées sont également un bon outil de préparation des étudiants candidats à des examens nationaux tels le test national MESSAGE d'entrée en LSG et LCCA et le test d'entrée en MAE.

TABLE DES MATIÈRES

Avant propos V

1 SÉRIES STATISTIQUES SIMPLES 1

Section 1 Définitions 2
1 La statistique 2
2 Variable statistique 2
3 Série statistique 3

Section 2 Présentation d'une série statistique de variable discrète 4
1 Distribution présentée sous forme de tableau et de graphe 4
2 Fonction de répartition de la variable statistique x 6
3 Valeurs caractéristiques d'une répartition 9

Section 3 Description d'une série statistique issue d'une variable continue 12
1 Représentation de la distribution sous forme de tableau et de graphe 13
2 Fonction de répartition associée à une distribution en classes 16
3 Valeurs caractéristiques de la répartition 18

Section 4 Indicateurs de tendance centrale 20
1 Moyenne arithmétique $\overline{x}$ 20
2 Autres moyennes usuelles 21
3 Le mode 22

Section 5 Indicateurs de dispersion 23
1 La dispersion en termes d'écarts moyens 24
2 La dispersion en termes d'intervalles 26

Section 6 Indicateurs de forme 27
1 Coefficients d'asymétrie et d'aplatissement 27
2 Coefficients de dispersion relative de la distribution 27

© Dunod. La photocopie non autorisée est un délit.

Section 7 Représentations graphiques 28
1 Diagramme circulaire 28
2 Diagramme en barres 28
Section 8 Indicateurs de concentration 29
1 Courbe de Lorentz 31
2 Médiale 32
3 Indices de concentration 33
Section 9 Analyses statistiques avec logiciels 34
1 Traitements avec le logiciel EXCEL 34
2 Traitements avec le logiciel SPSS 38

2 LES SÉRIES STATISTIQUES DOUBLES 43

Section 1 Séries doubles à indices simples 44
Section 2 Séries doubles à double indices 45
1 Cas de variables discrètes 46
2 Cas mixte de variable discrète et continue 50
Section 3 Ajustement linéaire par la méthode des moindres carrés 51
1 Série double à indice simple 52
2 Série double à deux indices 55
Section 4 Ajustements non linéaires de séries doubles 57
1 Ajustement logarithmique 57
2 Ajustement exponentiel 57
Section 5 Analyses statistiques avec EXCEL 58

3 ANALYSE INDICIAIRE DE SÉRIES TEMPORELLES 61

Section 1 Les indices simples 62
1 Indices simples de prix et de volume 62
2 Propriétés des indices simples 63
Section 2 Indices synthétiques 64
1 Indices de volumes 64
2 Indices de prix 65

4 ANALYSE DES SÉRIES CHRONOLOGIQUES 68

Section 1 Les composantes d'une série chronologique 69
Section 2 Les modèles de décomposition 70
Section 3 Détermination des composantes temporelles par méthodes empiriques 71
1 Méthode de la moyenne mobile 72
2 Méthode des pourcentages au trend 75

Section 4 Désaisonnalisation par la méthode analytique de Buys-Ballot 75
1 Modèle additif 75
2 Modèle multiplicatif à trend exponentiel 77
Section 5 Méthode de lissage exponentiel 77
1 Lissage simple pour les séries sans trend 77
2 Lissage de Holt-Winters pour modèle avec trend et saisonnalité 78
Section 6 Traitements avec logiciels 78
1 Traitements avec SPSS 78
2 Traitements avec Excel 80

5 ÉLABORATION ET FIABILITÉ D'UN QUESTIONNAIRE 84

Section 1 Construction d'un questionnaire 85
1 Analyse du problème étudié et pré-test 85
2 Nature des questions 85
3 Catégories de questions fermées 86
Section 2 Ordre des questions et pré-test 90
Section 3 Fiabilité et élaboration des échelles de mesures 91
1 Processus d'élaboration 91
2 Mesures de fiabilité 92
3 Validité 95
Section 4 Procédure sous SPSS 95

6 ANALYSE COMBINATOIRE 97

Section 1 Permutations 97
Section 2 Arrangements 98
1 Arrangements avec répétition 98
2 Arrangements sans répétition 98
Section 3 Combinaisons 99
Section 4 Répartition d'éléments non différentiables 100
Section 5 Formule de Poincaré 100

7 NOTIONS DE PROBABILITÉS 102

Section 1 Notions essentielles 102
1 Référentiel 102
2 Événements 102
Section 2 Probabilité définie sur un référentiel 103
Section 3 Composition d'événements 104
1 Réalisation simultanée ou conjointe de deux événements 105
2 Réalisation d'un événement au moins 105

© Dunod. La photocopie non autorisée est un délit.

3 Événement contraire 105
4 Formule de Poincaré 105
5 Implication, *i.e.* inclusion 106

Section 4 Probabilités conditionnelles 106
1 Définition 106
2 Formule de Bayes 107

8 VARIABLES ALÉATOIRES RÉELLES 109

Section 1 Définition 110

Section 2 Variables aléatoires discrètes 110
1 Lois de probabilité discrètes 110
2 Fonction de répartition de X 111
3 Valeur moyenne m ou espérance mathématique $E(X)$ 112
4 Variance $V(X)$ et autres moments centrés μ_r 113
5 Médiane, p-quantiles d'une loi discrète 113

Section 3 Variables aléatoires continues 114
1 Lois de probabilité continues dont le support est un intervalle 114
2 Valeur moyenne m ou espérance mathématique $E(X)$ 115
3 Variance $V(X)$ et moments centrés μ_r 116
4 Fonction de répartition 116
5 Médiane, p-quantiles d'une loi continue 117

Section 4 Variables aléatoires du type $Y = \phi(X)$ 118

Section 5 Inégalités de Markoff et de Bienaymé-Tchébycheff 119
1 Inégalité de Markoff 119
2 Inégalité de Bienaymé-Tchébycheff 119

Section 6 Variables aléatoires indépendantes 119
1 Couple de *v.a.* réelles indépendantes 119
2 n-uple de *v.a.* réelles indépendantes 119

Section 7 Convergence en probabilité et une loi d'une suite de Z_n de *v.a.* réelles 120

Section 8 Échantillons *iid* 121
1 Définition 121
2 Moyenne $\overline{X}$ d'un échantillon *iid* 121

9 LES PRINCIPALES LOIS DE PROBABILITÉS 123

Section 1 Lois normales 124
1 Loi normale centrée réduite $\mathcal{N}(0; 1)$ 124
2 Loi normale $\mathcal{N}(m; \sigma^2)$ de moyenne m et d'écart-type σ 125

Section 2 Lois discrètes 127
1 Loi de Bernoulli $\mathcal{B}(p)$ 127
2 Loi binomiale $\mathcal{B}(m_0, p)$ 128
3 Loi géométrique $\mathcal{G}(p)$ 130

4 Loi hypergéométrique notée $\mathcal{H}(m_0; n_1, n_2)$ ou $\mathcal{H}(N; m_0, p)$ 131
5 Loi de Poisson $\mathcal{P}(\lambda)$ 132
6 Loi des rangs signés $W^+(n)$ 134

Section 3 Lois continues (suite) 135
1 Loi de Student-Fisher $\mathcal{S}t(n)$ 135
2 Loi du $\chi^2_{(n)}$ (lire khi-deux) à n degrés de liberté 137
3 Loi gamma $\gamma(\lambda, a)$ 138
4 Loi exponentielle Exp(a) 140
5 Loi uniforme $\mathcal{U}(a,b)$ 141
6 Loi de Fisher-Snédécor $\mathcal{F}(m,n)$ 142
7 Loi de Cauchy $\mathcal{C}(m_e, r)$ 143
8 Loi bêta $\beta(a,b)$ 144
9 Loi logistique $\mathcal{L}(m,p)$ 145
10 Loi log-normale ou loi de Galton LN(m,σ) 146
11 Loi de Weibull $\mathcal{W}(a,b)$ 147

Section 4 Lois multinomiales 148
1 Caractérisation et interprétation 148
2 Théorème fondamental 148

Section 5 Procédures avec EXCEL, SPSS 149
1 Traitements avec EXCEL 149
2 Traitements avec SPSS 150

10 ESTIMATION PONCTUELLE ET INTERVALLE DE CONFIANCE 152

Section 1 Variables aléatoires définies sur une population statistique P 153
1 Distribution de la mesure d'un caractère sur une population P 153
2 Tirage au hasard d'un élément dans la population P 154

Section 2 Constitution d'un échantillon 154
1 Échantillon prélevé sur une population statistique 155
2 Échantillon associé à n expérimentations identiquement réalisées 156
3 Caractéristiques des échantillons 157
4 Distributions associées à la moyenne aléatoire $\bar{X}$ d'un échantillon *iid* 157
5 Distributions associées à la variance standard $\boldsymbol{S^2}$ d'un échantillon *iid* 158
6 Distribution associée à la proportion aléatoire $\boldsymbol{F}$ d'un échantillon *iid* 159

Section 3 Estimation ponctuelle des paramètres 159
1 Définition 159
2 Estimateurs ponctuels et maximum de vraisemblance 160
3 Propriétés des estimateurs usuels 160
4 Valeurs des estimations ponctuelles 161

Section 4 Estimation par intervalles de confiance 161
1 Intervalle de confiance de la moyenne m_P et de l'écart-type σ_P d'une distribution lorsque l'échantillon est de grande taille 161
2 Intervalles de confiance de m et de σ lorsque $X_1, X_2, ..., X_n$ est un échantillon *iid* d'une loi normale $\mathcal{N}(m, \sigma^2)$ 165
3 Intervalle de confiance du paramètre d'une loi de Poisson 167

© Dunod. La photocopie non autorisée est un délit.

4 Intervalle de confiance de la valeur moyenne d'une loi exponentielle 168
5 Intervalle de confiance d'une proportion p 169
6 Intervalle de confiance d'un p-quantile ξ_p d'une loi 172

Section 5 Estimation avec SPSS et EXCEL 173
1 Traitements avec EXCEL 173
2 Traitements avec SPSS 174

11 TESTS D'HYPOTHÈSES PARAMÉTRIQUES 178

Section 1 Méthodologie des tests 179
1 Les différents types d'hypothèses 179
2 Méthode de Neyman 180
3 Procédure pour la réalisation d'un test paramétrique 181

Section 2 Tests relatifs à la valeur de la moyenne d'une distribution 182
1 Tests relatifs à la valeur de m_P avec échantillon de grande taille 182
2 Tests relatifs à la valeur de m_P avec échantillon de petite taille 186

Section 3 Tests relatifs à la valeur de l'écart-type d'une distribution 188
1 Tests relatifs à la valeur de σ avec échantillon de grande taille 188
2 Tests relatifs à la valeur de σ avec échantillon de petite taille 190

Section 4 Tests de valeur d'une proportion 191

Section 5 Test de symétrie d'une distribution 194

Section 6 Test de la moyenne avec SPSS et EXCEL 195
1 Traitements avec EXCEL 195
2 Traitements avec SPSS 196

12 TESTS DE COMPARAISON 199

Section 1 Tests paramétriques de comparaison 200
1 Tests paramétriques de comparaison de moyennes 200
2 Tests de comparaison d'écarts-types 205
3 Tests de comparaison de proportions 209
4 Test d'homogénéité de proportions 211

Section 2 Tests de comparaison de deux distributions 212
1 Test du khi-deux 213
2 Test de Wilcoxon-Mann-Whitney 213

Section 3 Tests avec Excel et SPSS 216
1 Traitements statistiques avec EXCEL 217
2 Traitements statistiques avec SPSS 220

13 COUPLES ALÉATOIRES ET TESTS D'INDÉPENDANCE 226

Section 1 Lois bivariées discrètes 227
1 Distribution de probabilité 227
2 Moments centrés et matrice des variance-covariance 229

Section 2 Lois bivariées continues 230

Section 3 Tests d'indépendance par la méthode du khi-deux 230
1 Distribution d'un couple aléatoire sur une population 230
2 Indépendance 231
3 Test d'indépendance par la méthode du khi-deux 231

Section 4 Mesures d'association entre deux variables 234
1 Test de Spearman 234
2 Test de Bloomqvist 236

Section 5 Traitements sous Excel et SPSS 240
1 Traitements statistiques avec SPSS 240
2 Traitements statistiques avec EXCEL 243

14 TESTS D'AJUSTEMENT 245

Section 1 Tests d'ajustement de Kolmogorov-Smirnov 246
1 Ajustement sur une loi continue 246
2 Ajustement sur une loi discrète 247

Section 2 Tests d'ajustement du khi-deux 247

Section 3 Tests d'appartenance à une distribution normale 249
1 Tests de symétrie 249
2 Test du kurtosis 251
3 Statistique de Kolmogorov-Smirnov 252

Section 4 Outliers ou recherche de valeurs discordantes 252
1 Test de valeur discordante d'un échantillon dont la nature de la distribution est inconnue 253
2 Tests de valeurs discordantes sur une distribution supposée normale $N(m,\sigma^2)$ 253

Section 5 Traitement avec Excel et SPSS 255
1 Traitements statistiques avec SPSS 255
2 Traitements sous Excel 256

15 ANALYSE DE VARIANCES 258

Section 1 Analyse de variances à un facteur 258
1 Test de Fisher 255
2 Test par les rangs de Jonckheere-Terpstra 261
3 Test par les rangs de Kruskal-Wallis 262

© Dunod. La photocopie non autorisée est un délit.

Section 2 Analyse de variances à deux facteurs 263
1 Test par les rangs de Friedman 266
2 Test de Fisher 266

Section 3 Test de concordance de Kendall 267

Section 4 Traitements sous Excel et SPSS 269
1 Analyse de variance à un facteur avec Excel 269
2 Analyse de variance à un facteur avec SPSS (test de Fisher) 270
3 Test de Kruskal-Wallis avec SPSS 271
4 Test de Friedman avec SPSS 272

16 TESTS SUR LA RÉGRESSION LINÉAIRE 274

Section 1 Régression d'une *v.a. Y* sur une variable certaine *X* 275
1 Le modèle avec résidus aléatoires 275
2 Tests et intervalles de confiance des coefficients 276
3 Modèle prévisionnel et erreur de prévision 278

Section 2 Régression d'une *v.a. Y* sur une variable aléatoire *X* 279

Section 3 Tests d'autocorrélation des erreurs 280

Section 4 Régression d'une *v.a. Y* sur *k* variables certaines x_i 281
1 Régression linéaire sur *k* variables explicatives 281
2 Le modèle de régression linéaire multiple avec aléas 282
3 Tests et intervalles de confiance des coefficients 283
4 Évaluation prévisionnelle 285

Section 5 Multicolinéarité 286

Section 6 Traitement sous Excel et SPSS 265
1 Traitements statistiques avec Excel 265
2 Traitements statistiques avec SPSS 267

17 MODÈLE LOG-LINÉAIRE ET LOGIT 291

Section 1 Modèle log-linéaire 291
1 Modèle log-linéaire pour un tableau 2×2 291
2 Modèle log-linéaire pour un tableau à $I \times J \times K$ classes 296

Section 2 Les modèles logistiques 298
1 Modèles logit lorsque les variables exogènes sont continues 298
2 Modèle logit avec variables exogènes dichotomiques 300

Section 3 Procédures de traitement sous SPSS 302
1 Modèle log-linéaire 302
2 Modèle Logit 304

18 ACP & AFC 310

Section 1 Analyse en composantes principales 310
1 La Matrice des données 311
2 Nuage de points dans l'espace affine R^h des observations 312
3 Matrice d'inertie et détermination des axes principaux 313
4 Nuage des points individus projetés sur un plan Π 315
5 Nuage des variables et composantes principales 317
6 Représentation simultanée 318
7 Procédure de traitement sous SPSS 318

Section 2 Analyse factorielle des correspondances 321
1 Tableau de contingence 321
2 Nuage des points-lignes dans R^k 323
3 Nuage des points-colonnes dans R^h 326
4 Représentation simultanée 327
5 Procédures de traitement sous SPSS 328

CORRECTION DES EXERCICES 333

ANNEXES 367

Tables 373

Bibliographie 390

Index 391

© Dunod. La photocopie non autorisée est un délit.

1 SÉRIES STATISTIQUES SIMPLES

La statistique descriptive est une méthode de description des faits sociaux utilisant le nombre comme support objectif. La collecte de chiffres destinée à l'étude d'un caractère peut être complète c'est-à-dire étendue à toute la population ou partielle si l'on ne dispose que d'un échantillon. Les méthodes de la statistique descriptive servent à présenter sous forme interprétable un ensemble de données complexes. Elles utilisent pour cela des tableaux, des graphiques, des indicateurs.

Section 1	■	Les définitions des différents concepts employés
Section 2	■	La description d'une série statistique de variable discrète
Section 3	■	La description d'une série dans le cas d'un regroupement des valeurs en classes (cas notamment des variables continues)
Section 4	■	Les indicateurs de tendance centrale (moyenne, mode, . . .) qui permettent « de résumer la série »
Section 5	■	Les indicateurs de dispersion (écart-type, intervalle interquartile . . .)
Section 6	■	Les indicateurs de forme (coefficient d'asymétrie. . .) qui permettent de comparer la dispersion de séries hétérogènes
Section 7	■	Les représentations graphiques : diagrammes en barres et circulaires
Section 8	■	Les indicateurs de concentration qui caractérisent les inégalités dans les répartitions (médiale, coefficient de Gini . . .)
Section 9	■	Analyse des séries statistiques avec EXCEL et SPSS

© Dunod. La photocopie non autorisée est un délit.

Section 1

DÉFINITIONS

Définissons les termes d'unité, de population statistique, de variable statistique et de série statistique.

1 La statistique

La statistique est l'étude soit de phénomènes en général nombreux, préalablement rassemblés et exprimés sous forme numérique, soit de phénomènes ayant fait tout au moins l'objet d'un dénombrement.

L'unité statistique. – L'unité statistique est l'élément de l'ensemble que l'on veut étudier (*exemple* : une automobile est une unité statistique lorsque l'on étudie le parc automobile français). L'ensemble des unités statistiques est une *population statistique* (*exemple* : le parc automobile français).

Le caractère. – Le caractère est l'aspect de l'unité statistique que l'on retient dans l'analyse. Il peut être mesurable (*exemple* : la puissance fiscale d'une automobile) ou seulement dénombrable (*exemple* : la couleur de la carrosserie). On qualifie un caractère mesurable de quantitatif, un caractère seulement dénombrable de qualitatif.

Échantillon. – Il est un sous-ensemble d'une population statistique. L'échantillon est aléatoire lorsque son prélèvement dans la population statistique a été soumis aux lois du hasard.

2 Variable statistique

Elle est l'expression numérique du caractère observé sur les unités statistiques considérées, elle est souvent notée $\boldsymbol{x}$.

– La variable statistique $\boldsymbol{x}$ est dite ***discrète*** lorsqu'elle ne peut prendre que des valeurs numériques isolées : $x_1^*, x_2^*, \ldots, x_K^*$, l'indexation étant telle que $x_1^* < x_2^* < \ldots < x_K^*$ (*exemple* : concernant l'ensemble des assurés d'une société d'assurance automobile, à chaque adhérent est associé le nombre annuel d'accidents déclarés qui est un des entiers 0, 1, 2, 3…).

– La variable statistique $\boldsymbol{x}$ est dite ***continue*** lorsqu'elle peut prendre n'importe quelle valeur d'un intervalle (*exemple* : la distance parcourue par un véhicule au cours d'une année). Dans ce cas l'intervalle $[a,b]$ des valeurs possibles est divisé en K intervalles qui sont appelés classes : $[a_0,a_1],]a_1,a_2],\ldots,]a_{K-2},a_{K-1}],]a_{K-1},a_K]$ où $a = a_o < a_1 < a_2 < \ldots < a_{K-1} < a_K = b$.

3 Série statistique

Par série statistique, on désigne à la fois

– l'ensemble des valeurs $(x_1^*, x_2^*, \ldots, x_K^*)$ (respectivement des classes de valeurs $[a_0, a_1],]a_1, a_2], \ldots,]a_{K-2}, a_{K-1}],]a_{K-1}, a_K]$) de la variable $\boldsymbol{x}$;
– le nombre n_i d'observations associées à chaque valeur x_i^* (respectivement à chaque classe $]a_{i-1}, a_i]$) appelé effectif.

La somme des effectifs des classes est égale à n le nombre total d'observations :

$$n = n_1 + n_2 + \ldots + n_K \quad \text{noté} \quad \sum_{i=1}^{K} n_i$$

Une série statistique peut correspondre à un échantillon ou à la totalité d'une population statistique.

Étude d'une variable discrète

On a soumis un QCM comportant 10 sujets distincts aux 9 étudiants $e_1, e_2, \ldots, e_9$ d'un Master 2e année ayant choisi une option très cotée. À chaque étudiant e_i est associé le nombre de réponses exactes x_i : $x_1 = 5,\ x_2 = 6,\ x_3 = 5,\ x_4 = 6,\ x_5 = 7,\ x_6 = 8,\ x_7 = 9,\ x_8 = 9,\ x_9 = 5$.

Dans cet exemple la *population statistique* $\boldsymbol{P}$, sur laquelle est réalisée l'étude, est l'ensemble des étudiants ayant choisi l'option : $\boldsymbol{P} = \{e_1, e_2, \ldots, e_9\}$. Chaque élément e_i de $\boldsymbol{P}$, en l'occurrence chacun des 9 étudiants, est une *unité statistique*.

La *variable statistique* est l'application $\boldsymbol{x} : \boldsymbol{P} \mapsto \{0,1,2,\ldots,10\}$ qui, à chaque étudiant e_i, associe le nombre $\boldsymbol{x}(ei) = x_i$ de réponses exactes. On constate que $\boldsymbol{x}(\boldsymbol{P}) = \{5,6,7,8,9\}$, autrement dit : $x_1^* = 5, x_2^* = 6, x_3^* = 7, x_4^* = 8, x_5^* = 9$ et le nombre d'observations associées à chacune de ces valeurs est respectivement $n_1 = 3, n_2 = 2, n_3 = 1, n_4 = 1, n_5 = 2$.

Les résultats peuvent être regroupés sous forme d'un tableau statistique où figurent les valeurs observées et les effectifs n_i.

Nombre de réponses exactes x_i^*	5	6	7	8	9
Effectifs n_i	3	2	1	1	2

Lire « 3 étudiants ont eu 5 réponses exactes », « 2 étudiants ont eu 6 réponses exactes », etc.

© Dunod. La photocopie non autorisée est un délit.

Étude d'une variable continue

La société de location d'automobiles LOCAUTO dispose d'un parc de *5 000* véhicules. Sa direction a fait recensé au 1er janvier, par ses diverses agences, la distance annuelle parcourue par chaque véhicule au cours de l'année écoulée. À chaque véhicule automobile e_i de la population statistique $\boldsymbol{P} = \{e_1, e_2, \ldots, e_{5\,000}\}$ est associé sa distance parcourue d_i.

On définit ainsi une application $\boldsymbol{x} : \boldsymbol{P} \mapsto R^+$, $e_i \mapsto \boldsymbol{x}(e_i) = d_i$.

Les résultats obtenus figurent dans le tableau suivant :

Distance parcourue	$0 \leqslant d \leqslant 5$	$5 < d \leqslant 10$	$10 < d \leqslant 20$	$20 < d \leqslant 30$	$30 < d \leqslant 40$
Effectifs	500	2 000	1 500	750	250

Lire « il y a 500 véhicules qui ont parcouru une distance inférieure à 5 000 km », etc.

La variable $\boldsymbol{x}$ est continue, l'intervalle [0,40] des valeurs possibles de $\boldsymbol{x}$ se subdivise en 5 intervalles : $[a_0,a_1] = [0,5]$, $]a_1,a_2] =]5,10]$, $]a_2,a_3] =]10,20]$, $]a_3,a_4] =]20,30]$, $[a_4,a_5[=]30,40]$ et les effectifs respectifs associés sont $n_1 = 500$, $n_2 = 2\,000$, $n_3 = 1\,500$, $n_4 = 750$, $n_5 = 250$. Contrairement à l'exemple précédent, remarquons qu'à deux unités statistiques distinctes e_i et e_j correspondent deux valeurs statistiques d_i et d_j théoriquement différentes si la précision des mesures permettait d'évaluer les mètres, décimètres, centimètres, millimètres, etc.

Section 2

PRÉSENTATION D'UNE SÉRIE STATISTIQUE DE VARIABLE DISCRÈTE

Soit $\boldsymbol{x}$ une variable discrète qui prend les valeurs numériques $x_1^*, x_2^*, \ldots, x_K^*$ (où $x_1^* < x_2^* < \ldots < x_K^*$) et soit n_i les effectifs associés à chaque valeur x_i^*. Le nombre total d'observations $n = n_1 + n_2 + \ldots + n_K$.

1 Distribution présentée sous forme de tableau et de graphe

1.1 Présentation en tableau

La première méthode de description d'un ensemble de données consiste à déterminer la distribution des effectifs ou le nombre d'observations associées à chaque valeur. La série peut être présentée sous la forme d'un tableau où figurent les différentes valeurs x_i^* de la variable et le nombre de fois n_i où ces valeurs ont été observées, le nombre total d'observations étant égal à n.

Autrement dit, dans la population ou l'échantillon analysé, il y a n_i unités dont le caractère étudié mesure x_i^*. (*Cf.* lignes 1 et 2 du tableau ci-après).

La deuxième méthode de description d'une série consiste à indiquer la proportion du nombre total d'observations qui prennent la valeur x_i^*. Cette proportion appelée *fréquence relative* de la valeur x_i^* est notée $f_i = n_i/n$, exprimée en % elle indique le pourcentage d'observation qui prennent la valeur x_i^*.

Tableau 1.1

Valeurs de x	x_1^*	x_2^*	...	x_i^*	...	x_{K-1}^*	x_K^*
Effectifs	n_1	n_2	...	n_i	...	n_{K-1}	n_K
Fréquences relatives	$f_1 = n_1/n$	$f_2 = n_2/n$	...	$f_i = n_i/n$	...	$f_{K-1} = n_{K-1}/n$	$f_K = n_K/n$

La somme des fréquences relatives est égale à 1 soit $\sum_{i=1}^{K} f_i = f_1 + f_2 + \ldots + f_K = 1$.

1.2 Graphe de la distribution de la série statistique

La représentation graphique de la série statistique est souvent plus suggestive que la présentation sous forme de tableaux.

Diagramme en bâtons. Dans un système d'axe cartésien, portant en abscisses les valeurs de la variable $\{x_1^*, x_2^*, \ldots, x_K^*\}$ et en ordonnées les effectifs $(n_1, n_2, \ldots, n_k)$ ou les fréquences relatives $(f_1, f_2, \ldots, f_k)$ associés à chacune de ces valeurs, on obtient le graphe de la distribution appelé diagramme en bâtons (figure 1.1). La longueur du bâton est proportionnelle à l'effectif ou à la fréquence relative.

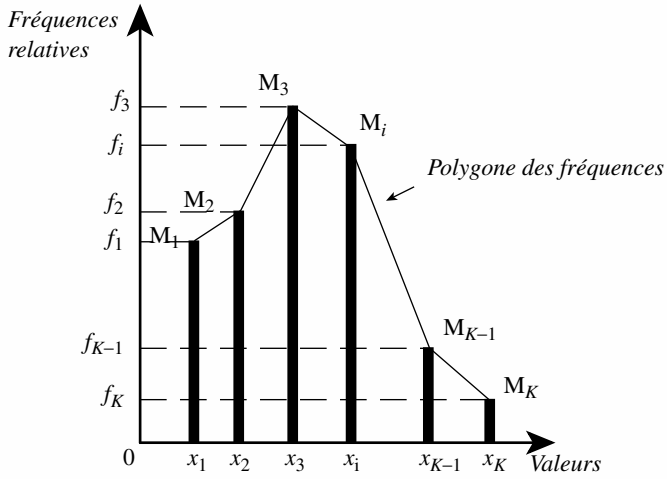


Figure 1.1 – *Diagramme en bâtons et polygone des fréquences relatives*

Le polygone des fréquences est obtenu en joignant les sommets successifs M_i du diagramme en bâtons des fréquences (voir figure 1.1).

© Dunod. La photocopie non autorisée est un délit.

Exemple

On a soumis un QCM comportant 10 questions distinctes aux 500 étudiants qui désirent s'inscrire en 2e année de master. Les 500 résultats sont synthétisés à l'aide du tableau ci-dessous où figurent l'éventail du nombre observé de réponses exactes x_i^*, de son effectif n_i et de sa fréquence relative f_i.

Nombre de réponses exactes x_i^*	5	6	7	8	9	10
Effectifs n_i	40	100	200	50	50	60
Fréquence relative $f_i = n_i/n$	0,08 = 40/500	0,20 = 100/500	0,40 = 200/500	0,10	0,10	0,12

Exprimant les fréquences relatives en %, on déduit du tableau ci-dessus que 8 % des étudiants ont 5 réponses exactes, 20 % ont 6 réponses exactes, etc.

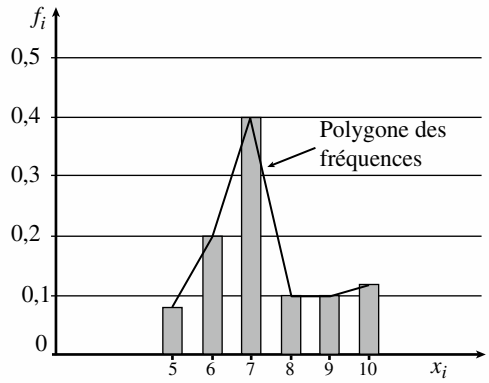


Figure 1.2 – *Diagramme en bâtons et polygone des fréquences relatives des réponses exactes*

Pour obtenir le diagramme en bâtons des fréquences, on porte en abscisse 5 la première valeur de x et on trace un bâton vertical qui a pour extrémité en ordonnée la valeur 0,08 (0,08 étant la fréquence relative correspondant à la valeur 5) puis on porte en abscisse 6 la deuxième valeur de x et on trace un bâton vertical qui a pour extrémité en ordonnée la valeur 0,2 (fréquence associée à la valeur 6) etc.

2 Fonction de répartition de la variable statistique x

Partant de la série statistique on peut déterminer les effectifs cumulés et les fréquences relatives cumulées :

Tableau 1.2

Valeurs de x	Effectifs n_i	Effectifs cumulés	Fréquences relatives f_i	Fréquences relatives cumulées
x_1^*	n_1	$\nu_1 = n_1$	$f_1 = n_1/n$	$\pi_1 = f_1$
x_2^*	n_2	$\nu_2 = n_1 + n_2$	$f_2 = n_2/n$	$\pi_2 = f_1 + f_2$
x_3^*	n_3	$\nu_3 = n_1 + n_2 + n_3$	$f_3 = n_3/n$	$\pi_3 = f_1 + f_2 + f_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_h^*	n_h	$\nu_h = n_1 + n_2 + \ldots + n_h$	$f_h = n_h/n$	$\pi_h = f_1 + f_2 + \ldots + f_h$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_K^*	n_K	$\nu_K = \sum_{i=1}^{K} n_i = n$	$f_K = n_K/n$	$\pi_K = 1$

L'effectif cumulé croissant jusqu'à une valeur x_h^* (ou effectif cumulé des valeurs de $\boldsymbol{x}$ inférieures ou égales à x_h^*) est $\nu_h = n_1 + n_2 + \ldots + n_h$, et est défini pour $h = 1,2,\ldots,K$.

Autrement dit, dans une population statistique P, il y a ν_h unités dont la valeur statistique est inférieure ou égale à x_h^*.

La fraction π_h du nombre total d'observations qui prennent une valeur inférieure ou égale à x_h^* est appelée *fréquence relative cumulée* des valeurs de $\boldsymbol{x}$ jusqu'à x_h^* :

$$\pi_h = \sum_{i=1}^{h} f_i = f_1 + f_2 + \ldots + f_h \text{ ou de façon équivalente } \pi_h = \nu_h/n$$

Exprimée en % elle désigne le pourcentage d'observations qui prennent une valeur inférieure ou égale x_h^*.

Fonction de répartition. La série statistique d'une variable $\boldsymbol{x}$ peut être caractérisée par sa fonction de répartition $F(t)$ qui désigne la proportion de l'ensemble des observations qui prennent une valeur inférieure ou égale à t :

$\forall\, t$ réel, $F(t) = \nu(t)/n$ où n est le nombre total d'observations et

$\nu(t)$ = nombre d'éléments de P dont la valeur statistique est inférieure ou égale à t.

REPÈRES : Propriétés

– La variable discrète $\boldsymbol{x}$ ne prenant aucune valeur inférieure à x_1^* on a $F(t) = 0\ \forall\ t < x_1^*$.
– $\forall\, h = 1, 2, \cdots, K$ on a $F(x_h^*) = \pi_h$. Ainsi $F(x_1^*) = \pi_1$, $F(x_2^*) = \pi_2$ etc.
– La variable discrète $\boldsymbol{x}$ ne prend aucune valeur entre x_i^* et x_{i+1}^*, aussi pour $x_i^* \leqslant t < x_{i+1}^*$ on a $F(t) = F(x_i^*)$ et ce quelque soit i. Ainsi pour $x_1^* \leqslant t < x_2^*$ on a $F(t) = F(x_1^*) = \pi_1$, pour $x_2^* \leqslant t < x_3^*$, on a $F(t) = F(x_2^*) = \pi_2$ etc.
– La variable discrète $\boldsymbol{x}$ ne prend aucune valeur supérieure à x_K^* , aussi $F(t) = 1\ \forall t \geqslant x_K^*$.

© Dunod. La photocopie non autorisée est un délit.

Sa représentation graphique est celle d'une courbe en escalier (voir figure 1.3) où figurent en abscisse les valeurs de la variable et en ordonnée les fréquences relatives cumulées.

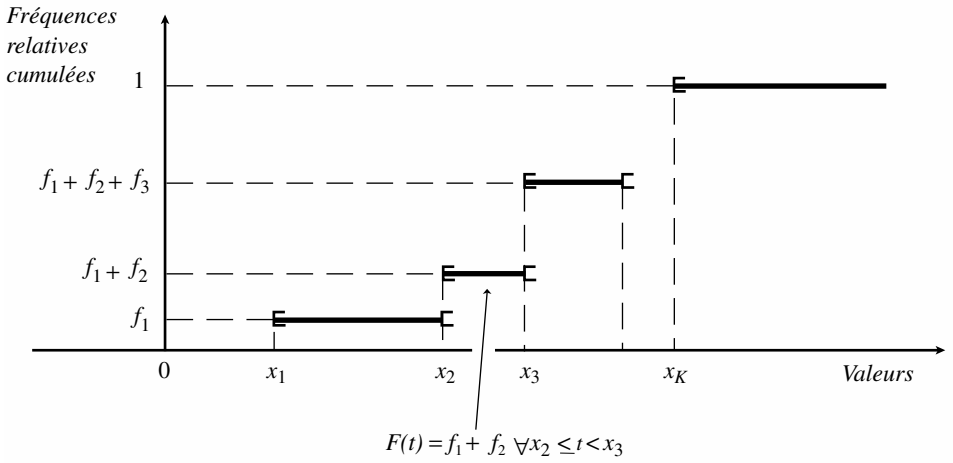


Figure 1.3 – *Diagramme en escalier des fréquences relatives cumulées*

Exprimée en %, $F(t)$ désigne le pourcentage d'observations dont la valeur est inférieure ou égale à t.

Exemple

Dans le cas de l'épreuve de QCM comportant 10 questions distinctes et qui a été proposée aux 500 étudiants désirant s'inscrire en 2e année de master on obtient le diagramme en escalier (*cf.* fig. 1.4) de la fonction de répartition $F(t)$ à partir des valeurs des fréquences relatives cumulées $F(x_i^*)$:

Nombre de réponses exactes x_i^*	5	6	7	8	9	10
Fréquences relatives cumulées $F(x_i^*) = \pi_i$	0,08	0,28	0,68	0,78	0,88	1,00

Ainsi $\forall t < 5 \ F(t) = 0$ aussi on porte un trait horizontal d'ordonnée 0 jusqu'à la valeur 5 exclue ;

– $\forall t$ tel que $x_1^* = 5 \leqslant t < x_2^* = 6$ on a $F(t) = F(x_i^*) = 0{,}08$ donc on porte un trait horizontal d'ordonnée 0,08 débutant en abscisse au point 5 *inclus* symbolisé par le crochet fermé [et allant en abscisse jusqu'au point 6 *exclu*, l'exclusion étant notée [(intervalle [5,6[fermé à gauche et ouvert à droite) ;

– $\forall t$ tel que $x_2^* = 6 \leqslant t < x_3^* = 7$ on a $F(t) = F(x_2^*) = 0{,}28\ldots$;

– $\forall t \geqslant 10$ on a $F(t) = 1$ aussi on porte un trait horizontal d'ordonnée 1 débutant en abscisse au point 10 inclus symbolisé par [.

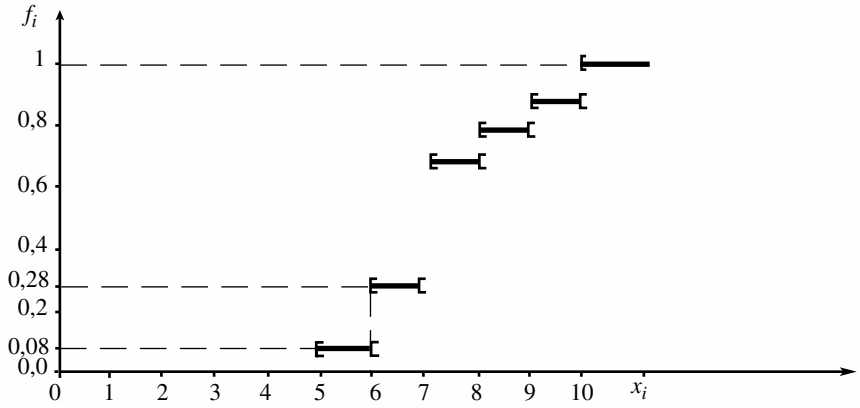


Figure 1.4 – *Diagramme en escalier des fréquences relatives cumulées*

3 Valeurs caractéristiques d'une répartition

La méthode de calcul de la médiane et des quartiles est différente selon que l'on parte de la série des n observations $x_1, x_2, \ldots, x_n$ ou de sa présentation sous forme de tableau où les n observations sont regroupés en K classes $\{x_1^*, x_2^*, \ldots, x_K^*\}$ avec leurs effectifs $n_1, n_2, \ldots, n_K$ $(n = n_1 + n_2 + \ldots + n_K)$.

3.1 Cas d'observations non groupées

Soit une série statistique de n observations $x_1, x_2, \ldots, x_n$ connues, distinctes ou non et classées par ordre de valeurs croissantes : $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$.

La valeur médiane de la série statistique. On appelle *médiane* la valeur m_e telle que : (nombre d'observations inférieures ou égales à m_e) $\geqslant n/2$ et (nombre d'observations supérieures ou égales à m_e) $\geqslant n/2$. Autrement dit, il y a au moins 50 % des observations qui ont une valeur inférieure ou égale à m_e la médiane et il y a au moins 50 % des observations qui ont une valeur supérieure ou égale à m_e.

Sa définition dépend de la parité de n :

– si n est un nombre impair $(n = 2p + 1)$, la *valeur médiane* $m_e = x_{p+1}$;

– si n est un nombre pair $(n = 2p)$, la *valeur médiane* $m_e = (x_p + x_{p+1})/2$.

La valeur médiane présente l'avantage d'être unique et de ne pas être affectée par les valeurs extrêmes.

Exemple

Afin de réaliser une étude sur la rentabilité sectorielle dans le secteur S1, une enquête permet d'obtenir les taux de rentabilité en % de 8 entreprises : 2,5 , 3,0 , 3,1 , 4,1 , 5,4 ,

© Dunod. La photocopie non autorisée est un délit.

4,5 , 5,2 , 2,5. La valeur médiane de cet échantillon est obtenue en classant ces 8 valeurs par ordre croissant :

Rang de l'observation *i*	1	2	3	4	5	6	7	8
Rentabilité financière (en %) x_i	2,5	2,5	3,0	3,1	4,1	4,5	5,2	5,4

Le nombre d'observations $n = 8$ étant un nombre pair ($n = 2p$ avec $p = 4$) la *valeur médiane* m_e correspond à $(x_p + x_{p+1})/2$ soit $(x_4 + x_5)/2 = (3{,}1 + 4{,}1)/2 = 3{,}6$.

Pour comparer la rentabilité des firmes de ce secteur S1 à la rentabilité des firmes du secteur S2, on a prélevé dans ce dernier un échantillon de 9 entreprises : 3,0, 3,5 , 3,8, 4,1, 4,1, 4,5, 5,5, 5,5, 4,2 . Pour obtenir la valeur médiane de ce deuxième échantillon, on réécrit les valeurs de cette série dans l'ordre croissant :

Rang de l'observation *i*	1	2	3	4	5	6	7	8	9
Rentabilité financière (en %) x_i	3,0	3,5	3,8	4,1	4,1	4,2	4,5	5,5	5,5

La taille de l'échantillon étant un nombre impair ($n = 2p + 1$ avec $p = 4$), la valeur médiane m_e correspond à x_{p+1} : $m_e = x_5 = 4{,}1$.

Les quartiles de la série statistique. Les trois quartiles q_1, q_2 et q_3 ($q_1 \leqslant q_2 \leqslant q_3$) sont des nombres qui partagent les n observations en 4 classes qui ont chacune un effectif sensiblement égal à $n/4$. Le second quartile coïncide évidemment avec la médiane.

• *Premier quartile*. Le premier quartile q_1 est défini comme suit :

$q_1 = (x_{n/4} + x_{n/4+1})/2$ si $n/4$ est un entier,

$q_1 = x_{[n/4]+1}$ (où $[n/4]$ est la partie entière de $n/4$) si $n/4$ n'est pas un entier.

On a la propriété : « (nombre d'observations inférieures ou égales à q_1) $\geqslant 0{,}25n$ et (nombre d'observations supérieures ou égales à q_1) $\geqslant 0{,}75n$ ». Autrement dit, il y a au moins 25 % des observations qui ont une valeur inférieure ou égale à q_1 et il y a au moins 75 % des observations qui ont une valeur supérieure ou égale à q_1.

• *Troisième quartile.* Le troisième quartile q_3 est défini comme suit :

$q_3 = (x_{3n/4} + x_{3n/4+1})/2$ si $3n$ est divisible par 4

$q_3 = x_{[3n/4]+1}$ (où $[3n/4]$ est la partie entière de $3n/4$) si $3n$ n'est pas divisible par 4.

On a la propriété : « (nombre d'observations inférieures ou égales à q_3) $\geqslant 0{,}75n$ et (nombre d'observations supérieures ou égales à q_3) $\geqslant 0{,}25n$ ».

Exemple

Ainsi pour l'échantillon des 8 firmes issues de S1, le nombre d'observations n étant égal à 8, $n/4$ est l'entier 2.

Par suite $q_1 = (x_{n/4} + x_{n/4+1})/2 = (x_2 + x_3)/2 = (2{,}5 + 3)/2 = 2{,}75$.

$3n/4$ étant l'entier 6, on a $q_3 = (x_{3n/4} + x_{3n/4+1})/2 = (x_6 + x_7)/2 = (4{,}5 + 5{,}2)/2$.

Pour les 9 firmes issues de S2 le nombre d'observations n n'est pas divisible par 4 car $n/4 = 9/4 = 2{,}25$. Et $[n/4]$ la partie entière de $n/4$ étant égale à 2, on a $q_1 = x_{[n/4]+1} = x_3 = 3{,}8$. De même, on a $q_3 = x_{[3n/4]+1} = x_7 = 4{,}5$.

Les quartiles sont représentés par le logiciel SPSS sous la forme d'une boîte à moustaches.

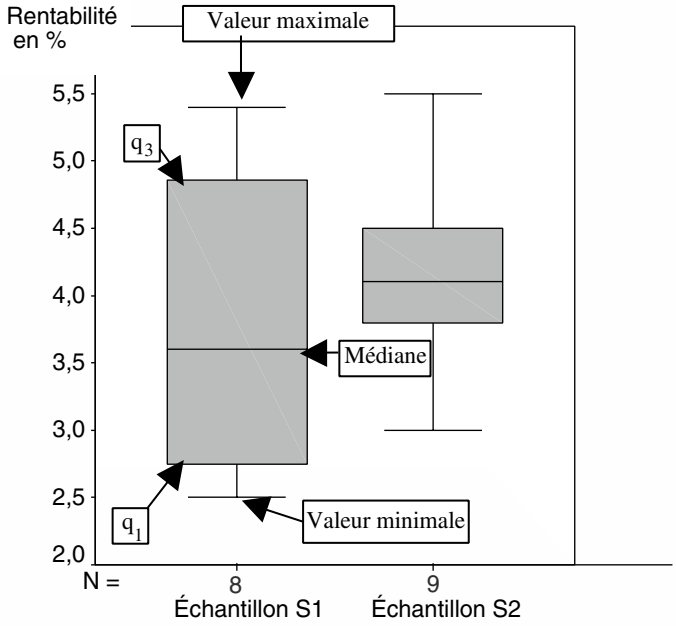


Figure 1.5 – *Boîte à moustaches*

Déciles, centiles. Lorsque n est grand, on définit :

– 9 déciles d_1, d_2,..., d_9 qui partagent les n observations en 10 classes qui ont chacune un effectif sensiblement égal à $n/10$ soit pour $h = 1,2,\ldots,9$
$d_h = x_{[hn/10]+1}$ si $hn/10$ n'est pas un entier et $d_h = (x_{hn/10} + x_{hn/10+1})/2$ si $hn/10$ est un entier ;

– 99 centiles $c_1, c_2, \ldots, c_{99}$ qui partagent les n observations en 100 classes qui ont chacune un effectif sensiblement égal à $n/100$ soit pour $h = 1,2,\ldots,99$ $c_h = x_{[hn/100]+1}$ si $hn/100$ n'est pas un entier, $c_h = (x_{hn/100} + x_{hn/100+1})/2$ si $hn/100$ est un entier.

3.2 Cas d'observations groupées par valeurs

Dans le cas où les n observations d'une variable discrète sont regroupées sans perte d'informations en K classes $\{x_1^*, x_2^*, \ldots, x_K^*\}$ où $x_1^* < x_2^* < \ldots < x_K^*$ (*cf.* tableau 1.2), les quartiles et les médianes se déterminent à partir des valeurs des K fréquences relatives cumulées $\pi_h = f_1 + f_2 + \ldots + f_h$.

© Dunod. La photocopie non autorisée est un délit.

Pour la médiane m_e deux éventualités sont à envisager :
– il existe un entier i_0 tel que $\pi_{i_0} > 0{,}5$ et $\pi_{i_0-1} < 0{,}5$ alors $m_e = x^*_{i_0}$;
– il existe un entier i_0 tel que $\pi_{i_0} = 0{,}5$, alors on a $m_e = (x^*_{i_0} + x^*_{i_0+1})/2$.

Dans le cas du premier quartile q_1 deux éventualités sont à envisager :
– il existe un entier i_0 tel que $\pi_{i_0} > 0{,}25$ et $\pi_{i_0-1} < 0{,}25$ alors $q_1 = x^*_{i_0}$;
– il existe un entier i_0 tel que $\pi_{i_0} = 0{,}25$, alors on a $q_1 = (x^*_{i_0} + x^*_{i_0+1})/2$.

Dans le cas du troisième quartile q_3, le raisonnement est le même que pour le premier quartile, le seuil étant cependant de 0,75 au lieu de 0,25.

Exemple

En consultant le tableau de la page 8 on souhaite connaître la valeur médiane et les autres quantiles des 500 notes relatives à l'épreuve de QCM.

On observe que la fréquence relative cumulée des valeurs inférieures à $x^*_3 = 7$ est $\pi_3 = 0{,}68 > 0{,}5$ et que la fréquence relative cumulée des valeurs inférieures à $x^*_2 = 6$ est $\pi_2 = 0{,}28 < 0{,}5$ donc la médiane est égale à x^*_3 : $m_e = 7$. De même,

– constatant que $\pi_2 = 0{,}28 > 0{,}25$ et $\pi_1 = 0{,}08 < 0{,}25$, on déduit que le premier quartile est égal à $x^*_2 = 6$;
– observant que $\pi_4 = 0{,}78 > 0{,}75$ et $\pi_3 = 0{,}68 < 0{,}75$, le troisième quartile est $x^*_4 = 8$.

Section 3 DESCRIPTION D'UNE SÉRIE STATISTIQUE ISSUE D'UNE VARIABLE CONTINUE

Les précédentes définitions concernant **une série de n observations $x_1, x_2, \ldots, x_n$ non regroupées en classes** s'appliquent à un échantillon de n valeurs d'une variable continue, valeurs théoriquement distinctes les unes des autres.

Mais lorsque n est grand, l'analyse de la série non groupée est fastidieuse, il convient alors de subdiviser l'intervalle (a,b), champ des valeurs numériques de x, en K classes consécutives d'amplitudes égales ou non : $[a_0,a_1],]a_1,a_2], \ldots,]a_{K-2},a_{K-1}],]a_{K-1},a_K]$ avec $a_0 = a$ et $a_K = b$ et de grouper toutes les observations qui appartiennent à une même classe. Si dans la série des n observations, on a trouvé n_i valeurs comprises entre a_{i-1} et a_i alors n_i représente l'effectif de la i-ème classe $]a_{i-1},a_i]$.

1 Représentation de la distribution sous forme de tableau et de graphe

1.1 Tableau de la série

La distribution de la variable statistique peut être présentée sous la forme de tableaux où figurent les effectifs n_i de chaque classe ou les fréquences relatives f_i.

Tableau 1.3

Valeurs de x	$a_0 < x \leqslant a_1$	$a_1 < x \leqslant a_2$	...	$a_{i-1} < x \leqslant a_i$	...	$a_{K-1} < x \leqslant a_K$
Effectifs	n_1	n_2	...	n_i	...	n_K
Fréquences relatives f_i	$f_1 = n_1/n$	$f_2 = n_2/n$	...	$f_i = n_i/n$	...	$f_K = n_K/n$

– n_i est l'effectif de la classe $]a_{i-1}, a_i]$.

– La *fréquence relative* de la i-ème classe $]a_{i-1}, a_i]$ est $f_i = n_i/n$.

Exemple

Considérons à un instant donné l'ensemble P du parc automobile de l'entreprise LOCAUTO de locations de voitures. Le tableau récapitulatif des valeurs $x_i = x(e_i)$ des distances parcourues (en milliers de km) par ses 5 000 véhicules $e_1, e_2, \ldots, e_{5\,000}$ est rappelé ci-dessous.

Répartition des valeurs de x	$0 \leqslant x \leqslant 5$	$5 < x \leqslant 10$	$10 < x \leqslant 20$	$20 < x \leqslant 30$	$30 < x \leqslant 40$
Effectifs n_i	500	2 000	1 500	750	250
Fréquences relatives $f_i (= n_i/n)$	0,10	0,40	0,30	0,15	0,05

Ainsi 10 % des véhicules ont parcouru une distance comprise en 0 et 5 000 km, 40 % (0,4 × 100 %) des véhicules ont parcouru une distance comprise en 5 000 et 10 000 km.

1.2 Histogramme

Il est suggestif de représenter dans un repère orthogonal l'effectif (ou la fréquence relative) de chaque classe par *la surface d'un rectangle* qui lui est égale ou proportionnelle. Cette représentation est nommée histogramme.

- Construction de *l'histogramme des effectifs* (voir figure 1.6). L'effectif n_i de la i-ème classe $]a_{i-1}, a_i]$ devant être égal à la surface du rectangle $A_{i-1}A_i\,B_iB_{i-1}$, on porte en abscisse les valeurs a_{i-1}, a_i et en ordonnée la valeur $y_i = n_i/(a_i - a_{i-1})$. L'aire du rectangle $A_{i-1}A_i\,B_iB_{i-1}$ est en effet alors égale à n_i.
- Il est également possible de construire *l'histogramme des fréquences* en portant en ordonnée $y'_i = f_i/(a_i - a_{i-1})$. Remarquons alors que dans un repère orthonormé, la surface totale du domaine hachuré (voir figure 1.7) est égale à 1 puisque $\sum_{i=1}^{K} f_i = 1$.

© Dunod. La photocopie non autorisée est un délit.

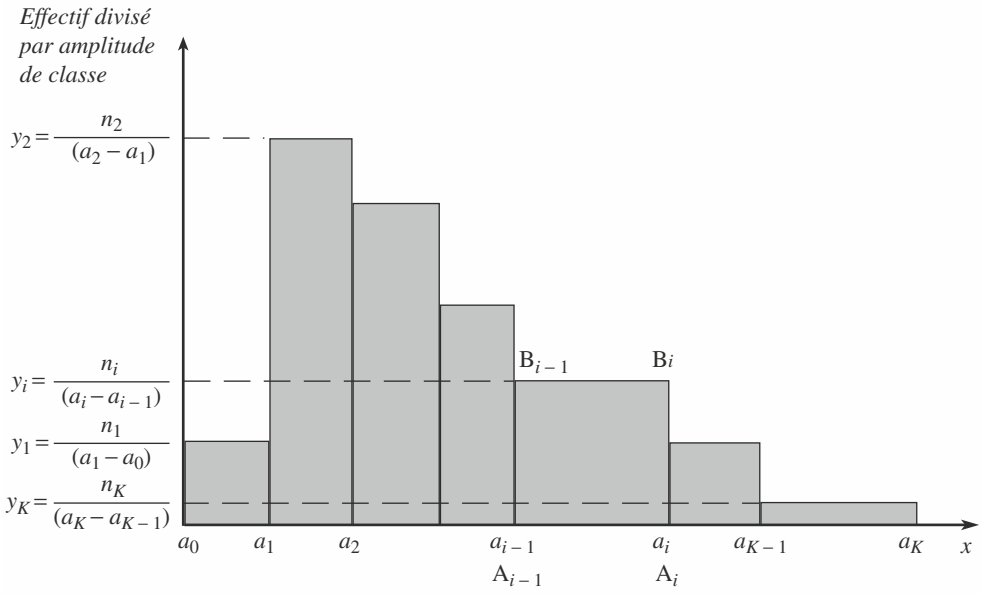


Figure 1.6 – *Histogramme des effectifs*

Exemple

La construction manuelle des histogrammes des fréquences associées au tableau récapitulatif des valeurs de $\boldsymbol{x}$ concernant les distances parcourues par les 5 000 véhicules de la société LOCAUTO, est réalisée à partir des valeurs y_i' ci-après.

Distance (en 10^3 km)	$0 \leqslant x \leqslant 5$	$5 < x \leqslant 10$	$10 < x \leqslant 20$	$20 < x \leqslant 30$	$30 < x \leqslant 40$
Effectifs n_i	500	2 000	1 500	750	250
Fréquences relatives f_i	0,10	0,40	0,30	0,15	0,05
Hauteurs des classes $y_i' = f_i/(a_i - a_{i-1})$	0,02	0,08	0,03	0,015	0,005

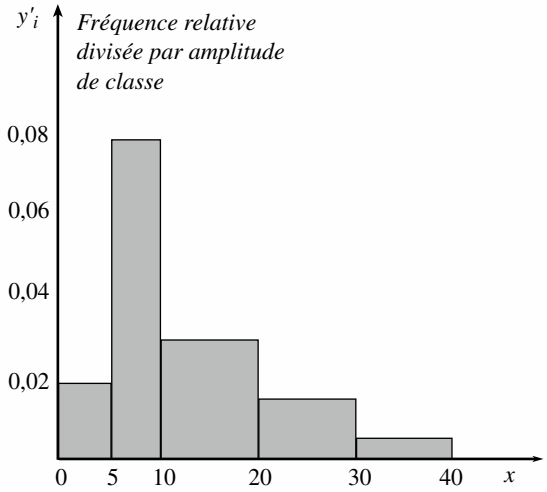


Figure 1.7 – *Histogramme des fréquences relatives*

Ainsi pour la première classe on porte en abscisse 0 et 5 et en ordonnée $y'_1 = 0{,}02$, pour la deuxième classe on porte en abscisse 5 et 10 et en ordonnée $y'_2 = 0{,}08$.

1.3 Polygone des fréquences

• *Lorsque les classes* $[a_0,a_1[,[a_1,a_2[,\ldots,[a_{K-2},a_{K-1}[,[a_{K-1},a_K]$ *ont une même amplitude* h ($a_i - a_{i-1} = h$ une constante positive), on joint par des segments de droites les milieux des sommets des rectangles de l'histogramme des fréquences.

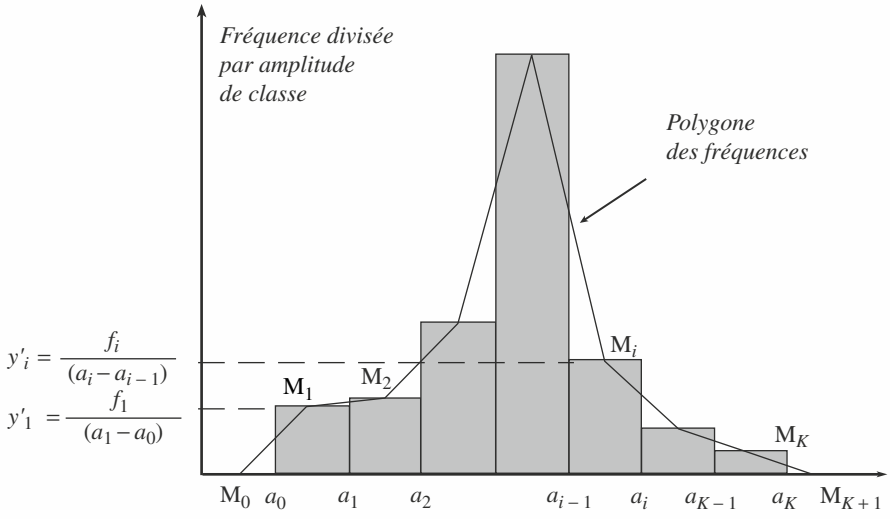


Figure 1.8 – *Polygone des fréquences*

© Dunod. La photocopie non autorisée est un délit.

Pour chaque $i = 1,2,\ldots K$, on construit les segments $[M_i M_{i+1}]$ où le point M_i a pour abscisse $x_i = (a_{i-1} + a_i)/2$ et pour ordonnée $y'_i = f_i/(a_i - a_{i-1})$. Aux extrémités on ajoute les segments $[M_0 M_1]$ et $[M_K M_{K+1}]$ où M_0 et M_{K+1} ont respectivement pour coordonnées $(a_0 - h/2,0)$ et $(a_K + h/2,0)$. Voir figure 1.8.

- *Lorsque les classes n'ont pas la même amplitude*, on considère parmi les K intervalles, celui ou ceux qui ont la plus petite longueur θ.
  Une fois déterminé $\theta = \inf_i\{(a_i - a_{i-1})\}$ et les valeurs $y'_i = f_i/(a_i - a_{i-1})$ pour $i = 1,2,\ldots K$, on construit le polygone des fréquences en joignant successivement par des segments de droites les points $M_0 = (a_0 - \theta/2,0)$,
  $M_1 = (a_0 + \theta/2, y'_1)$, $M'_1 = (a_1 - \theta/2, y'_1)$,
  $M_2 = (a_1 + \theta/2, y'_2)$, $M'_2 = (a_2 - \theta/2, y'_2)$, ...,
  $M_{K-1} = (a_{K-1} + \theta/2, y'_K)$, $M'_K = (a_K - \theta/2, y'_K)$, $M_{K+1} = (a_K + \theta/2,0)$.

La surface du domaine délimité par le polygone des fréquences et l'axe Ox est égale à 1.

2 Fonction de répartition associée à une distribution en classes

Du tableau de la distribution de la variable statistique on déduit, pour chaque a_i, les expressions des effectifs cumulés, des fréquences relatives cumulées et donc de la fonction de répartition F.

Tableau 1.4

Valeurs de x	$a_0 < x \leqslant a_1$	$a_1 < x \leqslant a_2$	...	$a_{h-1} < x \leqslant a_h$	...	$a_{K-1} < x \leqslant a_K$
Effectifs n_i	n_1	n_2	...	n_h	...	n_K
Effectifs cumulés	$\nu_1 = n_1$	$\nu_2 = n_1 + n_2$	...	$\nu_h = n_1 + n_2 + \ldots + n_h$	...	$\nu_K = n$
Fréquences relatives f_i	$f_1 = n_1/n$	$f_2 = n_2/n$	...	$f_h = n_h/n$	...	$f_K = n_K/n$
Fréquences relatives cumulées	$\pi_1 = f_1$	$\pi_2 = f_1 + f_2$	...	$\pi_h = f_1 + f_2 + \ldots + f_h$	...	$\pi_K = 1$

Dans la population (ou l'échantillon), il y a n_i unités dont la mesure du caractère est comprise entre a_{i-1} et a_i

- Le nombre cumulé d'observations qui ont une valeur inférieure ou égale à a_h étant notée ν_h : $\nu_h = \sum_{i=1}^{h} n_i = n_1 + n_2 + \ldots + n_h$.

  Cette expression définie pour $h = 1,2,\ldots,K$ correspond à l'effectif cumulé des valeurs de x inférieures ou égales à a_h (extrémité supérieure de la classe $]a_{h-1}, a_h]$).

• La *fréquence cumulée* des valeurs de $\boldsymbol{x}$ inférieures ou égales à a_h étant notée π_h :

$$\pi_h = \sum_{i=1}^{h} f_i = f_1 + f_2 + \ldots + f_h = \nu_h / n\,.$$

Exprimée en %, elle désigne le pourcentage d'observations qui prennent une valeur au plus égale à a_h. La fréquence cumulée des valeurs de $\boldsymbol{x}$ inférieures ou égales à a_k est égale à 1, soit $\pi_K = 1$.

Fonction de répartition. La fonction $F(t)$ qui associe à tout t la fréquence relative des observations inférieures ou égales à t est nommée fonction de répartition.

À l'intérieur de chaque classe, la répartition des valeurs de $\boldsymbol{x}$ étant *inconnue* il est d'usage de déterminer la valeur de la fonction de répartition $F(t)$ pour chaque extrémité a_i de classe puis de faire une interpolation linéaire. Autrement dit, on détermine la valeur de la fonction de répartition aux extrémités de classe :

$$F(a_0)=0,\ F(a_1)=\pi_1=f_1,\ F(a_2)=\pi_2=f_1+f_2,\ F(a_3)=\pi_3=f_1+f_2+f_3,\ldots,$$

$$F(a_h)=\pi_h=\sum_{i=1}^{h} f_i = f_1+f_2+\ldots+f_h,\ldots,F(a_K)=\pi_K=1$$

puis on obtient l'interpolation de $F(t)$ notée $F^*(t)$ en procédant comme suit :

$F^*(t) = 0$ pour $t \leqslant a_0$,

$F^*(t) = F(a_{h-1}) + f_h(t-a_{h-1})/(a_h-a_{h-1})$ pour $a_{h-1} < t \leqslant a_h\ \forall\, h = 1,2,..K$,

ainsi

$F^*(t) = F(a_0) + f_1(t-a_0)/(a_1-a_0)$ pour $a_o < t \leqslant a_1$

$F^*(t) = 1$ pour $t \geqslant a_K$

Sa représentation graphique s'obtient en joignant successivement les points P_h de coordonnées $(a_h, F(a_h))$ par des segments de droites (fig. 1.9).

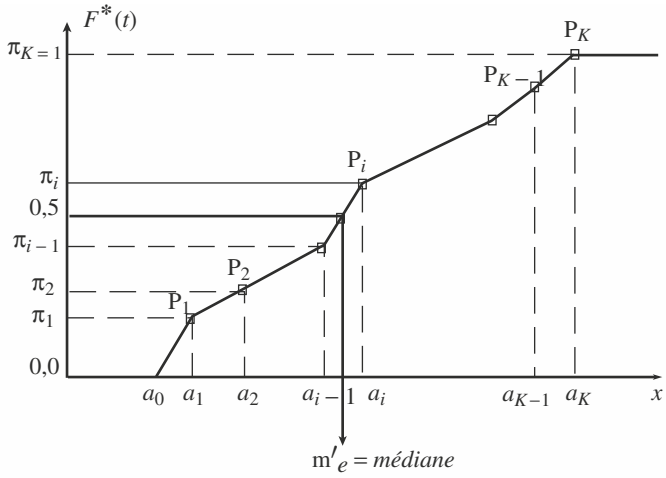


Figure 1.9 – *Fonction de répartition*

© Dunod. La photocopie non autorisée est un délit.

Exemple

Ci-dessous le tableau récapitulatif de la distance x parcourue par 5 000 véhicules :

Valeurs de x	$0 \leqslant x \leqslant 5$	$5 < x \leqslant 10$	$10 < x \leqslant 20$	$20 < x \leqslant 30$	$30 < x \leqslant 40$
Effectifs n_i	500	2 000	1 500	750	250
Fréquences relatives f_i	0,10	0,40	0,30	0,15	0,05
Fréquences cumulées $F(a_h)$	0,10	0,50	0,80	0,95	1,00

On déduit les coordonnées des points P_i : $P_0 = (0, 0)$; $P_1 = (5, 0,1)$; $P_2 = (10, 0,50)$; $P_3 = (20, 0,80)$; $P_4 = (30, 0,95)$; $P_5 = (40,1)$, puis le graphe de la fonction de répartition F^*. Voir figure 1.10.

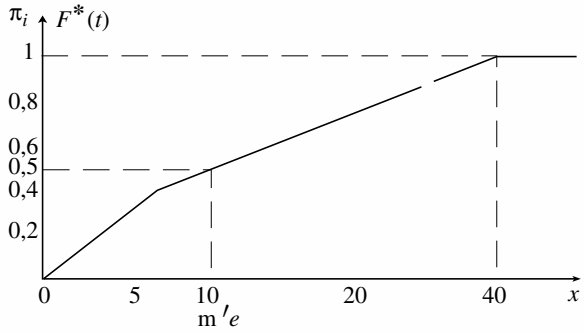


Figure 1.10 – *Fonction de répartition*

Remarquons que pour $t \leqslant 0$, $F^*(t) = 0$,
pour $0 < t \leqslant 5$, $F^*(t) = F^*(0) + 0,1 \times (t - 0)/(5 - 0) = 0 + 0,1 \times t/5$,
pour $5 < t \leqslant 10$ $F^*(t) = F(5) + 0,4 \times (t - 5)/(10 - 5) = 0,1 + 0,08 \times (t - 5)$;
pour $10 < t \leqslant 20$ $F^*(t) = F(10) + 0,3 \times (t - 10)/(20 - 10) = 0,5 + 0,03 \times (t - 10)$ ...

3 Valeurs caractéristiques de la répartition

- Si la série statistique de la variable continue comporte n observations $x_1, x_2, \ldots, x_n$ connues, distinctes ou non (*bien que théoriquement distinctes*) et classées par ordre de valeurs croissantes ($x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$) on détermine la médiane, les quartiles, déciles, etc. comme au § 3.1 de la section 2 (*cf.* p. 9).
- Dans le cas où les n observations d'une variable continue sont regroupées en K classes $]a_0, a_1],]a_1, a_2], \ldots,]a_{K-1}, a_K]$ on a *une perte d'informations*.

3.1 Valeur médiane

Il est d'usage d'attribuer à la valeur médiane de la série statistique sa valeur estimée m'_e définie par $F^*(m'_e) = 0,5$. Autrement dit, la médiane correspond graphiquement à l'abscisse du point de la fonction de répartition qui a pour ordonnée 0,5 (*cf.* fig.1. 10).

La *classe médiane* est par définition la classe $]a_{i-1},a_i]$ telle que :

– la fréquence relative cumulée jusqu'à l'extrémité supérieure de la classe est supérieure ou égale à 0,5 soit $F(a_i) \geqslant 0{,}5$ et que

– la fréquence relative cumulée jusqu'à l'extrémité inférieure de la classe est inférieure à 0,5 soit $F(a_{i-1}) < 0{,}5$.

Par interpolation linéaire on a :

$$m'_e = a_{i-1} + (a_i - a_{i-1}) \times (0{,}5 - F(a_{i-1}))/(F(a_i) - F(a_{i-1})) .$$

3.2 Quartiles

• *Premier quartile*. Sa valeur estimée q'_1 est définie par $F^*(q'_1) = 0{,}25$. La classe qui contient le premier quartile est la classe $]a_{i-1},a_i]$ telle que $F(a_i) \geqslant 0{,}25$ et $F(a_{i-1}) < 0{,}25$. Par interpolation linéaire on constate que :

$$q'_1 = a_{i-1} + (a_i - a_{i-1}) \times (0{,}25 - F(a_{i-1}))/(F(a_i) - F(a_{i-1})) .$$

• *Troisième quartile*. Il est d'usage d'attribuer au troisième quartile la valeur estimée q'_3 définie par $F^*(q'_3) = 0{,}75$. La classe du troisième quartile est la classe $]a_{i-1},a_i]$ telle que $F(a_i) \geqslant 0{,}75$ et $F(a_{i-1}) < 0{,}75$. Par interpolation linéaire on a :

$$q'_3 = a_{i-1} + (a_i - a_{i-1}) \times (0{,}75 - F(a_{i-1}))/(F(a_i) - F(a_{i-1})) .$$

De même on attribue à chaque décile $d_h (h = 1,2,\ldots,9)$ sa valeur estimée d'_h définie par $F^*(d'_h) = h/10$.

Exemple

On s'intéresse à la durée de service d'un guichet qui sert un client à la fois. On a relevé la durée de service de mille clients consécutifs ($n = 1\,000$), l'unité de temps étant la seconde. Les résultats sont consignés ci-dessous.

Répartition des valeurs de x	$0 \leqslant x \leqslant 30$	$30 < x \leqslant 60$	$60 < x \leqslant 90$	$90 < x \leqslant 150$	$150 < x \leqslant 240$
Effectifs n_i	369	251	148	163	69
Fréquences relatives f_i	0,369	0,251	0,148	0,163	0,069
Fréq. cum. $F(a_i) = \pi_i$	0,369	0,62	0,768	0,931	1

– La classe médiane est la classe $]a_{i-1},a_i]$ qui doit satisfaire aux conditions « $F(a_i) \geqslant 0{,}5$ et $F(a_{i-1}) < 0{,}5$ ».

$F(60) = 0{,}62 \geqslant 0{,}5$ et $F(30) = 0{,}369 < 0{,}5$ donc]30,60] est la classe médiane et $m'_e = 30 + (60 - 30) \times (0{,}5 - 0{,}369)/(0{,}62 - 0{,}369) = 45{,}657$.

Environ 50 % des observations ont une valeur inférieure à 45,657.

© Dunod. La photocopie non autorisée est un délit.

– La classe du troisième quartile est la classe $]a_{i-1},a_i]$ telle que $F(a_i) \geqslant 0{,}75$ et $F(a_{i-1}) < 0{,}75$ soit la classe]60,90].
Donc $q'_3 = 60 + (90 - 60) \times (0{,}75 - 0{,}62)/\ (0{,}768 - 0{,}62) = 86{,}35$.
Il y a donc environ 75 % des observations qui ont une valeur inférieure ou égale à 86,35.

Section 4

INDICATEURS DE TENDANCE CENTRALE

Ce sont des indicateurs qui se proposent de synthétiser l'ensemble d'une série statistique en faisant ressortir une position centrale de la valeur du caractère étudié. On utilise généralement la valeur médiane (*cf.* § précédent), différents types de moyennes et la valeur modale.

REPÈRES

Le statisticien Yule a énoncé les propriétés souhaitables pour les indicateurs d'une série statistique (indicateurs de tendance centrale, de dispersion…). Un bon indicateur doit :

– être défini de façon objective et indépendante de l'observateur ;

– dépendre de toutes les observations ;

– être de signification concrète ;

– être simple à calculer ;

– être peu sensible aux fluctuations d'échantillonnage ;

– se prêter aisément au calcul algébrique.

1 Moyenne arithmétique $\overline{x}$

La valeur moyenne d'une série statistique est la somme de toutes les observations divisée par le nombre total d'observations. Autrement dit, la *valeur moyenne* $\overline{x}$ d'une série de n observations $x_1,x_2,\ldots,x_n$ est $\overline{x} = (x_1 + x_2 + \ldots + x_n)/n$.

• Lorsque les n observations d'une variable discrète sont regroupées, sans perte d'information, en K classes $\{x_1^*\},\{x_2^*\},\ldots,\{x_K^*\}$ d'effectifs respectifs $n_1,n_2,\ldots,n_K$, on vérifie aisément que :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{K} n_i x_i^*.$$

La valeur moyenne est la moyenne des valeurs $\{x_1^*\},\{x_2^*\},\ldots,\{x_K^*\}$ pondérées par leurs effectifs respectifs $n_1,n_2,\ldots,n_K$.

• Si les n observations d'une variable continue sont regroupées en K classes $]a_0,a_1],]a_1,a_2],\ldots,]a_{K-1},a_K[$ d'effectifs respectifs $n_1,n_2,\ldots,n_K$ on a une perte d'informations. Il est d'usage d'attribuer à la valeur moyenne $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ sa valeur estimée $\overline{x}'$ définie par $\overline{x}' = \frac{1}{n}\sum_{i=1}^{K} n_i x_i^*$ où x_i^* est le centre de la i-ème classe : $x_i^* = (a_{i-1} + a_i)/2$.

REPÈRES : Propriétés de la moyenne $\overline{x}$

La moyenne a l'avantage d'être unique et de prendre en compte l'ensemble des observations. Elle présente également des propriétés algébriques intéressantes :

i) La somme des écarts des valeurs de x à la moyenne $\overline{x}$ est nulle : $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

ii) La somme $\sum_{i=1}^{n}(x_i - a)^2$ est minimale lorsque $a = \overline{x}$.

iii) Si l'on considère la nouvelle variable statistique $y = \alpha x + \beta$ où α et β sont des réels donnés alors on a : $\overline{y} = \alpha\overline{x} + \beta$.

Elle peut en revanche mal décrire la tendance centrale lorsqu'il y a des valeurs extrêmes exceptionnelles (valeurs exceptionnellement petites ou grandes).

2 Autres moyennes usuelles

• *Moyenne harmonique H*. Elle est l'inverse de la moyenne arithmétique des inverses des valeurs x_i :

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} = \frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right).$$

Dans le cas où les n observations sont regroupées en K classes $\{x_1^*\},\{x_2^*\},\ldots,\{x_K^*\}$ d'effectifs respectifs $n_1,n_2,\ldots,n_K$ on a

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{K}\frac{n_i}{x_i^*} = \frac{1}{n}\left(\frac{n_1}{x_1^*} + \frac{n_2}{x_2^*} + \ldots + \frac{n_K}{x_K^*}\right).$$

Sa définition présuppose que $\boldsymbol{x}$ ne prend pas la valeur 0. Elle est utile tout particulièrement pour les taux de change, les taux d'équipement ...

© Dunod. La photocopie non autorisée est un délit.

- *Moyenne géométrique G*. La moyenne géométrique est la racine n-ème du produit des n valeurs x_i supposées positives. La moyenne géométrique de K valeurs positives $\{x_1^*,x_2^*,\ldots,x_K^*\}$ munies des pondérations $n_1,n_2,\ldots,n_K$ est définie par [1]

$$G = \left[(x_1^*)^{n_1} \times (x_2^*)^{n_2} \times \ldots \times (x_K^*)^{n_K}\right]^{1/n}.$$

  Cet indicateur est utilisé pour évaluer les changements de prix, de salaire ...
- *Moyenne quadratique Q*. Q est la racine carrée de la moyenne arithmétique des carrés des valeurs x_i^* :

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{K} n_i (x_i^*)^2} = \sqrt{\frac{1}{n}(n_1(x_1^*)^2 + n_2(x_2^*)^2 + \ldots + n_K(x_K^*)^2)}$$

  Quelle que soit la série statistique considérée où tous les x_i^* sont positifs on a $H \leqslant G \leqslant \overline{x} \leqslant Q$.

Exemple

Partant des 12 relevés mensuels consécutifs de vente de l'article référencé VB dans une grande surface on souhaite déterminer successivement les moyennes mensuelle arithmétique, géométrique, quadratique et harmonique. La série des ventes mensuelles est récapitulée ci-dessous :

Quantités vendues x_i^*	1	2	3	4	5
Nombre de mois n_i	2	3	4	2	1

Lire : il y a deux mois dans l'année durant lesquels la grande surface vend un seul VB, trois mois durant lesquels elle vend deux VB, etc. Le nombre d'observations $n = 12$.

Moyenne arithmétique :

$\overline{x} = (1/12) \times (2 \times 1 + 3 \times 2 + 4 \times 3 + 2 \times 4 + 1 \times 5) = 2{,}75$.

Moyenne géométrique : $G = (1^2 \times 2^3 \times 3^4 \times 4^2 \times 5^1)^{1/12} = 2{,}47$.

Moyenne quadratique :

$Q = [(1/12) \times (2 \times 1^2 + 3 \times 2^2 + 4 \times 3^2 + 2 \times 4^2 + 1 \times 5^2)]^{1/2} = 2{,}986$.

Moyenne harmonique H définie par

$1/H = (1/12) \times (2/1 + 3/2 + 4/3 + 2/4 + 1/5) = 0{,}461$ soit $H = 2{,}16$.

3 Le mode

Le mode est une mesure de tendance centrale non influencée par les valeurs extrêmes de la distribution. Il est particulièrement employé pour des données nominales.

1 Le logarithme de la moyenne géométrique est égal à la moyenne arithmétique des logarithmes des valeurs x_i^*.

Variable discrète. Dans le cas d'une variable discrète, le mode est la valeur des observations qui apparaît le plus fréquemment.

Dans le cas où les n observations sont regroupées en K classes $\{x_1^*\},\{x_2^*\},\ldots,$ $\{x_K^*\}$ d'effectifs respectifs $n_1,n_2,\ldots,n_K$, le mode m_0 est la valeur $x_{i_0}^*$ de la variable statistique pour laquelle l'effectif n_{i_0} est maximal : $m_0 = x_{i_0}^*$.

Variable continue distribuée en classes. La *classe modale* est celle qui, dans l'histogramme des fréquences relatives, a la plus grande ordonnée : l'intervalle $]a_{i_0-1},a_{i_0}]$ est la classe modale si et seulement si $y'_{i_0} = f_{i_0}/(a_{i_0} - a_{i_0-1}) > y'_i = f_i/(a_i - a_{i-1})\ \forall\, i \neq i_0$.

Le milieu de la classe modale $m_0 = (a_{i_0} + a_{i_0-1})/2$ est par convention la valeur modale. Dans la plupart des cas concrets, il y a une seule valeur modale, la distribution est dite alors unimodale.

Si la distribution sur une population comporte deux ou plusieurs modes (distribution bimodale ou multimodale) cela laisse à penser qu'il existe deux ou plusieurs groupes distincts dans la population.

Exemple d'une série discrète

La valeur modale de la répartition de l'exemple précédent p. 22 est égale à 3 car pour cette valeur, l'effectif est maximal.

Exemple d'une série continue

Reprenons l'exemple de la société LOCAUTO (p. 14)

La classe]5,10] est la classe modale puisque la hauteur $y'_2 = 0{,}08$ de l'histogramme des fréquences relatives est la plus élevée des cinq hauteurs y'_i. Le mode est le centre de la classe : $x_{m_o} = 7{,}5$.

Section **5**

INDICATEURS DE DISPERSION

Exemple

On dispose des relevés hebdomadaires des ventes du lave-linge référencé MAMXRT dans deux grandes surfaces A et B sur deux années. Les séries étudiées comportent donc un total de 100 relevés compte tenu des périodes de fermeture égales à deux semaines par an.

Les deux séries statistiques ci-dessous ont même valeur moyenne égale à 2.

	Série A			Série B		
Quantités vendues	1	2	3	1	2	3
Effectif ou nombre de semaines	30	40	30	5	90	5

© Dunod. La photocopie non autorisée est un délit.

Il semble que la série B soit « plus centrée » que la série A sur sa valeur moyenne et donc « moins dispersée ». Aussi est-il nécessaire d'utiliser des unités de mesure de la dispersion. Toutes les caractéristiques de dispersion reposent sur une analyse d'intervalles ou d'écarts par rapport à une origine convenablement choisie.

1 La dispersion en termes d'écarts moyens

L'origine retenue pour le calcul des écarts moyens est généralement une caractéristique de tendance centrale telle la valeur médiane m_e ou la valeur moyenne $\overline{x}$. Pour que la moyenne des écarts par rapport à la valeur centrale retenue traduise correctement la dispersion, on considère leur valeur absolue ou leur puissance paire.

1.1 Écarts absolus moyens [1]

• *Écart absolu moyen par rapport à la valeur médiane* m_e :

$$e_{m_e} = \frac{1}{n}\sum_{i=1}^{n}|x_i - m_e| = \frac{1}{n}\left(|x_1 - m_e| + |x_2 - m_e| + \ldots + |x_n - m_e|\right)$$

où n est le nombre d'observations, x_i est la valeur de la i-ème observation, m_e est la médiane de la série statistique.

Dans le cas de regroupement des n valeurs x_i en K classes sans perte d'information

$$e_{m_e} = \frac{1}{n}\sum_{i=1}^{K} n_i|x_i^* - m_e| = \frac{1}{n}\left(n_1|x_1^* - m_e| + n_2|x_2^* - m_e| + \ldots + n_K|x_K^* - m_e|\right)$$

• *Écart absolu moyen par rapport à la moyenne arithmétique* $\overline{x}$:

$$e_{\overline{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| = \frac{1}{n}\sum_{i=1}^{K} n_i|x_i^* - \overline{x}|$$

dans le cas du regroupement des valeurs x_i en K classes sans perte d'information.

Exemple

Reprenant l'exemple de la série A on a

$$e_{\overline{x}} = \frac{1}{n}\sum_{i=1}^{K} n_i|x_i^* - \overline{x}| = (1/100)\times(30\times|1-2| + 40\times|2-2| + 30\times|2-1|) = 0{,}6$$

et dans le cas de la série B $e_{\overline{x}} = 0{,}1$, cela prouve que la série B est moins dispersée.

---

1 **Propriété.** L'écart absolu moyen par rapport à c, $e_c = \frac{1}{n}\sum_{i=1}^{n}|x_i - c|$ est minimal lorsque $c = m_e$.

1.2 Variance et écart-type

Ils mesurent la dispersion ou l'étalement des données autour de la moyenne.

Définitions. La variance $V(x)$ est la moyenne arithmétique des carrés des écarts des valeurs de $\boldsymbol{x}$ à leur moyenne arithmétique :

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\left[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\right].$$

- Si les n observations d'une variable discrète sont regroupées, sans perte d'information, en K classes $\{x_1^*\},\{x_2^*\},\ldots,\{x_K^*\}$ d'effectifs respectifs $n_1,n_2,\ldots,n_K$, alors

$$V(x) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i^* - \overline{x})^2 = \frac{1}{n}\left[n_1(x_1^* - \overline{x})^2 + n_2(x_2^* - \overline{x})^2 + \ldots + n_K(x_K^* - \overline{x})^2\right]$$

- Lorsque les n observations d'une variable continue sont regroupées en K classes $]a_0,a_1],]a_1,a_2],\ldots,]a_{K-1},a_K[$ d'effectifs respectifs $n_1,n_2,\ldots,n_K$, on a une perte d'information. Il est d'usage d'attribuer à $V(x)$ sa valeur estimée $\frac{1}{n}\sum_{i=1}^{K} n_i(x_i^* - \overline{x}')^2$ où $x_i^* = \frac{(a_{i-1} + a_i)}{2}$ et $\overline{x}' = \frac{1}{n}\sum_{i=1}^{K} n_i x_i^*$.

L'écart-type ou écart quadratique σ_x est la racine carrée de la variance : $\sigma_x = \sqrt{V(x)}$

REPÈRES : Propriétés de la variance

Elle est simple à calculer, tient compte de l'ensemble des observations et présente certaines propriétés intéressantes.

i) Formule de Köenig :

$$V(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{n}\left[(x_1)^2 + (x_2)^2 + \cdots + (x_n)^2\right] - \overline{x}^2$$

ou dans le cas du regroupement des n valeurs x_i en K classes :

$$V(x) = \frac{1}{n}\sum_{i=1}^{K} n_i(x_i^*)^2 - \overline{x}^2 = \frac{1}{n}\left[n_1(x_1^*)^2 + n_2(x_2^*)^2 + \cdots + n_K(x_K^*)^2\right] - \overline{x}^2$$

ii) La multiplication de chaque valeur de x par une même constante a multiplie la variance par le carré de cette constante : $V(ax) = a^2 V(x)$.

iii) L'addition d'une même constante b à chaque valeur de $\boldsymbol{x}$ ne modifie pas la variance :

$$V(x + b) = V(x).$$

De *ii)* et *iii)* il résulte que $V(ax + b) = a^2 V(x)$.

iv) $\forall\, t > 1$, on a « effectif des valeurs comprises entre $]\overline{x} - t\sigma_x[$ et $]\overline{x} + t\sigma_x[$ $\geqslant n(1 - 1/t^2)$ » (***inégalité de Bienaymé-Tchébycheff***)

v) $V_a = \frac{1}{n}\sum_{i=1}^{m}(x_i - a)^2$ est minimal lorsque $a = \overline{x}$.

© Dunod. La photocopie non autorisée est un délit.

Exemple

Partant des relevés mensuels des ventes de l'article référencé MAMXRT dans une grande surface C au cours d'une année, on souhaite calculer la variance de la série.

Quantités vendues	1	2	3	4	5
Nombre de mois	2	3	3	2	2

Moyenne arithmétique : $\overline{x} = (1/12) \times (2 \times 1 + 3 \times 2 + \ldots + 2 \times 5) = 2{,}916$.

Variance : $V(x) = (1/12) \times (2 \times 1^2 + 3 \times 2^2 + \ldots + 2 \times 5^2) - 2{,}916^2 = 1{,}743$.

L'écart-type $\sigma_x = \sqrt{V(x)} = 1{,}32$.

1.3 Moment centré d'ordre *h*

Le moment centré d'ordre h est la moyenne arithmétique des écarts des valeurs de $\boldsymbol{x}$ à leur moyenne arithmétique élevés à la puissance h :

$$\mu_{x,h} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^h = \frac{1}{n}\left[(x_1 - \overline{x})^h + (x_2 - \overline{x})^h + \ldots + (x_n - \overline{x})^h\right]$$

Lorsque les n observations d'une variable discrète sont regroupées en K classes $\{x_1^*\},\{x_2^*\},\ldots,\{x_K^*\}$ d'effectifs respectifs $n_1, n_2, \ldots, n_K$, on a :

$$\mu_{x,h} = \frac{1}{n}\sum_{i=1}^{K} n_i (x_i^* - \overline{x})^h = \frac{1}{n}\left[n_1(x_1^* - \overline{x})^h + n_2(x_2^* - \overline{x})^h + \ldots + n_K(x_K^* - \overline{x})^h\right]$$

Remarquer que la variance est le moment centré d'ordre 2.

Exemple

Dans l'exemple précédent le moment centré d'ordre 4

$$\mu_{x,4} = (1/12) \times (2 \times (1 - 2{,}916)^4 + 3 \times (2 - 2{,}916)^4 + 3 \times (3 - 2{,}916)^4 + 2 \times (4 - 2{,}916)^4 + 2 \times (5 - 2{,}916)^4) = 5{,}794$$

2 La dispersion en termes d'intervalles

L'intervalle de variation ou étendue est la différence entre les valeurs extrêmes de la variable (soit la différence entre la plus grande et la plus petite valeur) ou entre les centres des classes extrêmes. Il a pour principal défaut de ne tenir compte que des valeurs extrêmes qui sont parfois accidentelles ou exceptionnelles.

L'intervalle interquartile est l'intervalle (q_1, q_3) qui contient approximativement 50 % des observations, le reste se répartissant pour 25 % « à gauche » de q_1 et pour 25 % « à droite » de q_3 . L'indicateur retenu est la différence $(q_3 - q_1)$.

De même, l'intervalle interdécile (d_1, d_9) contient approximativement 80 % des observations.

Section 6

INDICATEURS DE FORME

Les indicateurs précédemment définis dépendent des unités de mesures utilisées et ne permettent donc pas de comparer la dispersion de séries hétérogènes. Pour éviter cet inconvénient on utilise des indicateurs de forme.

1 Coefficients d'asymétrie et d'aplatissement

Le coefficient d'asymétrie de Fisher (ou *skewness*) : $\gamma = \mu_{x,3}/(\sigma_x^3)$

où $\mu_{x,3} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$ est le moment centré d'ordre 3 et σ_x^3 est l'écart-type élevé au cube. Il est nul si la distribution des fréquences relatives est symétrique autour de la moyenne et est positif si la distribution est plus étalée à droite qu'à gauche de $\overline{x}$.

Le coefficient d'aplatissement Γ de Fisher (ou *Kurtosis*) : $\Gamma = \mu_{x,4}/(\sigma_x^4) - 3$

où $\mu_{x,4} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$ est le moment centré d'ordre 4 et σ_x^4 est l'écart-type élevé à la puissance 4.

Remarque. Dans le cas d'une distribution symétrique : moyenne = médiane.

2 Coefficients de dispersion relative de la distribution

- *coefficient interquartile* $= \dfrac{q_3 - q_1}{q_3 + q_1}$
- *coefficient de variation* $= \dfrac{\sigma_x}{\overline{x}}$. Cet indicateur est utile pour comparer la dispersion d'un caractère sur des populations ayant des valeurs moyennes très différentes ou exprimées dans des unités de mesure différente (monnaies différentes . . .).

© Dunod. La photocopie non autorisée est un délit.

Section 7

REPRÉSENTATIONS GRAPHIQUES

Les diagrammes en barres et circulaires sont utilisés pour donner une représentation visuelle des différentes parties d'un ensemble, ils sont notamment utiles pour caractériser la distribution de caractères qualitatifs (couleur d'une voiture . . .).

1 Diagramme circulaire

Chaque partie est représentée par un secteur dont la surface est proportionnelle à sa « mesure » (généralement les effectifs).

2 Diagramme en barres

Chaque partie est représentée par un rectangle dont l'aire est proportionnelle à sa « mesure ».

Exemple

En l'an N, la consommation des ménages d'une région donnée est de 2 758 unités monétaires et se répartit ainsi (en unités monétaires) : produits alimentaires 559 ; énergie 264 ; produits industriels 798 ; services 1 137, soit en pourcentages, respectivement 20,3 % ; 9,6 % ; 28,9 % et 41,2 %.

– Dans le cas d'une représentation circulaire, l'on porte sur le disque successivement (voir ci-après) : *angle* (OA,OB) $= 0{,}203 \times 360^\circ = 73^\circ 08$; *angle* (*OB*, OC) $= 0{,}0957 \times 360^\circ = 34^\circ 56$; *angle* (OC, OD) $= 0{,}289 \times 360^\circ = 104^\circ 04$ et l'on a nécessairement *angle* (OD, OA) $= 148^\circ 32$.

  La surface du secteur (AOB) repréente la part des produits alimentaires dans la consommation totale, elle-même représentée par la surface du disque : surface du secteur (AOB)/ surface du disque = 559/2 758 = 20,3 %. Il en est de même pour chacun des trois autres secteurs.

– Dans le cas de diagramme en barre, la hauteur de la barre est proportionnelle au nombre d'unités monétaires que représentent les différents secteurs.

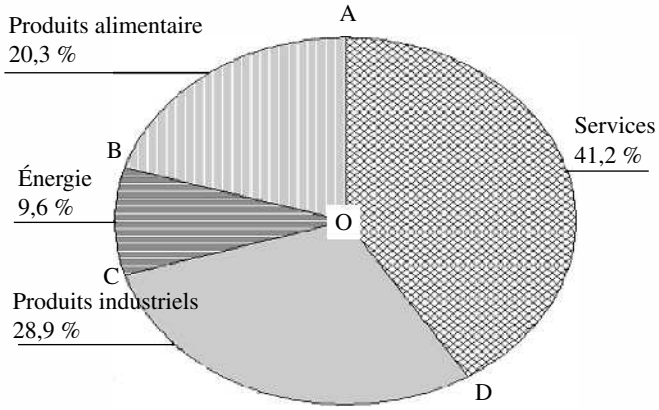


Figure 1.11 – *Diagramme circulaire*

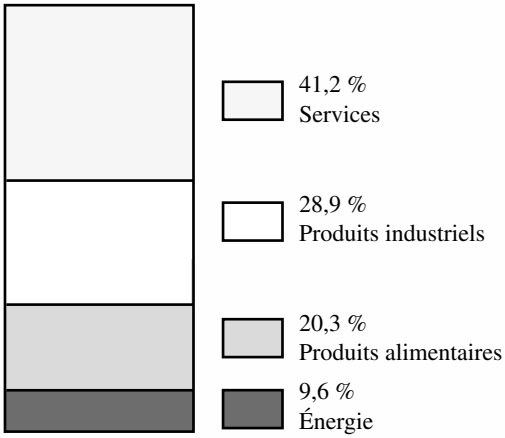


Figure 1.12 – *Diagramme en barres*

Section **8**

INDICATEURS DE CONCENTRATION

Ils mesurent les inégalités dans les répartitions d'une grandeur cumulative $\boldsymbol{x}$ (répartition des revenus, du patrimoine . . .).

Exemple

Les primes x_h attribuées à un service comprenant 25 employés se répartissent ainsi :

Primes x_h (en 10^2 euros)	8	30	40	60	100
Effectifs	9	7	4	3	2

© Dunod. La photocopie non autorisée est un délit.

Aussi peut-on s'intéresser au caractère plus ou moins égalitaire de la répartition des primes au sein du service. Pour cela il convient de déterminer, pour $h = 1,2,3,4,5$, quelle est la part $\rho(x_h^*)$ du total des primes M qui a été octroyé au π_h pourcentage d'employés qui ont perçu au plus x_h^*.

x_i^*	8	30	40	60	100	Total
Effectifs	9	7	4	3	2	25
Effectifs cumulés ν_h	9	16	20	23	25	
Fréquences relatives cumulées $F(x_h^*) = \pi_h$ **ou proportion d'individus ayant une prime** $\leqslant$ **à** x_h^*	0,36 (=9/25)	0,64 (=16/25)	0,80	0,92	1	
Valeur globale des primes distribuées $n_i \times x_i^*$	72	210	160	180	200	M =822
Valeur globale cumulée des primes distribuées à ceux ayant une prime inférieure ou égale à x_h^* $M_h = \sum_{i=1}^{h} n_i x_i^*$	72	282	442	622	822	
$\rho(x_h^*) = M_h/M$ **part du montant total des primes attribuées à ceux dont la prime est** $\leqslant$ **à** x_h^*	0,087 (=72/822)	0,343 (=282/822)	0,538 (=442/822)	0,757	1	

Grandeur cumulée. D'une façon générale, la série statistique comprenant n individus, on associe à chaque individu e_i l'expression numérique x_i d'un caractère et ordonnant les valeurs $x_1,x_2,\ldots,x_n$ par ordre croissant $(x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n)$ on considère la somme des x_i ou grandeur cumulée $M = \sum_{i=1}^{n} x_i = n \times \overline{x}$.

Lorsque les n observations de $\boldsymbol{x}$ sont regroupées en K classes $\{x_1^*\},\{x_2^*\},\ldots,\{x_K^*\}$ (où $x_1^* < x_2^* < \ldots < x_K^*$) d'effectifs respectifs $n_1,n_2,\ldots,n_K$ alors $M = n \times \overline{x} = \sum_{i=1}^{K} n_i x_i^*$.

S'intéressant à la façon dont se répartit le total M entre les n individus, il convient de déterminer les couples $(F(x_h^*),\rho(x_h^*)) = (\pi_h, \sum_{i=1}^{h} n_i x_i^*/M)$ pour $h = 1,2,\ldots,K$ où

$$F(x_h^*) = \pi_h = f_1 + f_2 + \ldots + f_{h-1} + f_h$$

est la fréquence relative cumulée ou proportion d'observations dont la valeur prise est inférieure ou égale x_h^* et

$$\rho(x_h^*) = \sum_{i=1}^{h} n_i x_i^*/M = \frac{n_1 x_1^* + n_2 x_2^* + \ldots + n_h x_h^*}{n_1 x_1^* + n_2 x_2^* + \ldots + n_h x_h^* + \ldots + n_K x_K^*}$$

est la part de M que représentent les observations qui prennent une valeur inférieure ou égale à x_h^*.

– S'il s'agit d'une variable continue dont les valeurs sont regroupées en K classes $]a_0,a_1],]a_1,a_2],\ldots,]a_{K-1},a_K[$ d'effectifs respectifs $n_1,n_2,\ldots,n_K$ on estime la valeur de M par $M \cong \sum_{i=1}^{K} n_i x_i^*$ et $\rho(a_h)$ par $\rho(a_h) \cong \left(\sum_{i=1}^{h} n_i x_i^*\right)/M$ où x_i^* est

le centre de la i-ème classe. On détermine les couples $(F(a_h^*), \rho(a_h)) = (\sum_{i=1}^{h} f_i, \sum_{i=1}^{h} n_i x_i^*/M)$.

1 Courbe de Lorentz

D'une façon générale, pour toute valeur t vérifiant $0 \leqslant t \leqslant x_K^*$, on s'intéresse à la proportion $\rho(t)$ de la masse $M = \sum_{i=1}^{K} n_i x_i^*$ détenue par la proportion $F(t)$ des éléments e_i de P pour lesquels $\boldsymbol{x}(e_i) \leqslant t$. Lorsque $t = x_h^*$ on retrouve $F(x_h^*) = \pi_h$ et $\rho(x_h^*) = \left(\sum_{i=1}^{h} n_i x_i^*\right)/M$.

Choisissant un repère orthonormé où figurent en abscisse $F(t)$ et en ordonnée $\rho(t)$, on construit la courbe de concentration en joignant par des segments de droites les points $M_1 = (F(x_1^*), \rho(x_1^*)),\ M_2 = (F(x_2^*), \rho(x_2^*)), \ldots,\ M_K = (F(x_K^*)\rho(x_K^*)),\ = (1,1)$

Exemple

Reprenant l'exemple précédent on fait figurer en abscisse $F(t)$ « la fréquence relative de salariés dont la prime est $\leqslant t$ » et en ordonné $\rho(t)$ « la part du total des primes perçus par les salariés dont la prime est $\leqslant t$ ». Pour construire la courbe de concentration, on joint les différents points de coordonnées $(F(x_h^*), \pi_h)$: $M_0(0,0)$, $M_1(0{,}36, 0{,}087)$, $M_2(0{,}64, 0{,}343)$, $M_3(0{,}8, 0{,}538)$, $M_4(0{,}92, 0{,}757)$, $M_5(1,1)$.

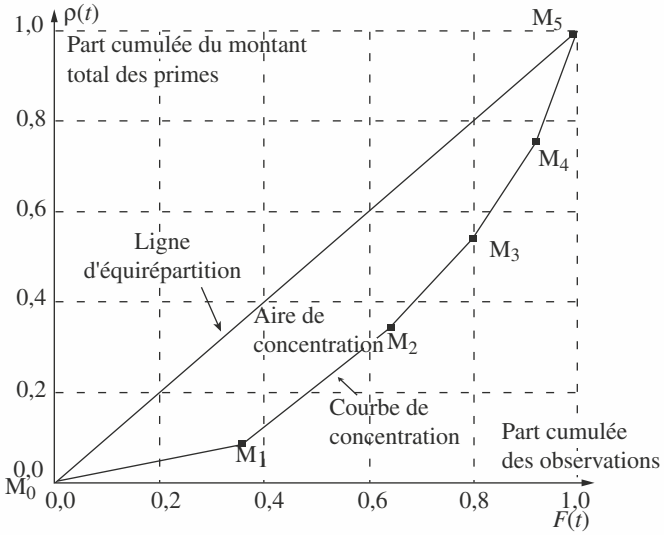


Figure 1.13 – *Courbe de concentration*

© Dunod. La photocopie non autorisée est un délit.

L'aire de concentration est la surface du domaine délimité par la courbe de concentration et la ligne d'équirépartition qui est caractérisée par le segment OM_5 où $O = (0,0)$ et $M_5 = (1,1)$.

En considérant le point M_2 on constate que 64 % des salariés reçoivent moins de 34,3 % des primes distribuées. La répartition s'écarte de l'équipartition parfaite représentée par la diagonale OM_5.

2 Médiale

Les valeurs $x_1, x_2, \ldots, x_n$ étant classées par ordre croissant $(x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n)$, la valeur médiale m_d est la valeur x_r telle que $\sum_{i=1}^{r} x_i \geqslant M/2$ et $\sum_{i=1}^{r-1} x_i < M/2$.

C'est la première valeur de la série ordonnée telle que la somme des valeurs des observations qui lui sont inférieures ou égales atteignent ou dépasse 50 % de la masse totale M.

Les n observations de la variable étant regroupées en K classes $x_1^*, x_2^*, \ldots, x_K^*$ (où $x_1^* < x_2^* < \ldots < x_k^*$) d'effectifs respectifs $n_1, n_2, \ldots, n_K$, la valeur médiale m_d est la valeur x_r^* telle que

$$\rho(x_r^*) \geqslant 0{,}5 \text{ et } \rho(x_{r-1}^*) < 0{,}5 \text{ (soit } \sum_{i=1}^{r} n_i x_i^*/M \geqslant 0{,}5 \text{ et } \sum_{i=1}^{r-1} n_i x_i^*/M < 0{,}5)$$

Si les n observations d'une variable continue sont regroupées en K classes $]a_0, a_1],]a_1, a_2], \ldots,]a_{K-1}, a_K[$ d'effectifs respectifs $n_1, n_2, \ldots, n_K$, on attribue à la valeur médiale, sa valeur estimée m'_d définie par $\rho(m'_d) = 0{,}5$ après les interpolations linéaires. La classe médiale est la classe $]a_{i-1}, a_i]$ telle que $\rho(a_i) \geqslant 0{,}5$ et $\rho(a_{i-1}) < 0{,}5$. La valeur médiale estimée

$$m'_d = a_{i-1} + (a_i - a_{i-1}) \times (0{,}5 - \rho(a_{i-1}))/(\rho(a_i) - \rho(a_{i-1}))\,.$$

Exemple

Reprenant l'exemple précédent, on a $M = 822 \times 10^2$ d'où $M/2 = 411 \times 10^2$ et l'on constate que la part prise par les primes d'au plus 40×10^2 euros est $\rho(40) = 0{,}538 \geqslant 0{,}5$ et que $\rho(30) = 0{,}343 < 0{,}5$, donc la médiale est 40. Autrement dit, 5 des 25 employés les mieux rémunérés se répartissent $1 - 0{,}538 = 46{,}2$ % du total des primes attribuées.

3 Indices de concentration

Les indices de concentration sont des indicateurs qui prennent des valeurs comprises entre 0 et 1 avec en commun la caractéristique de prendre la valeur 0 pour une parfaite homogénéité de la répartition et la valeur 1 pour une concentration totales sur un élément.

3.1 Indice de Gini

L'indice de concentration de Gini I_G est défini par

$$I_G = (\textit{Aire de concentration})/(\textit{aire du triangle sous la diagonale})$$
$$= 2\times\ (\textit{Aire de concentration})$$

Si $I_G = 0$ il y a égalité parfaite, la courbe de concentration est confondue avec la première bissectrice. Lorsque I_G augmente, la concentration s'accroît et la répartition devient plus inégalitaire.

Repères

De façon analytique $I_G = \dfrac{1}{2(n-1)M}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_j - x_i\right| = \dfrac{1}{(n-1)M}\sum_{1\leqslant i<j\leqslant n}(x_j - x_i)$.

Les n observations étant regroupées en K classes $x_1^*, x_2^*, \cdots, x_K^*$ (où $0 \leqslant x_1^* < x_2^* < \cdots < x_K^*$)

$I_G = \dfrac{1}{2(n-1)M}\sum_{i=1}^{K}\sum_{j=1}^{K} n_i n_j \left|x_j^* - x_i^*\right| = \dfrac{1}{(n-1)M}\sum_{1\leqslant i<j\leqslant K} n_i n_j (x_j^* - x_i^*)$.

3.2 Indice d'Herfindahl

L'indice de concentration d'Herfindahl I_H est défini par :

$I_H = \dfrac{1}{1-1/n}\sum_{i=1}^{n}(x_i' - 1/n)^2$ où $x_i' = x_i/M$. On a : $I_H = 1 - J_H$ où $J_H = \dfrac{1}{1-1/n}\left(1 - \sum_{i=1}^{n} x_i'^2\right)$ est l'indice de diversification.

Exemple

Reprenant l'exemple précédent, on a : $I_G = [(9\times 7\times(30-8) + 9\times 4 \times(40-8) + \ldots + 3\times 2\times(100-60)]/(24\times 822) = 0{,}428$.

Pour déterminer J_H, remarquer que $\sum(x_i')^2 = (9\times 8^2 + 7\times 30^2 + 4\times 40^2 + 3\times 60^2 + 2\times 100^2)/822^2 = 0{,}065$ d'où l'on déduit $J_H = 0{,}974$ puis $I_H = 0{,}026$.

La disparité entre les valeurs prises par I_G et I_H résulte du fait qu'ils ne satisfont pas au « critère d'objectivité ». À I_G et I_H on doit substituer $I_G^* = \phi(I_G)$ et $I_H^* = \phi(I_H)$ qui prennent des valeurs voisines (*Cf.* P. Ch. Pupion CNRS, *Revue d'Économie industrielle* n° 76, 2ème trimestre 1996, p. 115-123)

© Dunod. La photocopie non autorisée est un délit.

Section 9

ANALYSES STATISTIQUES AVEC LOGICIELS

1 Traitements avec le logiciel EXCEL

Les fonctions statistiques assurées sont :

– **la gestion des données** (la création et la mise à jour des fichiers de données), **icône accueil** ;

– **la mise en forme des informations** pour une aide à l'interprétation (édition de tableaux, des rapports et des graphiques), menus **Insertion et Données** ;

– **le traitement statistique des données** à l'aide de procédures (correspondant aux méthodes statistiques usuelles), menu **Insertion**, menu **Données** (sous menu utilitaire d'analyse).

Les principales commandes des menus figurent ci-dessous.

Accueil : pour créer un nouveau fichier excel, enregistrer un fichier, ouvrir et fermer un fichier existant et imprimer, déterminer la forme des paragraphes et de la police

Insertion : permet de réaliser des tableaux et graphiques

Mise en page : mise en page du document (sauts de page…) définir le format des cellules, des lignes, des colonnes

Formule : une bibliothèque de fonctions disponibles (financières, mathématiques…)

Données : trier les données, filtrer et réaliser certains traitements statistiques

Révision : permet d'insérer des commentaires, de chercher des synonymes

Affichage : pour organiser les fenêtres ou passer d'une fenêtre à une autre.

Exemple

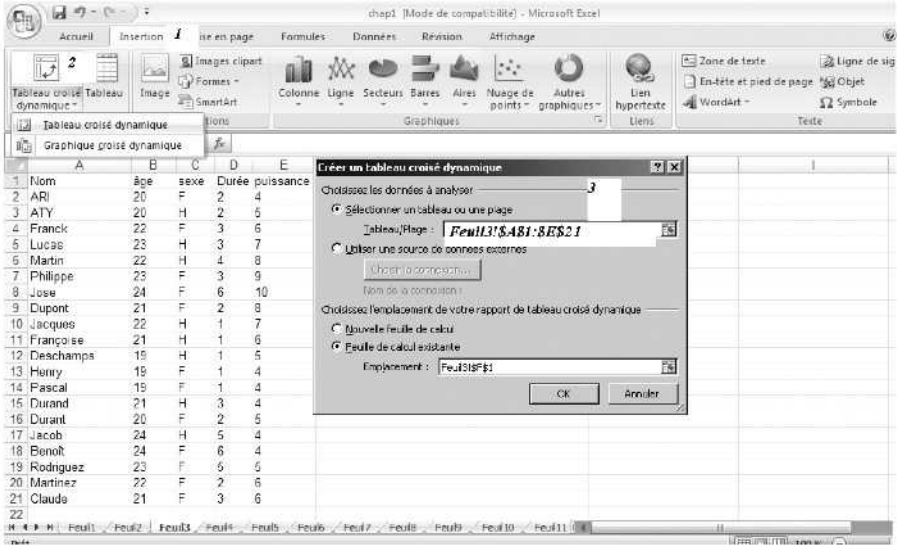
Dans un fichier d'assurance automobile figurent entre autres le nom, l'âge, le sexe du conducteur, le nombre d'années écoulées depuis le premier contrat souscrit, la puissance fiscale du véhicule. Prélevant dans ce fichier 20 conducteurs de moins de 25 ans, on indique dans la colonne A leur nom, dans la colonne B leur âge… Ainsi est obtenue la feuille Excel présentée ci-dessous.

a) Avec le logiciel Excel, il est possible de dresser le tableau correspondant à la série statistique sur le sexe des conducteurs faisant partie de l'échantillon :

Procédure.

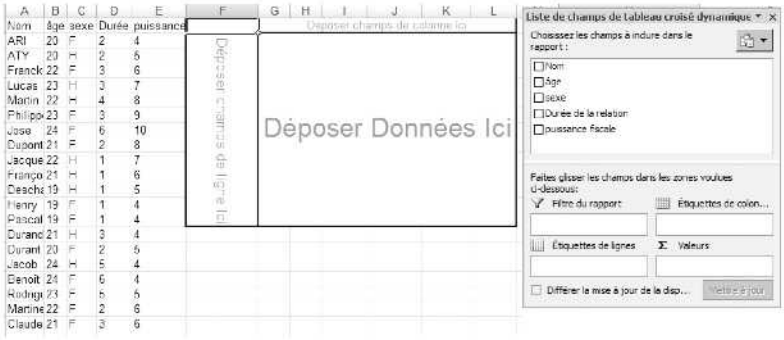
1. On clique sur (Insertion)

2. On clique sur (Tableau croisé dynamique) et on opte pour (Tableau croisé dynamique)



3. On sélectionne les cellules correspondant aux données (\$A\$1 :\$E\$21) et on clique sur (OK).

4. Apparaît la fenêtre ci-dessous



Procédure.

5. On clique sur (sexe) que l'on déplace sur (Déposer champs de ligne) puis on clique sur (Nom) que l'on déplace sur (Déposer Données ici). On obtient le tableau suivant :

F	G
Nombre de Nom	
sexe	Total
F	12
H	8
Total général	20

Il y a un caractère qualitatif le sexe Femme ou Homme et les effectifs correspondant sont 12 et 8.

© Dunod. La photocopie non autorisée est un délit.

b) On peut réaliser un tableau croisé avec en ligne le sexe et en colonne l'âge de l'assuré.

Procédure.

Pour croiser les variables âge et sexe, on fait glisser la variable (âge) sur (Déposer champs de colonne) à la cinquième étape. On obtient :

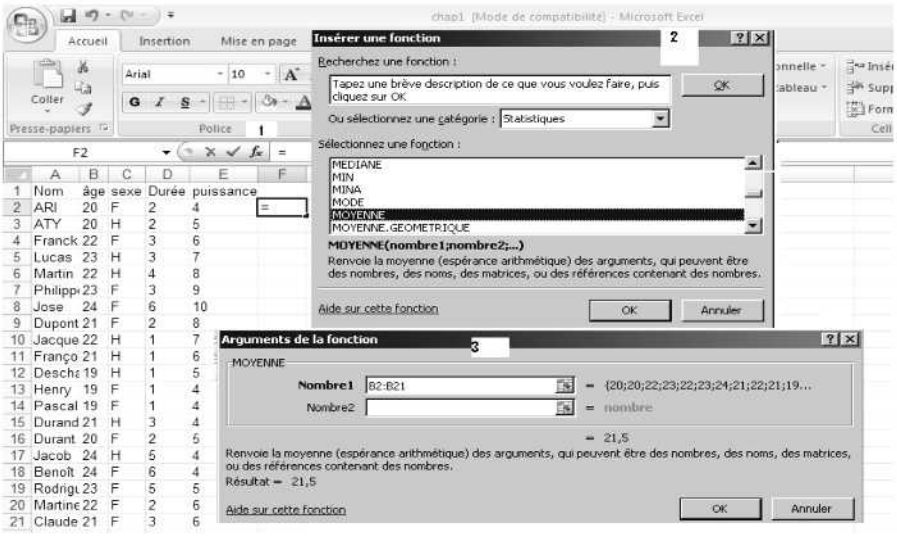
Nombre de Nom	âge						
sexe	19	20	21	22	23	24	Total général
F	2	2	2	2	2	2	12
H	1	1	2	2	1	1	8
Total général	3	3	4	4	3	3	20

c) Construire le diagramme en bâtons de la distribution par sexe des assurés.

Procédure.

Pour obtenir un diagramme de la distribution par sexe on part du premier tableau et on clique sur l'icône .

d) On peut calculer également l'âge moyen, l'âge médian, le premier et troisième quartile en partant de la feuille de données.



Procédure.

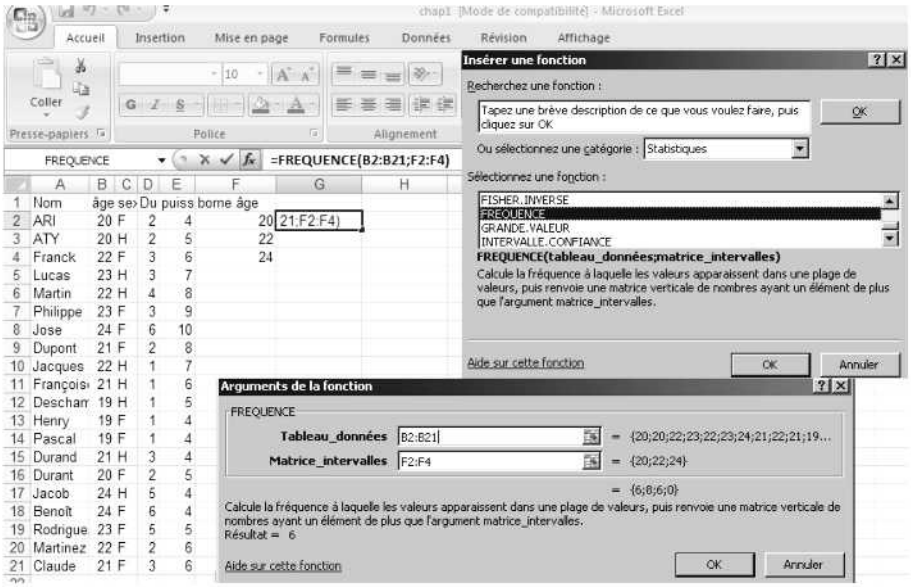
1. On clique sur (Accueil) puis (*fx*).

2. Dans le menu Insérer une fonction, on sélectionne en déroulant ▲ une (catégorie) les **Statistiques** et on déroule ▲ pour sélectionner la (fonction) souhaitée : **MOYENNE**

3. Dans le menu **Arguments de la fonction** on sélectionne la plage de données (B2 :B21) (B2 :B21) qui correspond à l'ensemble des valeurs situées entre la colonne B ligne 2 et la colonne B ligne 21 et on lit le résultat 21.5. On suit la même démarche pour les autres indicateurs.

Pour obtenir la médiane, le premier ou le troisième quartile il faut sélectionner dans l'étape 2 la fonction souhaitée

e) Regroupant les assurés par classe d'âge : [18, 20],]20,22] et]22,24] on détermine les fréquences de ces classes.



	A	B	C	D	E	F	G	H
1	Nom	âge	se›	Du	puiss	borne	âge	
2	ARI	20	F	2	4	20	21;F2:F4)	
3	ATY	20	H	2	5	22		
4	Franck	22	F	3	6	24		
5	Lucas	23	H	3	7			
6	Martin	22	H	4	8			
7	Philippe	23	F	3	9			
8	Jose	24	F	6	10			
9	Dupont	21	F	2	8			
10	Jacques	22	H	1	7			
11	François	21	H	1	6			
12	Deschan	19	H	1	5			
13	Henry	19	F	1	4			
14	Pascal	19	F	1	4			
15	Durand	21	H	3	4			
16	Durant	20	F	2	5			
17	Jacob	24	H	5	4			
18	Benoît	24	F	6	4			
19	Rodrigue	23	F	5	5			
20	Martinez	22	F	2	6			
21	Claude	21	F	3	6			

Procédure.

1. On entre les bornes souhaitées dans une colonne (F) puis comme précédemment on clique sur (Insertion) puis (Fonction).

2. On obtient la fenêtre **insérer une fonction**. On sélectionne (Statistiques) et la fonction (FREQUENCE).

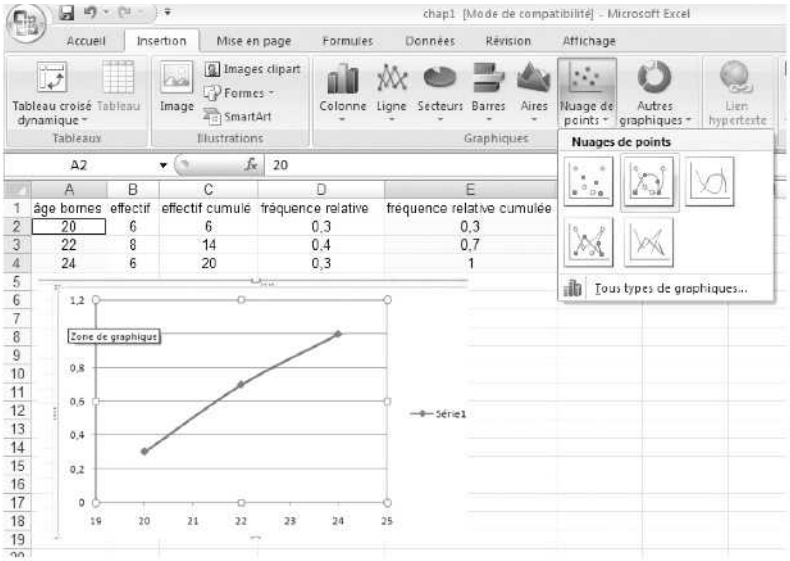
3. On sélectionne alors dans (tableau_données) les cellules correspondant aux données (B2 :B21) puis dans (matrice_intervalles) les cellules (F2 :F4) correspondant aux bornes.

4. Le résultat figure dans le coin droit de la fenêtre soit ({6,8,6,0}) et l'on clique sur (Fin). Pour restituer ces résultats il suffit de sélectionner la cellule de sortie et deux cellules adjacentes en dessous, puis d'appuyer au clavier **F2** et enfin d'appuyer simultanément sur **Ctrl Majuscule** et **Entrée**. Le tableau obtenu a la forme suivante :

X	[18, 20]	]20,22]	]22,24]
n_i	6	8	6

f) Représenter la courbe des fréquences relatives cumulées de la variable âge à partir d'une feuille excel où figurent les fréquences relatives cumulées par âge.

© Dunod. La photocopie non autorisée est un délit.



Procédure.

1. On sélectionne simultanément avec la fonction Ctrl E2 :E4 (en premier l'ordonnée) et B2 :B4 (en abscisse) on clique Insertion et sur l'icône puis la forme souhaitée.

2. En cliquant sur l'abscisse et l'ordonnée on peut donner un titre à l'axe des abscisses (âge) et un titre à l'axe des ordonnées (fréquences relatives cumulées).

2 Traitements avec le logiciel SPSS

Présentation du logiciel. Les données peuvent être issues d'un autre logiciel (un SGBD, un tableur ou un traitement de texte) ou peuvent être saisies sous SPSS.

Les fonctions assurées sont :

– **la gestion des données** à analyser (la création et la mise à jour des fichiers de données), menu **Fichier** et **Données** ;

– **la mise en forme des informations** pour une aide à l'interprétation (édition de tableaux, des rapports et des graphiques), menus **Analyse** et **Graphes** ;

– **le traitement statistique des données** à l'aide de procédures (correspondant aux méthodes statistiques usuelles), menu **Analyse**.

Les principales commandes des menus SPSS figurent ci-dessous.

Fichier :

Nouveau pour créer un nouveau fichier SPSS ;

Ouvrir pour ouvrir un fichier SPSS existant ;

Lire les données du texte pour lire un fichier ASCII ; ***Afficher infos sur les données*** pour afficher les informations sur un certain fichier (description des variables, nombre d'enregistrements) ;

Enregistrer, ***Enregistrer sous*** pour sauvegarder des enregistrements ;

Imprimer pour imprimer ;

Arrêter le processeur pour arrêter le traitement.

Édition : pour réaliser les opérations d'édition habituelles (copier, etc.).

Affichage : pour afficher notamment les noms des variables ou les données.

Données : pour définir des variables, insérer des variables, insérer une observation, trier les observations, fusionner les fichiers (selon les observations ou selon les variables), diviser le fichier, sélectionner les observations, pondérer les observations.

Transformer : pour calculer, compter, recoder, transformer, remplacer les valeurs manquantes, utiliser des variables calculées à partir des valeurs d'une variable sélectionnée, etc.

Analyser : pour sélectionner les procédures statistiques. Il dispose de plusieurs menus secondaires : *rapport* (récapituler, caractéristiques), *statistiques descriptives*, *tabuler (tableaux)*, *régression*, etc.

Graphes : pour créer des graphiques.

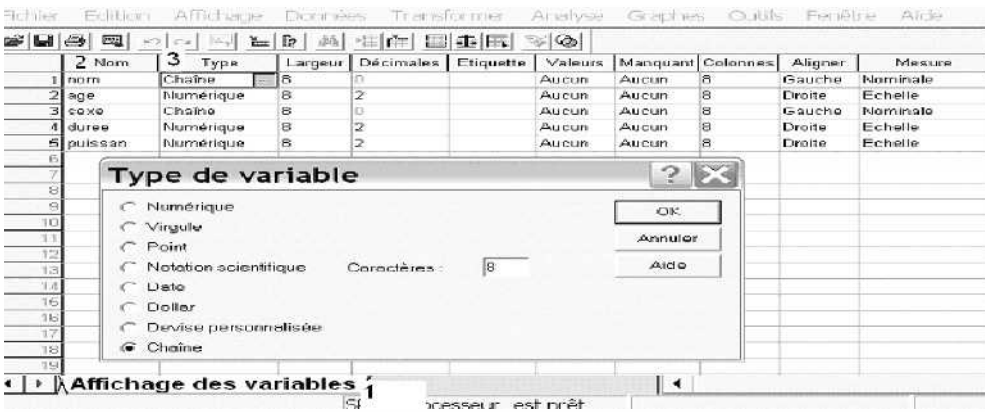
Outils : pour choisir une commande SPSS, préciser la police, etc.

Fenêtres : pour organiser les fenêtres ou passer d'une fenêtre à une autre.

Aide : pour obtenir de l'aide.

2.1 Procédure de saisie des données sous SPSS

On part du même fichier d'assurance automobile dans lequel on a prélevé 20 conducteurs âgés de moins de 25 ans et où figurent entre autres l'âge du conducteur, le nombre d'années écoulées depuis le premier contrat souscrit, la puissance fiscale de la voiture, le sexe du contractant.



© Dunod. La photocopie non autorisée est un délit.

Procédure de saisie 1. On clique en bas de la fenêtre sur (Affichage des variables),

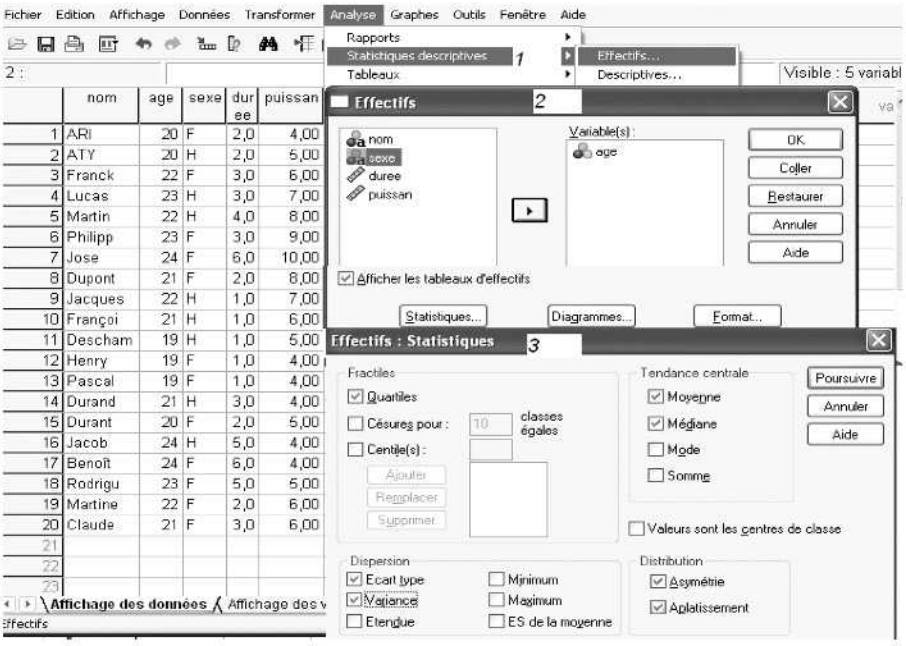
2. On définit les noms des variables (nom de variable ayant moins de huit caractères) soit ici les variables intitulées « nom », « age », « sexe », « duree », « puissan ».

3. Dans (Type) (menu déroulant) on opte entre variable (Numérique), . . ., ou une (Chaîne de caractères).

4. Puis cliquant sur (Affichage des données)**,** on entre les 20 observations en colonnes.

2.2 Traitements statistiques

a) Calculer les indicateurs de dispersion et de tendance centrale de la variable âge puis représenter sa distribution.



	nom	age	sexe	duree	puissan
1	ARI	20	F	2,0	4,00
2	ATY	20	H	2,0	5,00
3	Franck	22	F	3,0	6,00
4	Lucas	23	H	3,0	7,00
5	Martin	22	H	4,0	8,00
6	Philipp	23	F	3,0	9,00
7	Jose	24	F	6,0	10,00
8	Dupont	21	F	2,0	8,00
9	Jacques	22	H	1,0	7,00
10	Françoi	21	H	1,0	6,00
11	Descham	19	H	1,0	5,00
12	Henry	19	F	1,0	4,00
13	Pascal	19	F	1,0	4,00
14	Durand	21	H	3,0	4,00
15	Durant	20	F	2,0	5,00
16	Jacob	24	H	5,0	4,00
17	Benoît	24	F	6,0	4,00
18	Rodrigu	23	F	5,0	5,00
19	Martine	22	F	2,0	6,00
20	Claude	21	F	3,0	6,00
21					
22					
23					

Procédure.

1. On clique sur (Analyse), (Statistiques descriptives) et (Fréquences).

2. On clique sur la variable (age) que l'on envoie dans (Variable).

3. Cliquant sur (Statistiques) on obtient le menu **Effectifs : statistiques** et on spécifie les indicateurs souhaités (quartiles. . .) avant de cliquer sur (Poursuivre).

4. Cliquant sur (Diagramme) .on obtient le menu **Effectifs : Diagramme** et on sélectionne (Diagramme en bâtons) avant de cliquer sur (Poursuivre) .

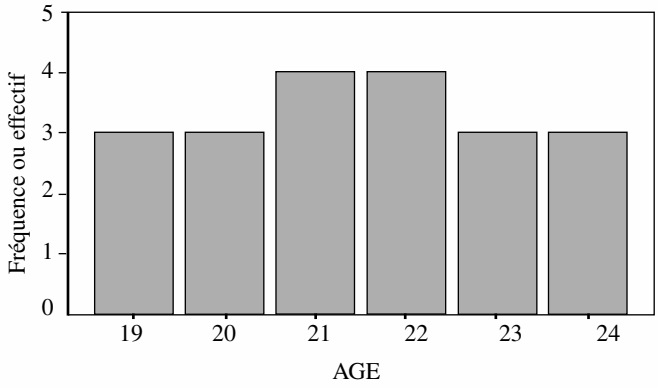
Les résultats statistiques obtenus sont les suivants :

1– la distribution des effectifs sous forme de tableau et de diagramme en bâtons

Tableau des effectifs et fréquence relative exprimée en % pour la variable âge

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	19	3	15	15	15
	20	3	15	15	30
	21	4	20	20	50
	22	4	20	20	70
	23	3	15	15	85
	24	3	15	15	100
	Total	20	100	100	

Diagramme en bâtons



À la lecture du graphe on constate que la distribution est symétrique autour de 21 et 22 ans, la distribution est bimodale avec pour valeurs modales 21 et 22 ans.

2– le tableau où figurent les indicateurs de tendance centrale et de dispersion

Statistiques	AGE				
N	Valide	20,00			
	Manquante	0,00	Erreur std. d'asymétrie		0,51
Moyenne		21,50	Aplatissement		–1,08
Médiane		21,50	Erreur std. d'aplatissement		0,99
Écart-type		1,67	Centiles	25	20,00
Variance		2,79		50	21,50
Asymétrie		0,00		75	23,00

Commentaires. Le premier indicateur de tendance centrale montre que l'âge moyen des conducteurs de l'échantillon est 21,5 ans. La médiane est 21,5, le premier quartile (25 centiles) est égal à 20, le troisième quartile ou 75 centiles est 23. La distribution est symétrique puisque le coefficient d'asymétrie est égal à 0, etc. La variance est la variance standard de l'échantillon (défini dans le chapitre 9) soit $s_x^2 = \dfrac{n}{n-1} V_x$ et l'écart-type standard de l'échantillon est $s_x = 1{,}67$.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 333)

Exercice 1

Le caractère x représente la note sur 20 attribuée à chacun des 100 candidats soumis à un test d'embauche réalisé par le cabinet de recrutement CR. Les valeurs et effectifs correspondants sont donnés par le tableau suivant :

Valeurs	[0, 5]	]5, 8]	]8, 10]	]10, 12]	]12, 15]	]15, 20]
Effectifs	25	18	12	20	15	10

1. Déterminer la classe modale puis donner les estimations respectives des valeurs de m_e, q_1 et q_3, médiane, premier et troisième quartile de l'échantillon.

2. Construire l'histogramme et le polygone des fréquences sur un même graphique.

3. Construire la courbe cumulative des fréquences sur un autre graphique.

Exercice 2

Les salaires mensuels $\boldsymbol{x}$ (en euros) des 400 salariés d'une entreprise délocalisée en République *M* se répartissent, par tranche, conformément au tableau suivant :

x	[750-800]	]800- 850]	]850-900]	]900-1100]	]1100-1500]	]1500-2000]
Effectifs	60	80	105	110	35	10

1. Évaluer l'écart-type et la moyenne de la série.

2. Déterminer S_h la part cumulée des h premières classes dans le total des salaires versés par l'entreprise pour $h = 1,\ldots,6$; faire un diagramme où figurent en abscisse les fréquences cumulées F_h de x et en ordonnée S_h puis commenter.

Exercice 3

Présentant un nouveau produit à un échantillon de 25 consommateurs, on leur demande de déterminer le prix qu'ils considèrent comme étant normal.

Valeurs	10	12	13	14	15	16	17
Effectifs	2	5	5	6	3	2	2

1. Représenter le diagramme en bâtons de la série.

2. Calculer le mode, la moyenne arithmétique, la médiane et les premier et troisième quartiles de la série.

3. Évaluer l'écart-type, l'étendue et l'intervalle interquartile de la série.

2 LES SÉRIES STATISTIQUES DOUBLES

Les séries statistiques doubles traduisent au moyen de tables ou de graphiques le rapprochement que l'on effectue entre deux caractères $\boldsymbol{x}$ et $\boldsymbol{y}$. Ces caractères peuvent correspondre à deux aspects d'une même unité statistique (*exemple* : la taille $\boldsymbol{x}$ et le poids $\boldsymbol{y}$ d'une personne) ou concerner des phénomènes distincts, mais plus ou moins liés (*exemple* : la production industrielle et les importations, la consommation et le revenu…). Enfin, l'un des caractères peut n'avoir qu'une signification de repère, $\boldsymbol{x}$ désignant par exemple le temps. La technique d'ajustement par les moindres carrés ordinaires permet d'optimiser le choix d'une relation de type $y = ax + b$ entre deux variables x et y.

Exemple

À chaque unité e_i d'un échantillon E de ménages, on associe son revenu mensuel $x_i = \boldsymbol{x}(e_i)$ (évalué en euros) et le nombre $y_i = \boldsymbol{y}(e_i)$ d'individus composant la famille.

• ***Si la taille de l'échantillon est petite***, on relève les couples de résultats au fur et à mesure des observations avec le cas échéant une répétition de certaines valeurs du couple $(\boldsymbol{x},\boldsymbol{y})$. Il s'agit alors d'une série à indice simple

• ***Si la taille de l'échantillon est grande***, les valeurs des variables $\boldsymbol{x}$ et $\boldsymbol{y}$ sont regroupées par classes et la présentation se fait avec un tableau à 2 entrées (*cf.* p. 46).

Section 1 ■ Les séries doubles à indice simple
Section 2 ■ Les séries doubles à indice double
Section 3 ■ Ajustement linaire par la méthode des moindres carrés
Section 4 ■ Ajustement non linaire de séries doubles

© Dunod. La photocopie non autorisée est un délit.

Section 1

SÉRIES DOUBLES À INDICES SIMPLES

Lorsque le nombre n d'observations est petit (cas notamment des petits échantillons) la série double est décrite par l'énumération des couples (x_i, y_i) où x_i et y_i représentent respectivement les valeurs prises par $\boldsymbol{x}$ et $\boldsymbol{y}$ pour la i-ème observation :

Tableau 2.1

$\boldsymbol{x}$	x_1	x_2	…	x_i	…	x_n
$\boldsymbol{y}$	y_1	y_2	…	y_i	…	y_n

La recherche d'un lien entre la valeur prise par $\boldsymbol{y}$ et la valeur prise par $\boldsymbol{x}$ nécessite la représentation du nuage de points $M_1 = (x_1, y_1)$, $M_2 = (x_2, y_2), \ldots,$ $M_n = (x_n, y_n)$ dans un repère orthogonal (*cf.* figure 2.1).

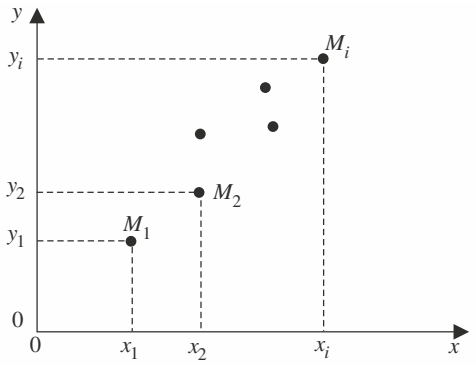


Figure 2.1 – *Nuage de points*

De la série statistique présentée dans le tableau 1 on déduit :

• *les moyennes* $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ des valeurs de $\boldsymbol{x}$ et $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ des valeurs de $\boldsymbol{y}$;

• *la variance* de $\boldsymbol{x}$ notée $V(\boldsymbol{x}) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ et l'*écart-type* $\sigma(\boldsymbol{x}) = \sqrt{V(\boldsymbol{x})}$,

la variance de $\boldsymbol{y}$ notée $V(\boldsymbol{y}) = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ et l'*écart-type* $\sigma(\boldsymbol{y}) = \sqrt{V(\boldsymbol{y})}$;

• la *covariance* des valeurs prises par le couple $(\boldsymbol{x},\boldsymbol{y})$ qui est définie comme la moyenne arithmétique des produits des écarts de chaque variable à sa moyenne :

$$\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

C'est une mesure d'association linéaire entre variables qui est positive si les valeurs de $\boldsymbol{x}$ et de $\boldsymbol{y}$ varient dans le même sens.

• Le *coefficient de corrélation linéaire* qui lie les valeurs prises par le couple $(\boldsymbol{x},\boldsymbol{y})$ est une mesure de l'intensité de la relation linéaire entre ces variables :

$$r(\boldsymbol{x},\boldsymbol{y}) = \mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})/[\sigma(\boldsymbol{x}) \times \sigma(\boldsymbol{y})].$$

Exemple

Sur un échantillon de 10 ménages, on associe le revenu mensuel $\boldsymbol{x}(e_i)$ du ménage (évalué en euros) et le nombre $\boldsymbol{y}(e_i)$ d'individus qui composent la famille :

Revenu x	633	1 106	623	802	1 206	700	900	1 000	1 200	1 050
Nombre y de personnes	2	3	1	3	3	3	4	5	4	4

Moyenne et variance des valeurs de $\boldsymbol{x}$ et de $\boldsymbol{y}$: $\bar{x} = 922$, $V(\boldsymbol{x}) = 45\,135{,}4$. $\bar{y} = 3{,}2$; $V(\boldsymbol{y}) = 1{,}16$.

Covariance des valeurs de $\boldsymbol{x}$ et de $\boldsymbol{y}$: $\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y}) = (1/10) \times [(633 - 922)(2 - 3{,}2) + (1\,106 - 922)(3 - 3{,}2) + \ldots + (1\,050 - 922)(4 - 3{,}2)] = 142{,}7$.

Coefficient de corrélation linéaire : $r(\boldsymbol{x},\boldsymbol{y}) = \mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})/[\sigma(\boldsymbol{x}) \times \sigma(\boldsymbol{y})] = 0{,}623$.

REPÈRES

Propriétés des variances et covariances. $\forall$ les réels a, b, c et d, on a :

$$V(ax + b) = V(ax) = a^2 V(x),\ V(cy + d) = V(cy) = c^2 V(y)$$

$$\mathrm{Cov}\,(ax + b, cy + d) = ac\,\mathrm{Cov}(x, y).$$

Autrement dit, si on considère les variables $x' = ax + b$ et $y' = cy + d$, on a

$$V(x') = a^2 V(x),\ V(y') = c^2 V(y),\ \mathrm{Cov}(x', y') = ac\,\mathrm{Cov}(x, y).$$

Propriétés du coefficient de corrélation linéaire

j) $-1 \leqslant r(x, y) \leqslant 1$

jj) $r(x, y) = 1$ (respectivement $r(x, y) = -1$) si et seulement si il existe deux nombres réels a_0 et b_0 avec $a_0 > 0$ (respectivement avec $a_0 < 0$) tels que $y = a_0 x + b_0$

jjj) $\forall$ les réels a, b, c et d, avec $a > 0$ et $c > 0$ on a : $r(ax + b, cy + d) = r(x, y)$.

Section **2**

SÉRIE DOUBLE À DOUBLE INDICES

1 Cas de variables discrètes

Les couples de valeurs observées sont présentés avec leur effectif.

© Dunod. La photocopie non autorisée est un délit.

1.1 Le tableau croisé

La présentation sous forme de tableau croisé est utilisée *lorsque le nombre n d'observations est grand*. Soit une série double à deux indices portant sur l'observation de n couples de valeurs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Regroupant sans perte d'informations les valeurs de $\boldsymbol{x}$ en K classes $x_1^*, x_2^*, \ldots, x_K^*$ et les valeurs de $\boldsymbol{y}$ en ℓ classes $y_1^*, y_2^*, \ldots, y_\ell^*$ on obtient le tableau suivant à deux entrées :

Tableau 2.2

Valeurs de $x \rightarrow$ / de $y \downarrow$	x_1^*	x_2^*	...	x_i^*	...	x_K^*	**Total ou distribution marginale de y**
y_1^*	n_{11}	n_{21}	...	n_{i1}	...	n_{K1}	$n_{\bullet 1}$
y_2^*	n_{12}	n_{22}	...	n_{i2}	...	n_{K2}	$n_{\bullet 2}$
...	...	...	...	...	...	...	...
y_j^*	n_{1j}	n_{2j}	...	n_{ij}	...	n_{Kj}	$n_{\bullet j}$
...	...	...	...	...	...	...	...
y_ℓ^*	$n_{1\ell}$	$n_{2\ell}$	...	$n_{i\ell}$	...	$n_{K\ell}$	$n_{\bullet \ell}$
Total ou distribution marginale de x	$n_{1\bullet}$	$n_{2\bullet}$	...	$n_{i\bullet}$	...	$n_{K\bullet}$	**Nombre total d'observations n**

Lire : parmi les n couples d'observations (x_i, y_i) on a recensé n_{11} couples égaux à (x_1^*, y_1^*).

Plus généralement, on désigne par :

– n_{ij} le nombre de couples égaux à (x_i^*, y_j^*), appelé effectif de la classe $\{x_i^*\} \times \{y_j^*\}$

– $n_{i\bullet} = \sum_{j=1}^{\ell} n_{ij} = n_{i1} + n_{i2} + \ldots + n_{i\ell}$ le nombre d'observations pour lesquelles $\boldsymbol{x} = x_i^*$

– $n_{\bullet j} = \sum_{i=1}^{K} n_{ij} = n_{1j} + n_{2j} + \ldots + n_{Kj}$ le nombre d'observations pour lesquelles $\boldsymbol{y} = y_j^*$

– $f_{ij} = n_{ij}/n$ la *fréquence relative* associée au couple de valeurs (x_i^*, y_j^*).

La somme de tous les effectifs est égale à n soit $\sum_{i=1}^{K} \sum_{j=1}^{\ell} n_{ij} = n$ et la somme des fréquences relatives est égale à 1 : $\sum_{i=1}^{K} \sum_{j=1}^{\ell} f_{ij} = 1$.

Dans ce tableau croisé on vérifie aisément que :

$$\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{\ell} n_{ij}(x_i^* - \bar{x})(y_j^* - \bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{\ell} n_{ij} x_i^* y_j^* - \bar{x}\bar{y} \text{ (formule de Köenig)}$$

Sur ce tableau à double entrée apparaissent *les distributions* marginales de $\boldsymbol{x}$ et de $\boldsymbol{y}$, *les distributions liées ou conditionnelles* de $\boldsymbol{y}$ par $\boldsymbol{x}$ et de $\boldsymbol{x}$ par $\boldsymbol{y}$.

1.2 Les distributions marginales

Il s'agit de la distribution de l'une des variables indépendamment de l'autre.

• *La distribution marginale* de $\boldsymbol{x}$ se lit sur la dernière ligne du tableau 2. Elle concerne les effectifs des valeurs de $\boldsymbol{x}$ considérées isolément :

Valeurs de x	x_1^*	x_2^*	...	x_i^*	...	x_K^*	Total
Effectifs	$n_{1\bullet}$	$n_{2\bullet}$	...	$n_{i\bullet}$	...	$n_{K\bullet}$	n

$n_{i\bullet}$ = *nombre d'observations pour lesquelles* $\boldsymbol{x} = x_i^*$

= *somme des éléments en colonne* $= n_{i1} + n_{i2} + \ldots + n_{i\ell} = \sum_{j=1}^{\ell} n_{ij}$

La moyenne et la variance se calculent aisément à partir de ce tableau :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{K} n_{i\bullet} x_i^* \text{ et } V(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{K} n_{i\bullet}(x_i^* - \bar{x})^2$$

• La *distribution marginale* de $\boldsymbol{y}$ se lit dans la dernière colonne de la table à double entrée.

Valeurs de y	y_1^*	y_2^*	...	y_j^*	...	y_ℓ^*
Effectifs	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet j}$	...	$n_{\bullet \ell}$

$n_{\bullet j}$ = *nombre d'observations pour lesquelles* $\boldsymbol{y} = y_j^*$

= *somme des éléments en ligne* $= \sum_{i=1}^{K} n_{ij} = n_{1j} + n_{2j} + \ldots + n_{Kj}$

On a évidemment $\bar{y} = \frac{1}{n}\sum_{j=1}^{\ell} n_{\bullet j} y_j^*$ et $V(\boldsymbol{y}) = \frac{1}{n}\sum_{j=1}^{\ell} n_{\bullet j}(y_j^* - \bar{y})^2$

1.3 Les distributions conditionnelles

La valeur d'une variable étant fixée, on étudie la distribution de l'autre variable.

• ***La distribution liée de y par la condition*** $\boldsymbol{x} = x_i^*$ (distribution de $\boldsymbol{y}$ sachant que $\boldsymbol{x} = x_i^*$) concerne les effectifs des valeurs de $\boldsymbol{y}$ associées à la valeur particulière x_i^* de $\boldsymbol{x}$. Elle s'obtient en utilisant la i-ème colonne du tableau de contingence :

Valeurs de y sachant que $x = x_i^*$	y_1^*	y_2^*	...	y_j^*	...	y_ℓ^*	
Effectifs	n_{i1}	n_{i2}	...	n_{ij}	...	$n_{i\ell}$	Total $n_{i\bullet}$

© Dunod. La photocopie non autorisée est un délit.

Ainsi, parmi les $n_{1\bullet}$ observations (x_i, y_i) telles que $x_i = x_1^*$, il y en a n_{11} telles que $y_i = y_1^*$, n_{12} telles que $y_i = y_2^*$,… , $n_{1\ell}$ telles que $y_i = y_\ell^*$.

Pour chacune de ces K distributions conditionnelles (ou liées) et partant des ces tableaux, on peut calculer la moyenne et la variance appelées respectivement :

– *moyenne liée ou conditionnelle* de $\boldsymbol{y}$ pour $\boldsymbol{x} = x_i^*$ (moyenne de $\boldsymbol{y}$ sachant que $\boldsymbol{x} = x_i^*$) : $\bar{y}_i^* = \dfrac{1}{n_{i\bullet}} \sum_{j=1}^{\ell} n_{ij} y_j^* = \dfrac{1}{n_{i\bullet}} (n_{i1} y_1^* + n_{i2} y_2^* + \ldots + n_{i\ell} y_\ell^*)$

– *variance liée ou conditionnelle* de $\boldsymbol{y}$ pour $\boldsymbol{x} = x_i^*$

$$V(\boldsymbol{y}/\boldsymbol{x} = x_i^*) = \frac{1}{n_{i\bullet}} \sum_{j=1}^{\ell} n_{ij} (y_j^* - \bar{y}_i^*)^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{\ell} n_{ij} (y_j^*)^2 - (\bar{y}_i^*)^2$$

• ***La distribution liée de x** par **la condition** $\boldsymbol{y = y_j^*}$* concerne les effectifs des valeurs de $\boldsymbol{x}$ associées à la valeur particulière y_j^* de $\boldsymbol{y}$. Elle est caractérisée par le tableau

Valeurs de x	x_1^*	x_2^*	…	x_K^*	Total
Effectifs lorsque $y = y_j^*$	n_{1j}	n_{2j}		n_{Kj}	$n_{\bullet j}$

Pour chacune de ces ℓ distributions on définit

– la *moyenne liée ou conditionnelle* de $\boldsymbol{x}$ pour $\boldsymbol{y} = y_j^*$: $\bar{x}_j^* = \dfrac{1}{n_{\bullet j}} \sum_{i=1}^{K} n_{ij} x_i^*$

– la *variance* de $\boldsymbol{x}$ sachant que $\boldsymbol{y} = y_j^*$ dite variance *intraclasse* :

$$V(\boldsymbol{x}/\boldsymbol{y} = y_j^*) = \frac{1}{n_{\bullet j}} \sum_{i=1}^{K} n_{ij} (x_i^* - \bar{x}_j^*)^2 = \frac{1}{n_{\bullet j}} \sum_{i=1}^{K} n_{ij} (x_i^*)^2 - (\bar{x}_j^*)^2$$

REPÈRES : Propriétés des distributions conditionnelles

– La moyenne $\bar{x}$ est la moyenne pondérée des ℓ moyennes liées $\bar{x}_j^*$: $\bar{x} = \dfrac{1}{n} \sum_{j=1}^{\ell} n_{\bullet j} \bar{x}_j^*$

la variance $V(\boldsymbol{x})$ peut s'exprimer comme étant la somme de la ***v**ariance **d**es **m**oyennes liées* (ou *variance interclasse*) et de la ***m**oyenne **d**es **v**ariances liées* (ou *moyenne des variances intraclasse*) :

$V(\boldsymbol{x}) = \dfrac{1}{n} \sum_{j=1}^{\ell} n_{\bullet j} (\bar{x}_j^* - \bar{x})^2 + \dfrac{1}{n} \sum_{j=1}^{\ell} n_{\bullet j} \times V(x \,/\, y = y_j^*)$; on dit que $V(\boldsymbol{x})$ = VDM + MDV

– De même, $\bar{y} = \dfrac{1}{n} \sum_{i=1}^{K} n_{i\bullet} \bar{y}_i^*$; $V(y) = \dfrac{1}{n} \sum_{i=1}^{K} n_{i\bullet} (\bar{y}_i^* - \bar{y})^2 + \dfrac{1}{n} \sum_{i=1}^{K} n_{i\bullet} \times V(y \,/\, x = x_i^*)$.

Exemple

Sur un échantillon de 86 ménages, on associe le nombre $\boldsymbol{x}(e_i)$ d'individus qui composent la famille et le nombre $\boldsymbol{y}(e_i)$ de chambres dans l'habitation principale. Au tableau statistique, on joint les moyennes et variances conditionnelles $\bar{x}_j^*$, $\bar{y}_i^*$ $V(\boldsymbol{x}/\boldsymbol{y} = y_j^*)$, $V(\boldsymbol{y}/\boldsymbol{x} = x_i^*)$.

	$\boldsymbol{x}=1$	$\boldsymbol{x}=2$	$\boldsymbol{x}=3$	$\boldsymbol{x}=4$	$\boldsymbol{x}=5$	$\boldsymbol{x}=6$	$\boldsymbol{x}=7$	Distribution marginale de $\boldsymbol{y}$	Moyennes liées $\bar{x}_j^*$	Variance intraclasse $V(\boldsymbol{x}/\boldsymbol{y}=y_j^*)$
$\boldsymbol{y}=1$	22	7	1	1			0	$n_{\bullet 1}=31$	1,39	0,50
$\boldsymbol{y}=2$	2	5	2	1				$n_{\bullet 2}=10$	2,20	0,76
$\boldsymbol{y}=3$		3	8	3	2	2	1	$n_{\bullet 3}=19$	3,74	1,98
$\boldsymbol{y}=4$	3	1	3	5	2	2	2	$n_{\bullet 4}=18$	3,89	3,43
$\boldsymbol{y}=5$			1	2	1	2	2	$n_{\bullet 5}=8$	5,25	1,94
Distribution marginale de $\boldsymbol{x}$	$n_{1\bullet}=27$	$n_{2\bullet}=16$	$n_{3\bullet}=15$	$n_{4\bullet}=12$	$n_{5\bullet}=5$	$n_{6\bullet}=6$	$n_{7\bullet}=5$	$n=86$	2,88	3,36
Moyennes liées $\bar{y}_i^*$	0,56	1,69	1,07	1,25	2,40	0,83	1,20			
$V(\boldsymbol{y}/\boldsymbol{x}=x_i^*)$ Variance intraclasse	2,58	1,53	9,13	11,94	9,24	15,97	16,76			

On détermine la distribution marginale de $\boldsymbol{x}$ en lisant le total des effectifs par colonne :

$\boldsymbol{x}$	1	2	3	4	5	6	7
Effectif $\boldsymbol{n_{i\bullet}}$	27	16	15	12	5	6	5

et on en déduit les moyennes et variances des valeurs $\boldsymbol{x}(e_i)$: $\bar{x} = 2{,}8837$; $V(\boldsymbol{x}) = (1/86) \times [27 \times 1^2 + \ldots + 5 \times 7^2] - (2{,}8837)^2 = 3{,}358$.

On obtient la *distribution marginale de* $\boldsymbol{y}$ en lisant le total des effectifs par ligne :

$\boldsymbol{y}$	1	2	3	4	5
Effectif $\boldsymbol{n_{\bullet i}}$	31	10	19	18	8

et : $\overline{y} = 2{,}5581$; $V(\boldsymbol{y}) = 1{,}944$.

Sur le tableau figurent les moyennes liées et variances *intraclasses*. Par exemple :

– la *moyenne liée ou conditionnelle* de $\boldsymbol{y}$ pour $\boldsymbol{x} = x_1^*$ (c'est-à-dire la moyenne des valeurs de y pondérée par les effectifs de la première colonne) :

$$\bar{y}_1^* = \frac{1}{n_{1\bullet}} \sum_{j=1}^{\ell} n_{1j} y_j^* = (1/27) \times (22 \times 1 + 2 \times 2 + \ldots + 0 \times 5) = 0{,}56 \ ;$$

la *variance liée ou conditionnelle* de $\boldsymbol{y}$ pour $\boldsymbol{x} = x_1^*$:

$$V(\boldsymbol{y}/\boldsymbol{x} = x_1^*) = \frac{1}{27} \sum_{j=1}^{5} n_{1j} (y_j^* - 0{,}56)^2 = \frac{1}{27} \sum_{j=1}^{5} n_{1j} (y_j^*)^2 - (0{,}56)^2$$

$$= (1/27) \times [22 \times 1^2 + 2 \times 2^2 + 0 \times 3^2 + 3 \times 4^2 + 0 \times 5^2) - 0{,}56^2 = 2{,}58 \ ;$$

© Dunod. La photocopie non autorisée est un délit.

2 Cas mixte de variable discrète et continue

Lorsque le nombre de valeurs prises par $\boldsymbol{x}$ est trop grand pour figurer dans un tableau de taille raisonnable, les valeurs prises par $\boldsymbol{x}$ sont regroupées en K classes.

Tableau 2.3

Valeurs de $\boldsymbol{x}\rightarrow$ / de $\boldsymbol{y}\downarrow$	$a_0 < \boldsymbol{x} \leqslant a_1$	$a_1 < \boldsymbol{x} \leqslant a_2$	...	$a_{i-1} < \boldsymbol{x} \leqslant a_i$	...	$a_{K-1} < \boldsymbol{x} \leqslant a_K$	Total ou distribution marginale de $\boldsymbol{y}$
y_1^*	n_{11}	n_{21}	...	n_{i1}	...	n_{K1}	$n_{\bullet 1}$
y_2^*	n_{12}	n_{22}	...	n_{i2}	...	n_{K2}	$n_{\bullet 2}$
...	...	...	...	...	...	...	...
y_j^*	n_{1j}	n_{2j}	...	n_{ij}	...	n_{Kj}	$n_{\bullet 2}$
...	...	...	...	...	...	...	...
y_ℓ^*	$n_{1\ell}$	$n_{2\ell}$	...	$n_{i\ell}$	...	$n_{K\ell}$	$n_{\bullet \ell}$
Total ou distribution marginale de $\boldsymbol{x}$	$n_{1\bullet}$	$n_{2\bullet}$	...	$n_{i\bullet}$	...	$n_{K\bullet}$	n

avec éventuellement $a_0 = -\infty$ et $a_h = +\infty$.

Parmi les n couples d'observations (x_i, y_i) on a recensé n_{11} couples tels que $a_0 < x \leqslant a_1$ et $y = y_1^*$. Plus généralement, on a recensé n_{ij} couples tels que $a_{i-1} < x \leqslant a_i$ et $y = y_j^*$.

Le fait de regrouper dans une même classe des valeurs x_i « voisines » mais distinctes crée une perte d'informations. Cela ne permet pas d'avoir des valeurs exactes mais seulement des estimations de $\bar{x}$, $V(\boldsymbol{x})$, $\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})$ et $r(\boldsymbol{x},\boldsymbol{y})$. Ces estimations sont obtenues en considérant que les valeurs qui sont regroupées dans la classe $a_{i-1} < x \leqslant a_i$ ont pour valeur moyenne $x_i^* = (a_{i-1} + a_i)/2$. En revanche le tableau 2.3 permet d'obtenir les valeurs exactes de $\bar{y}$ et de $V(\boldsymbol{y})$.

La distribution liée de y *par* ***la condition*** « la valeur de $\boldsymbol{x}$ appartient à la i-ème classe » concerne les effectifs des valeurs de $\boldsymbol{y}$ associées aux valeurs particulières de $\boldsymbol{x}$ qui sont comprises entre a_{i-1} et a_i.

Valeurs de $\boldsymbol{y}$	y_1^*	y_2^*	...	y_j^*	...	y_ℓ^*	
Effectifs	n_{i1}	n_{i2}	...	n_{ij}	...	$n_{i\ell}$	Total $= n_{i\bullet}$

On définit pour chacune de ces K distributions liées : la *moyenne liée* et la *variance liée.*

De même, on a une estimation de

– la *moyenne liée ou conditionnelle* de $\boldsymbol{x}$ pour $\boldsymbol{y} = y_j^*$ soit

$$\bar{x}_j^* \cong \frac{1}{n_{\bullet j}} \sum_{i=1}^{K} n_{ij} x_i^* \text{ où } x_i^* = (a_{i-1} + a_i)/2 \text{ est le centre de la classe } a_{i-1} < x \leqslant a_i$$

– de *la variance liée ou conditionnelle* de $\boldsymbol{x}$ pour $\boldsymbol{y} = y_j^*$ soit

$$V(\boldsymbol{x}/\boldsymbol{y} = y_j^*) \cong \frac{1}{n_{\bullet j}} \sum_{i=1}^{K} n_{ij}(x_i^* - \bar{x}_j^*)^2 .$$

Section 3

AJUSTEMENT LINÉAIRE PAR LA MÉTHODE DES MOINDRES CARRÉS

Exemple

Sont présentés dans le tableau ci-dessous, pour une même période, les relevés de consommation (noté $\boldsymbol{x}$) et de revenu disponible (noté $\boldsymbol{y}$) d'un échantillon de 5 ménages. Autrement dit cinq couples de valeurs $(x_1, y_1), (x_2, y_2), \ldots (x_5, y_5)$.

Numéro d'observation	1	2	3	4	5
$\boldsymbol{x}$	500	1 000	1 500	1 750	2 000
$\boldsymbol{y}$	500	900	1 100	1 200	1 700

La consommation $\boldsymbol{y}$ dépend évidemment du revenu $\boldsymbol{x}$ ($\boldsymbol{x}$ est la variable *exogène*, $\boldsymbol{y}$ la variable *endogène*).

Graphe. – On représente les points $P_1 = (500,500)$, $P_2 = (1\,000,900)$, $P_3 = (1\,500, 1\,100)$, $P_4 = (1\,750,1\,250)$, $P_5 = (2\,000,1\,700)$.

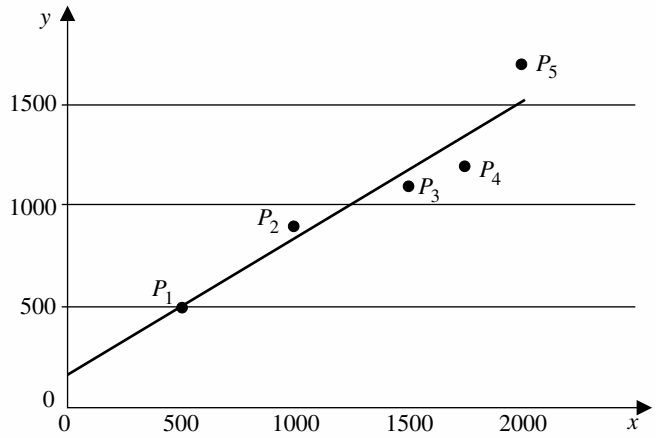


Figure 2.2 – *Ajustement du nuage de points*

Les points P_i semblent alignés. Traçant une droite qui « résume » le nuage de points, on obtient une relation théorique de type $\boldsymbol{y} = a\boldsymbol{x} + b$ (soit une relation linéaire affine). Il convient alors de déterminer quelle est la meilleure droite d'ajustement.

© Dunod. La photocopie non autorisée est un délit.

1 Série double à indice simple

Soit donc la série double à indice simple $(\boldsymbol{x},\boldsymbol{y})$ où $\boldsymbol{y}$ semble dépendre linéairement de $\boldsymbol{x}$:

Tableau 2.4

$\boldsymbol{x}$	x_1	x_2	...	x_{n-1}	x_n
$\boldsymbol{y}$	y_1	y_2	...	y_{n-1}	y_n

On place les n points $P_1(x_1,y_1), P_2(x_2,y_2), \ldots, P_n(x_n,y_n)$ dans un repère orthonormé d'abscisse x et d'ordonnée y et on cherche la droite d'équation $\boldsymbol{y} = a\boldsymbol{x} + b$ qui passe au plus près des points.

1.1 Méthode des moindres carrés

À chaque choix de la droite d'ajustement $\boldsymbol{y} = a\boldsymbol{x} + b$ (c'est-à-dire à chaque choix des constantes réelles a et b) correspond l'expression :

$$\mathbf{e}(a,b) = |P_1H_1|^2 + |P_2H_2|^2 + \ldots + |P_nH_n|^2 = \sum_{i=1}^{n}(y_i - y_i')^2$$

– où $|P_iH_i|$ est l'écart entre le point observé $P_i(x_i,y_i)$ et le point $H_i(x_i,y_i')$ projection verticale de P_i sur la droite $\boldsymbol{y} = a\boldsymbol{x} + b$ (voir figure 2.3 ci-après) ;

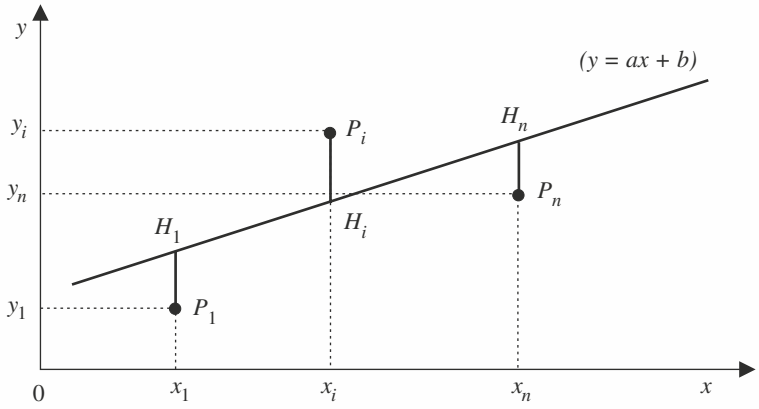


Figure 2.3

– $\mathbf{e}(a,b)$ représente la somme des carrés des écarts entre la valeur réelle y_i (valeur effective de $\boldsymbol{y}$ pour $\boldsymbol{x} = x_i$) et la valeur théorique $y_i' = ax_i + b$ (ou valeur de $\boldsymbol{y}$ pour $\boldsymbol{x} = x_i$ si $\boldsymbol{y} = a\boldsymbol{x} + b$).

Lorsque $\mathbf{e}(a,b) = 0$, cela signifie que les points sont tous alignés. Plus $\mathbf{e}$ est petit, meilleur sera l'ajustement c'est-à-dire le choix de a et b.

On doit chercher le couple de valeurs (a_0, b_0) qui minimise la fonction à deux variables $\mathbf{e}(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$ et l'on en déduit que cette droite a pour équation $\boldsymbol{y} = a_0\boldsymbol{x} + b_0$ où

$$a_0 = \mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})/V(\boldsymbol{x}) \quad \text{et} \quad b_0 = \bar{y} - a_0\bar{x}.$$

Elle passe donc par le point G de coordonnées $(\bar{x}, \bar{y})$.

Partant des coefficients de cette droite on définit :

– les ***valeurs estimées*** par l'ajustement linéaire : $y_i' = a_0x_i + b_0$;

– la ***variance expliquée*** $V_E(y)$ (ou variance des valeurs expliquées $y_1' = a_0x_1 + b_0$, $y_2' = a_0x_2 + b_0$, … , $y_n' = a_0x_n + b_0$), soit $V_E(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i' - \bar{y}')^2$ où $\bar{y}' = \bar{y}$;

– les ***écarts*** algébriques entre les valeurs estimées et les valeurs effectives $e_i = y_i' - y_i$;

– la ***variance résiduelle*** (ou variance des écarts), soit $V_R(y) = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2$

$= \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2$ puisque $\bar{e} = \frac{1}{n}\sum_{i=1}^{n}e_i = 0$

ou de façon équivalente $V_R(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - a_0x_i - b_0)^2$.

Repères

La meilleure droite d'ajustement par la méthode des moindres carrés a pour équation $y = a_0x + b_0$ où $a_0 = \mathrm{Cov}(x, y)/V(x)$ et $b_0 = \bar{y} - a_0\bar{x}$. Elle permet de déterminer chaque valeur $y' = a_0x_i + b_0$ qui est la valeur expliquée de y pour $x = x_i$. Peuvent alors être évaluées :

– la variance expliquée $V_E(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i' - \bar{y}')^2 = V(y) \times r^2(x, y)$ où

$r(x, y) = \mathrm{Cov}(x, y)/[\sigma(x) \times \sigma(y)]$;

– la variance résiduelle $V_R(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 = V(y) - V_E(y) = V(y) \times (1 - r^2(x, y))$

En dépit de la formulation hâtive « les variations de $\boldsymbol{x}$ expliquent partiellement les variations de $\boldsymbol{y}$ », l'existence d'une forte corrélation n'implique pas nécessairement un lien de causalité.

Remarque. La droite de régression de $\boldsymbol{x}$ sur $\boldsymbol{y}$ a pour équation $\boldsymbol{x} = a_0'\boldsymbol{y} + b_0'$ où $a_0' = \mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})/V(\boldsymbol{y})$ et $b_0' = \bar{x} - a_0'\bar{y}$. On a la relation $a_0 \times a_0' = r^2(\boldsymbol{x},\boldsymbol{y})$.

© Dunod. La photocopie non autorisée est un délit.

1.2 Qualité de l'ajustement et coefficient de corrélation linéaire

La valeur du coefficient de détermination $r^2(\boldsymbol{x},\boldsymbol{y}) = (\text{Cov}(\boldsymbol{x},\boldsymbol{y})/[\sigma(\boldsymbol{x}) \times \sigma(\boldsymbol{y})])^2$ est comprise entre 0 et 1. Une valeur $r^2(\boldsymbol{x},\boldsymbol{y}) = 1$ correspond à une variance résiduelle nulle ($V_R(y) = 0$) et à un ajustement linéaire parfait puisque $e_1 = e_2 = \ldots = e_n = 0$.

Une valeur $r^2(\boldsymbol{x},\boldsymbol{y}) = 0$ traduit *a contrario* l'inexistence d'une relation fonctionnelle de type $\boldsymbol{y} = a_0\boldsymbol{x} + b_0$.

Plus $r^2(\boldsymbol{x},\boldsymbol{y})$ est élevé et meilleure est la qualité de l'ajustement.

Exemple

Sont présentés pour une période déterminée, les relevés de consommation $\boldsymbol{y}$ et de revenu disponible $\boldsymbol{x}$ sur un échantillon de 10 ménages.

Valeurs de x	510	1 020	1 505	1 750	1 995	695	1 205	1 680	1 950	2 190
Valeurs de y	490	910	1 105	1 195	1 720	690	1 110	1 280	1 405	1 895

La consommation $\boldsymbol{y}$ dépend du revenu disponible $\boldsymbol{x}$ ($\boldsymbol{x}$ est la variable explicative ou *exogène* et $\boldsymbol{y}$ est la variable expliquée ou *endogène*). Pour obtenir l'équation de la droite d'ajustement de $\boldsymbol{y}$ sur $\boldsymbol{x}$ par la méthode des **m**oindres **c**arrés **o**rdinaire (MCO), on détermine au préalable les valeurs :

$\bar{x} = 1\,450$; $\bar{y} = 1\,180$; $V(\boldsymbol{x}) = 293\,910$, $V(\boldsymbol{y}) = 166\,330$,

$\text{Cov}(\boldsymbol{x},\boldsymbol{y}) = (1/10) \times [510 \times 490 + 1\,020 \times 910 + \ldots + 2\,190 \times 1\,895] - 1\,450 \times 1\,180 = 211\,107$.

On a $a_0 = \text{Cov}(\boldsymbol{x},\boldsymbol{y})/V(\boldsymbol{x}) = 0{,}718$ et $b_0 = \bar{y} - a_0\bar{x} = 138{,}5$ donc la droite d'ajustement a pour équation $\boldsymbol{y} = 0{,}718\boldsymbol{x} + 138{,}5$ soit une propension moyenne à consommer égale à 0,718 et un niveau de consommation fixe indépendant du revenu égal à 138,5.

Pour évaluer la qualité de l'ajustement on détermine la valeur du coefficient de détermination $R^2 = (r(\boldsymbol{x},\boldsymbol{y}))^2 = (\text{Cov}(\boldsymbol{x},\boldsymbol{y}))^2/[V(\boldsymbol{x}) \times V(\boldsymbol{y})] = 0{,}9116$. La valeur de R^2 étant proche de 1, l'ajustement linéaire obtenu est de bonne qualité.

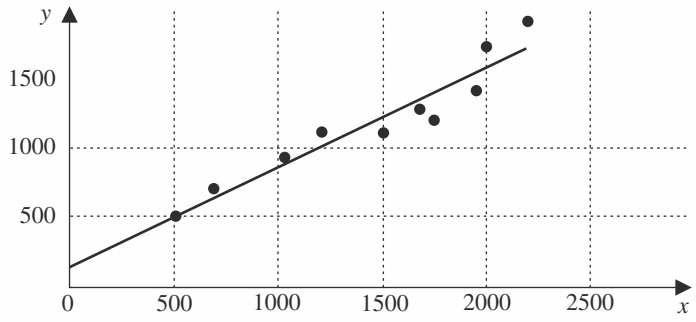


Figure 2.4 – *Graphe du nuage de points et de la droite d'ajustement*

2 Série double à deux indices

On dispose de n couples d'observations (x_1,y_1), (x_2,y_2), …, (x_n,y_n), regroupés en classes où n_{ij} est le nombre de couples d'observations égaux à (x_i^*,y_j^*) :

Valeurs de x → / de y ↓	x_1^*	x_2^*	…	x_i^*	…	x_K^*	**Total ou distribution marginale de y**
y_1^*	n_{11}	n_{21}	...	n_{i1}	...	n_{K1}	$n_{\bullet 1}$
...	...	...	...	...	...	...	...
y_j^*	n_{1j}	n_{2j}	...	n_{ij}	...	n_{Kj}	$n_{\bullet j}$
...	...	...	...	...	...	...	...
y_ℓ^*	$n_{1\ell}$	$n_{2\ell}$	...	$n_{i\ell}$	...	$n_{K\ell}$	$n_{\bullet \ell}$
Total ou distribution marginale de x	$n_{1\bullet}$	$n_{2\bullet}$	...	$n_{i\bullet}$	...	$n_{K\bullet}$	Nombre total d'observations n

À la valeur x_i^* correspondent ℓ valeurs y_1^*, y_2^*,… , y_ℓ^* avec les effectifs respectifs $n_{i,1}, n_{i,2},\ldots,n_{il}$, (éventuellement $n_{ij}=0$). On obtient ainsi un nuage de points que l'on souhaite ajuster par une droite (*cf.* figure 2.5).

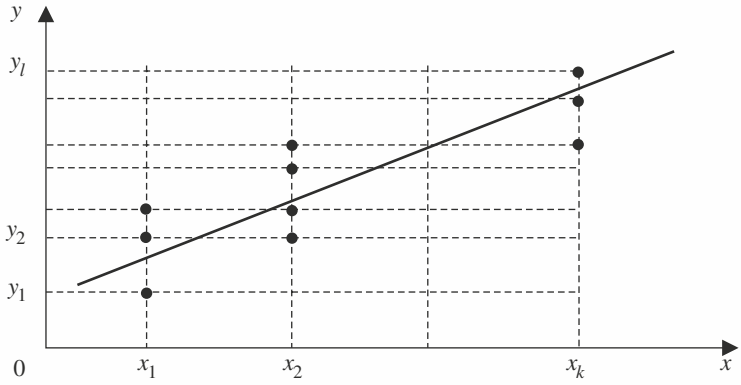


Figure 2.5

Par la méthode des moindres carrés, l'ajustement linéaire optimal $\boldsymbol{y} = a\boldsymbol{x} + b$ sera celui pour lequel $\mathbf{e}(a,b) = \sum_{i=1}^{K}\sum_{j=1}^{\ell} n_{ij}(y_j^* - ax_i^* - b)^2$ est *minimum*.

On obtient : $a_0 = \dfrac{\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})}{V(\boldsymbol{x})} = \dfrac{\sum_{i=1}^{K}\sum_{j=1}^{\ell} n_{ij}x_i^*y_j^* - n\bar{x}\times\bar{y}}{\sum_{i=1}^{K} n_{i\bullet}(x_i^*)^2 - n\bar{x}^2}$; $b_0 = \bar{y} - a_0\bar{x}$

La régression de $\boldsymbol{y}$ sur $\boldsymbol{x}$ peut faire apparaître une concomitance entre les variations des valeurs x_i^* de $\boldsymbol{x}$ et les variations des moyennes liées $\bar{y}_i^*$ (qui résument l'ensemble des $n_{i\bullet}$ valeurs de $\boldsymbol{y}$ associées à x_i^*). Si « $\forall i$, $x_{i+1}^* > x_i^* \Longrightarrow \bar{y}_{i+1}^* > \bar{y}_i^*$ »,

© Dunod. La photocopie non autorisée est un délit.

autrement dit si une augmentation des valeurs x_i^* s'accompagne d'une augmentation des valeurs $\bar{y}_i^*$ on dit que les variables $\boldsymbol{x}$ et $\boldsymbol{y}$ sont corrélées positivement.

Lorsque le nuage formé par les points correspondant aux couples $(x_i^*, \bar{y}_i^*)$ de l'ensemble de régression, est d'allure rectiligne, l'ajustement est de type linéaire.

Les propriétés présentées dans le cas de l'indexation simple s'appliquent évidemment.

Exemple

Sur un échantillon de 108 ménages dans une grande ville, on associe le revenu annuel du ménage $\boldsymbol{x}$ exprimé en dizaine de milliers d'euros et la surface $\boldsymbol{y}$ en m^2 de la résidence principale du ménage. Les résultats statistiques sont regroupés dans le tableau ci-après.

Dans une même classe peuvent figurer des couples de valeurs distincts. Par exemple (0,61, 14) et (0,18, 21) sont regroupés dans la classe $[0\,;2]\times[10\,;25]$. De ce fait on ne peut qu'obtenir une estimation des valeurs de $\bar{x}$, $V(\boldsymbol{x})$, $\bar{y}$, $V(\boldsymbol{y})$, $\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})$, a_0 et b_0. Faisant l'hypothèse d'une répartition uniforme au sein de chaque classe, on prend pour valeur de $\boldsymbol{x}$ et de $\boldsymbol{y}$ les valeurs des centres de classes $x_i^* = (a_{i-1} + a_i)/2$; $y_j^* = (b_{j-1} + b_j)/2$ qui figurent respectivement dans l'avant-avant dernière ligne et avant-avant dernière colonne.

	$x \leqslant 2$	$1{,}5 < x \leqslant 2{,}5$	$2{,}5 < x \leqslant 3{,}5$	$3{,}5 < x \leqslant 6{,}5$	Total	y_j^*	$n_{\bullet j} y_j^*$	$n_{\bullet j} y_j^{*2}$
$10 \leqslant y \leqslant 25$	25			3	$n_{\bullet 1} = 28$	17,5	490	8 575,0
$25 < y \leqslant 45$	7	25	3	1	$n_{\bullet 2} = 36$	35	1 260	44 100,0
$45 < y \leqslant 85$	1	2	8	11	$n_{\bullet 3} = 22$	65	1 430	92 950,0
$85 < y \leqslant 125$	1	1	3	10	$n_{\bullet 4} = 15$	105	1 575	165 375,0
$125 < y \leqslant 250$			2	5	$n_{\bullet 5} = 7$	187,5	1 312,5	246 093,8
Total	$n_{1\bullet} = 34$	$n_{2\bullet} = 28$	$n_{3\bullet} = 16$	$n_{4\bullet} = 30$	108		$\Sigma =$ 6 067,5	$\Sigma =$ 557 093,8
x_i^*	1	2	3	5				
$n_{i\bullet} x_i^*$	34	56	48	150	$\Sigma = 288$			
$n_{i\bullet}(x_i^*)^2$	34	112	144	750	$\Sigma = 1\,040$			

Calculs intermédiaires. Les moyennes et les variances de $\boldsymbol{x}$ et de $\boldsymbol{y}$ sont

$\bar{x} = 288/108 = 2{,}66$; $V(\boldsymbol{x}) = 1\,040/108 - 2{,}66^2 = 2{,}518$;

$\bar{y} = 6\,067{,}5/108 = 56{,}18$; $V(\boldsymbol{y}) = 557\,093{,}8 - 56{,}18^2 = 2\,002$;

De la valeur de la covariance entre $\boldsymbol{x}$ et $\boldsymbol{y}$: $\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{K}\sum_{j=1}^{\ell} n_{ij} x_i y_j - \bar{x}\bar{y}$

$= [25 \times (1 \times 17{,}5) + 7 \times (1 \times 35) + \ldots + 5 \times (5 \times 187{,}5)]/108 - 2{,}66 \times 56{,}18$
$= 44{,}32$ sont déduits $a_0 = \mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})/V(\boldsymbol{x}) = 17{,}6$ et $b_0 = \bar{y} - a_0\bar{x} = 9{,}244$.

La droite d'ajustement a pour équation $y = 17{,}6x + 9{,}244$.

Cet ajustement est de médiocre qualité car le carré du coefficient linéaire

$$R^2 = (\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y}))^2/V(\boldsymbol{x})(V(\boldsymbol{y}) = 0{,}39.$$

Section 4

AJUSTEMENTS NON LINÉAIRES DE SÉRIES DOUBLES

Soit une série double à indices simples.

Tableau 2.5

$\boldsymbol{x}$	x_1	x_2	...	x_i	...	x_n
$\boldsymbol{y}$	y_1	y_2	...	y_i	...	y_n

On peut envisager notamment des ajustements non linéaires de $\boldsymbol{y}$ sur $\boldsymbol{x}$ de type logarithmique « $y = a\ln(x) + b$ », exponentiel « $y = K\mathrm{e}^{ax}$ », parabolique « $y = ax^2 + bx + c$ ». Ces ajustements peuvent se ramener par des changements de variables à des ajustements linéaires.

1 Ajustement logarithmique

Sur une feuille quadrillée où l'abscisse est à l'échelle logarithmique et l'ordonnée à l'échelle arithmétique, portons les couples de valeurs (x_i, y_i). Si les points $P_1(x_1, y_1), P_2(x_2, y_2), \ldots, P_n(x_n, y_n)$ ainsi obtenus sont presque alignés, on peut envisager l'ajustement $\boldsymbol{y} = a\ln(\boldsymbol{x}) + b$ où a et b sont des constantes réelles.

Les valeurs de a et b s'obtiennent par la méthode des moindres carrés en introduisant la variable auxiliaire $\boldsymbol{x}' = \ln(\boldsymbol{x})$ et en considérant les couples (x'_1, y_1), $(x'_2, y_2), \ldots, (x'_n, y_n)$, où $x'_i = \ln(x_i)$. On a en effet $\boldsymbol{y} = a_0\boldsymbol{x}' + b_0$ et par la suite $a_0 = \mathrm{Cov}(\boldsymbol{x}', \boldsymbol{y}) / V(\boldsymbol{x})$, et $b_0 = \bar{y} - a_0\bar{x}'$ (*cf.* exercice 2).

2 Ajustement exponentiel

Si les points (x_i, y_i) semblent alignés sur une feuille quadrillée où l'abscisse est à l'échelle *arithmétique* et l'ordonnée à l'échelle *logarithmique*, on introduit la variable statistique $\boldsymbol{y}' = \ln(\boldsymbol{y})$ et on considère les n couples d'observation (x_1, y'_1), $(x_2, y'_2), \ldots, (x_n, y'_n)$ où $y'_i = \ln(y_i)$.

De l'ajustement linéaire : $\boldsymbol{y}' = a_0\boldsymbol{x} + b_0$ où $a_0 = \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}') / V(\boldsymbol{x})$, et $b_0 = \bar{y}' - a_0\overline{x}$, on déduit l'ajustement exponentiel $\boldsymbol{y} = K\mathrm{e}^{a_0 x}$ où $K = \mathrm{e}^{b_0}$.

© Dunod. La photocopie non autorisée est un délit.

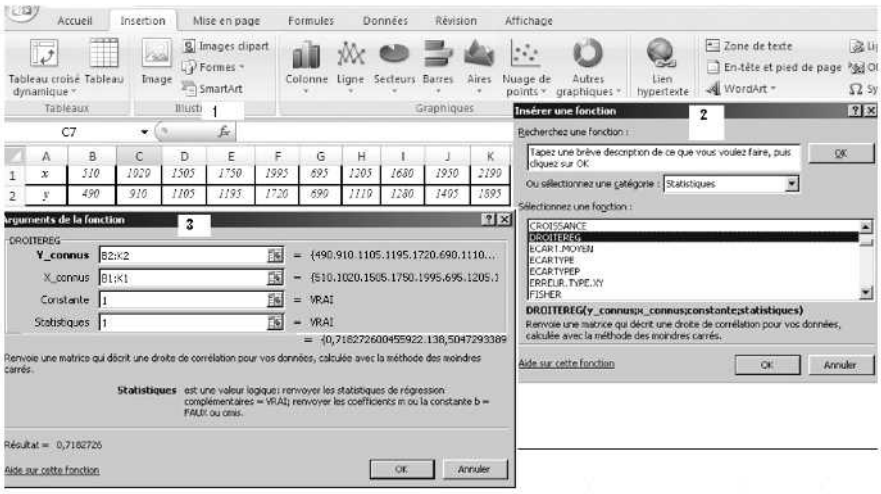
Section 5

ANALYSES STATISTIQUES AVEC EXCEL

Exemple

Sont présentés pour une période déterminée, les relevés de consommation y et de revenu disponible x sur un échantillon de 10 ménages.

Valeurs de x	510	1 020	1 505	1 750	1 995	695	1 205	1 680	1 950	2 190
Valeurs de y	490	910	1 105	1 195	1 720	690	1 110	1 280	1 405	1 895



La consommation y dépend du revenu disponible x (x est la variable explicative ou *exogène* et y est la variable expliquée ou *endogène*). Pour obtenir l'équation de la droite d'ajustement de y sur x par la méthode des moindres carrés ordinaire (MCO), on peut utiliser les fonctions de Excel.

Procédure.

1. On clique sur (Insertion) et on clique sur ($f\,x$)

2. On obtient le menu insérer une fonction on sélectionne une catégorie (Statistiques) puis une fonction (DROITEREG) en déroulant.

3. Dans arguments de la fonction on spécifie les plages des valeurs de Y soit (B2 : K2) et de X soit (B1 : K1) puis on met 1 dans constante pour dire qu'il y a une constante b et on clique sur (OK)

Résultat

On lit que $a = 0{,}718$ et que $b = 138{,}50$ et donc les coefficients de la droite de régression

$$y = 0{,}718x + 138{,}5$$

4. On peut au niveau de l'étape 2 déterminer le coefficient de corrélation en sélectionnant dans le menu Statistiques **COEFFICIENT.CORRELATION** puis **COEFFICIENT.DETERMINATON**

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 337)

Exercice 1

Une mission se rend à l'agence pour l'emploi afin d'évaluer un stage de 15 jours de formation de secrétaires portant sur le traitement de TEXTUEL VI. À partir d'une enquête statistique incluant plusieurs promotions, la mission a dressé le tableau suivant :

Nombre x de jours de stage	5	6	8	10	11	12	13	15
Nombre y d'erreurs de saisie par page	42	44	30	35	28	27	22	20

1. Au vu du tableau, le chef de mission présume que les deux variables sont liées par une relation de type affine $y = ax + b$. Afin de confirmer cette intuition :

a) déterminer la variable explicative et la variable expliquée ;

b) déterminer relation affine $y = a_0x + b_0$ déduite de la méthode des moindres carrés puis mesurer l'intensité de cette liaison ;

c) déterminer quelle proportion de la variance du nombre d'erreurs par page est expliquée par la relation affine et en déduire la valeur de la variance résiduelle.

2. Il souhaite estimer :

a) l'effet d'une journée supplémentaire de stage sur le nombre d'erreurs de saisie ;

b) le nombre d'erreurs de saisie que l'on peut attendre d'un nouveau stagiaire à son arrivée devant le poste informatique ;

c) le nombre de jours supplémentaires de stage nécessaires pour que les erreurs de saisie passent à 10 par page.

Exercice 2

Ce tableau présente l'évolution du chiffre d'affaires (CA) de la filiale bretonne de la firme américaine VIXON entre 2002 et 2006.

Années	2002	2003	2004	2005	2006
CA (10^6 en euros)	8	12	35	40	70

1. Établir par la méthode des moindres carrés la relation linéaire liant le chiffre d'affaires Y au temps t en considérant que $t = 1$ en 2002.

2. Vérifier la qualité de l'ajustement et déterminer les variances expliquée V_E et résiduelle V_R de Y.

3. Posant $t = 1$ en 2002 faire un ajustement logarithmique et comparer la qualité de l'ajustement avec celui de la question 1.

QCM. Soit la série statistique double (x_i, y_i) : (2, 7), (3, 5), (4, 3). Alors la covariance des variables $\boldsymbol{x}$ et $\boldsymbol{y}$ vaut : ① 41/3 ② – 4/3 ③ 86/3 ④ – 4 ⑤ aucune réponse ne convient.

3 ANALYSE INDICIAIRE DE SÉRIES TEMPORELLES

Les indices permettent de comparer à un instant donné des grandeurs exprimées au moyen d'unités différentes, de suivre leur évolution au cours du temps, enfin de décrire l'évolution temporelle de grandeurs simples (prix du fuel domestique, prix d'un service ou d'un bien vendu) ou composites (indice des prix, indice de production d'une entreprise, d'un secteur...). Disposant du relevé unitaire des prix et des quantités de biens et services quantifiables vendus à deux dates successives t_0 et t_1 on peut estimer aisément l'évolution du prix ou de la quantité vendue pour chacun de ces biens entre ces deux dates en recourant aux indices simples. Si au contraire, on veut évaluer l'évolution globale du prix d'un panier de biens et ser-vices, il faut recourir à des indices synthétiques.

Section 1 ■ Les indices simples
Section 2 ■ Les indices synthétiques

© Dunod. La photocopie non autorisée est un délit.

Section 1

LES INDICES SIMPLES

1 Indices simples de prix et de volume

– Le prix unitaire p d'un bien de type b, produit et vendu par une entreprise, dépend de l'instant t considéré : $p = p(t)$. Pour comparer le prix à la date t et le prix à la date de référence t_0, on considère le quotient

$$I(t/t_0) = p(t)/p(t_0) = \frac{(prix\ du\ bien\ à\ la\ date\ t)}{(prix\ du\ bien\ à\ la\ date\ t_0)}$$

appelé *indice de prix à la date t par rapport à la date de référence* t_0.

Lorsque $I(t/t_0) > 1$, il y a augmentation des prix.

– La quantité Q d'unités de biens b vendue par l'entreprise est une *fonction cumulative* du temps. Si l'on prend t_0 pour date de référence, $Q(t)$ désigne le nombre d'unités vendues depuis l'instant t_0 jusqu'à l'instant t (avec $t > t_0$).

Le temps étant équiréparti en périodes de même durée $(t_0,t_1),(t_1,t_2)$, ..., (t_{n-1},t_n), [où $t_i - t_{i-1} = h$ constante positive], notons q_i la quantité vendue durant la i-ème période : $q_i = Q(t_i) - Q(t_{i-1})$ (soit $i = 1$ pour (t_0,t_1), $i = 2$ pour (t_1,t_2),...).

L'évolution de la vente durant la i-ème période (t_{i-1},t_i) par rapport à la période de référence (t_0,t_1) ou première période s'exprime par le quotient :

$$I(i/1) = q_i/q_1 = \frac{(quantité\ produite\ ou\ consommée\ durant\ la\ i\text{-ème}\ période)}{(quantité\ produite\ ou\ consommée\ durant\ la\ première\ période)}$$

appelé indice simple de volume la i-ème période (t_{i-1},t_i) par rapport à la première période.

Si la période de référence est la période précédente, l'indice simple est

$$I(i/i-1) = q_i/q_{i-1}.$$

Exemple

Une entreprise spécialisée dans la fabrication de chaises en bois référencées « chaises arbois » souhaite connaître l'évolution des prix fixés et des quantités vendues depuis le début de l'année N. Le prix en début d'année est de 38 euros puis passe à 40 euros le 1-04-N, 45 le 1-07-N, etc.

Dates t	1–01–N	1–04–N	1–07–N	1–10–N	1–01–N+1
Quantité Q vendue depuis le 1/01/N	0	10 000	22 000	37 000	57 000
Prix p à la date t	38	40	45	46	48
Trimestre i		1	2	3	4
Ventes q_i du trimestre		10 000	12 000	15 000	20 000

L'indice de prix à la date 1-04-N par rapport à la date de référence 1-01-N est $I(\text{1-04-}N/\text{1-01-}N) = 40/38 = 1{,}0526$ soit 105,26 si l'on prend pour base l'indice 100 au 1-01-N.

Ainsi l'indice de l'évolution de la vente durant la 3-ème période par rapport à la période de référence 1 est $I(3/1) = q_3/q_1 = 15\,000/10\,000 = 1{,}5$ soit 150 si l'on prend pour base l'indice 100 au premier trimestre.

De même $I(3/2) = q_3/q_2 = 15\,000/12\,000 = 1{,}25$.

2 Propriétés des indices simples

D'une façon générale, le temps étant équiréparti en périodes de même durée $(t_0,t_1),(t_1,t_2),...,(t_{n-1},t_n)$, on appelle *indice élémentaire* de la grandeur G pour la n-ème période par rapport à la n_0-ème période de référence le rapport

$$I(n/n_0) = \frac{G_n}{G_{n_0}}$$

Repères

Cet indice possède les propriétés

i) de réversibilité : $I(n_0/n) = 1/[I(n/n_0)]$,

ii) de circularité: quelque soit la n'-ème période $I(n/n_0) = I(n/n') \times I(n'/n_0)$,

iii) d'enchaînement :

$$I(n/n_0) = I[n/(n-1)] \times I[(n-1)/(n-2)] \times I[(n-2)/(n-3)] \times \cdots \times I[(n_0+1)/n_0])$$

De ces propriétés il résulte que pour une période m quelconque on a :

$I(n/m) = I(n/1)/I(m/1)$. Autrement dit, partant de l'indice élémentaire $I(n/1)$ dont la période de référence est la première période on obtient un indice élémentaire $I(n/m)$ ayant pour période de référence la m-ème période en divisant l'indice $I(n/1)$ par $I(m/1)$.

Le *taux de croissance de G pour la période* (t_{n-1},t_n) est le quotient

$$\tau_n = I[n/n-1] - 1 = (G_n - G_{n-1})/G_{n-1}$$

Sachant que $G_n = G_{n-1}(1+\tau_n)$ pour $n > 1$ on a :

$$G_n = (1+\tau_2)(1+\tau_3) \times ... \times (1+\tau_n)G_1\,.$$

Si le taux de croissance est constant et égal à τ depuis la première période jusqu'à la n-ème période $(\tau_2 = \tau_3 = ... = \tau_n = \bar{\tau})$ on a $G_n = (1+\bar{\tau})^{n-1}G_1$.

$\bar{\tau}$ est alors appelé *taux moyen* de croissance de G pour la « longue » période (t_0,t_n).

© Dunod. La photocopie non autorisée est un délit.

Exemple

Partant du tableau précédent sur l'évolution des quantités q_n vendues par trimestre (soit $G_n = q_n$ avec $n = 1,2,3,4$), on détermine le taux de croissance τ_n associé à la n-ème période.

Trimestre n	1	2	3	4
Quantité vendue q_n	$10\,000 = q_1$	$12\,000 = q_2$	$15\,000 = q_3$	$20\,000 = q_4$
$\tau_n = (q_n - q_{n-1})/q_{n-1}$		0,20	0,25	0,33

Le taux moyen de croissance trimestriel $\bar{\tau}$ de la quantité Q sur les quatre trimestres (t_1,t_4) est défini par : $q_4 = (1+\bar{\tau})^{4-1} q_1$ soit $20\,000 = 10\,000(1+\bar{\tau})^{4-1}$ donc $\bar{\tau} = 26\ \%$.

Section 2

INDICES SYNTHÉTIQUES

En économie et en gestion, on agrége des ***quantités*** d'ensembles hétérogènes. Ainsi, pour évaluer la production globale d'un secteur agricole ou d'une entreprise diversifiée, on peut considérer le *vecteur* des quantités produites ou utiliser la monnaie comme étalon de référence et « faire la *somme monétaire* » des productions élémentaires.

1 Indices de volumes

Soit un panier de K types de biens et services quantifiables $b_1,b_2,....,b_K$ dont on souhaite étudier l'évolution au cours de périodes de même durée, *période* $1 = (t_0,t_1)$, *période* $2 = (t_1,t_2)$,..., *période* $n = (t_{n-1},t_n)$,...

Pour la période de base $n_0 = (t_{n_0-1},t_{n_0})$ on définit :

– les produits b_1, b_2,....,b_K

– leurs prix unitaires respectifs $p_{n_0}^{(1)}, p_{n_0}^{(2)}, ..., p_{n_0}^{(K)}$

– les quantités produites ou consommées respectivement : $q_{n_0}^{(1)}, q_{n_0}^{(2)}, ..., q_{n_0}^{(K)}$

Pour la n-ème période $= (t_{n-1},t_n)$ on recense :

– les prix unitaires des biens respectivement notés $p_n^{(1)}, p_n^{(2)} ..., p_n^{(K)}$

– les quantités produites respectivement notées $q_n^{(1)}, q_n^{(2)}, ..., q_n^{(K)}$

Les indices de volume de Laspeyres et de Paasche de la transaction sur la période $n = (t_{n-1},t_n)$ par rapport à la période de référence $n_0 = (t_{n_0-1},t_{n_0})$ sont ainsi définis :

– l'indice de volume de Laspeyres

$$L_q(n/n_0) = \sum_i q_n^{(i)} p_{n_0}^{(i)} / \sum_i q_{n_0}^{(i)} p_{n_0}^{(i)}$$

$$= \frac{(\textit{quantités de la période } n \textit{ valorisées aux prix de la période } n_0)}{(\textit{quantités de la période } n_0 \textit{ valorisées aux prix de la période } n_0)}$$

– l'indice de volume de Paasche

$$P_q(n/n_0) = \sum_i q_n^{(i)} p_n^{(i)} / \sum_i q_{n_0}^{(i)} p_n^{(i)}$$

$$= \frac{(\textit{quantités de la période } n \textit{ valorisées aux prix de la période } n)}{(\textit{quantités de la période } n_0 \textit{ valorisées aux prix de la période } n)}$$

Exemple

Le tableau ci-dessous donne les prix unitaires et les quantités consommées de trois articles en 2004, 2005 et 2006. À partir de ce tableau, on calcule pour la période 2006 par référence à la période 2004, les indices de volume de Laspeyres $L_q(2006/2004)$ et de Paasche $P_q(2006/2004)$.

	Prix			Quantités		
	$p_{n0}^{(i)}$		$p_n^{(i)}$	$q_{n0}^{(i)}$		$q_n^{(i)}$
Articles \ Années	2004	2005	2006	2004	2005	2006
A	30	40	45	100	120	150
B	5	6	6	20	30	50
C	40	50	45	10	20	30

– Indice de volume de Laspeyres $L_q(n/n_0) = \sum_i q_n^{(i)} p_{n_0}^{(i)} / \sum_i q_{n_0}^{(i)} p_{n_0}^{(i)}$ avec $n_0 = 2004$ et $n = 2006$:

$$L_q(2006/2004)$$

$$= \frac{(\textit{quantités de l' année 2006 valorisées aux prix de l'année 2004})}{(\textit{quantités de l'année 2004 valorisées aux prix de l'année 2004})}$$

$$= (150 \times 30 + 50 \times 5 + 30 \times 40)/(100 \times 30 + 20 \times 5 + 10 \times 40) = 1{,}7$$

– Indice de volume de Paasche $P_q(n/n_0) = \sum_i q_n^{(i)} p_n^{(i)} / \sum_i q_{n_0}^{(i)} p_n^{(i)}$ avec $n_0 = 2004$ et $n = 2006$:

$$P_q(2006/2004) = \frac{(\textit{quantités de l'année 2006 valorisées aux prix de l'année 2006})}{(\textit{quantités de l'année 2004 valorisées aux prix de l'année 2006})}$$

$$= (150 \times 45 + 50 \times 6 + 30 \times 45)/(100 \times 45 + 20 \times 6 + 10 \times 45) = 1{,}656$$

– **l'indice de volume de Fisher** est la moyenne géométrique des indices de Laspeyres et Paasche : $F_q(n/n_0) = \sqrt{L_q(n/n_0) \times P_q(n/n_0)}$.

2 Indices de prix

Pour connaître l'évolution des prix d'un catalogue donné de biens b_1 , b_2 ,..., b_K produits ou consommés en quantités respectives q_1 , q_2 ,..., q_K , on compare la valeur de ces biens durant la période n considérée $(t_n - 1, t_n)$ à la valeur de ces mêmes biens durant la période n_0 de référence (t_{n_0-1}, t_{n_0}). Les indices de prix de Laspeyres et de Paasche de la transaction sur la période $n = (t_{n-1}, t_n)$ par rapport à la période de référence $n_0 = (t_{n_0-1}, t_{n_0})$ sont ainsi définis :

© Dunod. La photocopie non autorisée est un délit.

– ***l'indice prix de Laspeyres***

$$L_p(n/n_0) = \sum_i p_n^{(i)} q_{n_0}^{(i)} / \sum_i p_{n_0}^{(i)} q_{n_0}^{(i)}$$

$$= \frac{(prix\ \ de\ \ la\ \ période\ n\ \ pondérés\ par\ les\ quantités\ \ considérées\ de\ la\ période\ n_0)}{(prix\ \ de\ \ la\ \ période\ n_0\ pondérés\ par\ les\ quantités\ \ considérées\ de\ la\ période\ n_0)}$$

– ***l'indice des prix de Paasche***

$$P_p(n/n_0) = \sum_i p_n^{(i)} q_n^{(i)} / \sum_i p_{n_0}^{(i)} q_n^{(i)}$$

$$= \frac{(prix\ \ de\ \ la\ \ période\ n\ \ pondérés\ par\ les\ quantités\ \ considérées\ de\ la\ période\ n)}{(prix\ \ de\ \ la\ \ période\ n_0\ pondérés\ par\ les\ quantités\ \ considérées\ de\ la\ période\ n)}$$

Exemple

Reprenons l'exemple précédent.

– Indice des prix de Laspeyres $L_p(n/n_0) = \sum_i p_n^{(i)} q_{n_0}^{(i)} / \sum_i p_{n_0}^{(i)} q_{n_0}^{(i)}$ avec $n_0 = 2004$ et $n = 2005$: $L_p(2005/2004)$
$= (40 \times 100 + 6 \times 20 + 50 \times 10)/(30 \times 100 + 5 \times 20 + 40 \times 10) = 1{,}32$

– Indice des prix de Paasche $P_p(n/n_0) = \sum_i p_n^{(i)} q_n^{(i)} / \sum_i p_{n_0}^{(i)} q_n^{(i)}$ avec $n_0 = 2004$ et $n = 2005$: $P_p(2005/2004)$
$= (40 \times 120 + 6 \times 30 + 50 \times 20)/(30 \times 120 + 5 \times 30 + 40 \times 20) = 1{,}314$

– **l'indice de Fisher des prix** est la moyenne géométrique des indices de Laspeyres et Paasche :

$$F_p(n/n_0) = \sqrt{L_p(n/n_0) \times P_p(n/n_0)}$$

Exercices (*cf.* corrections page 338)

Exercice 1

Une entreprise commercialise 4 types de magnétoscopes. Les prix pratiqués en euros et les chiffres d'affaires réalisés (en millions d'euros) en 2004 et 2006 sont :

	Prix		Chiffre d'affaires	
	Année 2004	**Année 2006**	**Année 2004**	**Année 2006**
PX1	800	700	20	21
PX2	720	680	72	34
VTS	650	650	32,5	48,75
VTR	360	290	54	58

1. Calculer les indices de prix de Paasche et de Laspeyres en prenant l'année 2004 comme base 100.

2. Calculer les indices de volume de Paasche et de Laspeyres en prenant l'année 2004 comme base 100.

Exercice 2

Entre janvier 2001 et avril 2002, le cours de l'euro en dollar ($) sur le marché des changes a connu l'évolution suivante :

	janvier 2006	juillet 2006	octobre 2006	avril 2007
euros en $	1,1821	1,2792	1,2737	1,3369

1. Déterminer pour cette période :

– l'indice de base 100 en juillet 2006 donnant l'évolution du cours du dollar ;

– le taux de variation *global* ; le taux de variation *mensuel moyen* $\bar{\tau}_m$.

2. Déterminer le taux mensuel moyen $\bar{\tau}^*$ de variation du cours du dollar entre juillet 2006 et octobre 2006.

QCM 1. Figurent ci-dessous les prix unitaires p et les quantités consommées q de deux articles A et B. Année 2004 : $p_A = 30$, $q_A = 100$; $p_B = 35$, $q_B = 200$; Année 2006 : $p_A = 40$, $q_A = 150$; $p_B = 50$, $q_B = 300$.

L'indice de prix de Lapeyres $L_p(2006/2004)$ évalué en % est égal à :

① 1,1 ② 1,2 ③ 1,3 ④ 1,4 ⑤ aucune réponse proposée ne convient.

QCM 2. L'évolution en quantité vendue d'un artcile au cours de 3 années consécutives a été la suivante : 100, 110, 121. Le taux moyen de croissance en % pour cette période est de :

① 7 ; ② 11 ; ③ 21 ; ④ 31/3 ; ⑤ aucune réponse proposée ne convient.

© Dunod. La photocopie non autorisée est un délit.

4 ANALYSE DES SÉRIES CHRONOLOGIQUES

Une série chronologique est une suite d'observations au cours du temps. En matière économique et en gestion, l'étude des chroniques, c'est-à-dire des séries statistiques temporelles, permet de mettre en évidence les composantes (trend, mouvement saisonnier...) qui régissent l'évolution d'une variable statistique au cours du temps. Selon le modèle retenu, les composantes sont ou non indépendantes. De nombreuses techniques empiriques telles que la méthode de la moyenne mobile ou la méthode des pourcentages moyens permettent de désaisonnaliser une série, autrement dit de la corriger de ses variations saisonnières.

Section 1 ■ Les composantes d'une série chronologique
Section 2 ■ Les modèles de décomposition
Section 3 ■ Désaisonnalisation par des méthodes empiriques
Section 4 ■ Désaisonnalisation par la méthode analytique de Buys-Ballot
Section 5 ■ Méthode de lissage exponentiel
Section 6 ■ Traitements avec SPSS et Excel

Section 1

LES COMPOSANTES D'UNE SÉRIE CHRONOLOGIQUE

Les n valeurs observées x_1 , x_2 , ... , x_n d'une variable statistique $\boldsymbol{x}(t)$ aux dates t_1 , t_2 , ... , t_n généralement séparées par des durées égales ou sensiblement égales (années, mois, jours), peuvent être représentées graphiquement dans un repère cartésien par n points $M_1 = (t_1 , x_1), M_2 = (t_2 , x_2) , ... , M_n = (t_n , x_n)$. Ces points peuvent être ajustés par une courbe d'équation $x = \varphi(t)$ où φ est une fonction continue et $x_i = \varphi(t_i)$.

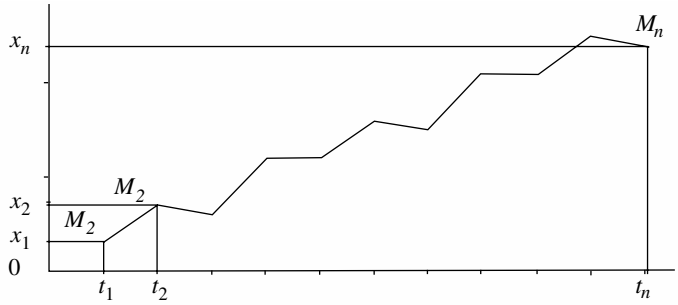


Figure 4.1 – *Une chronique*

Lorsque les observations sont rapprochées (relevés trimestriels, par exemple) et s'étendent sur une période assez longue (une ou deux décennies, par exemple), on peut isoler des « forces » dont les effets sont perceptibles à différents horizons temporels.

Les valeurs observées au cours du temps, $\boldsymbol{x}(t)$, peuvent être considérées comme la résultante (*cf.* figure 4.1) :

– ***d'un mouvement séculaire*** $\Omega(t)$ ou allure générale du phénomène sur une très longue période (en économie : une ou plusieurs décennies) ;

– ***d'un mouvement cyclique*** $\boldsymbol{C(t)}$ de grande amplitude qui traduit des oscillations approximativement périodiques autour de la tendance séculaire (en économie on distingue les cycles longs de Kondratiev d'une durée de 50 à 60 ans, de Kyznets de 20 à 25 ans, les cycles de Juglar d'une durée de 7 à 11 ans et les cycles de Kitchin de 3 ou 4 ans) ;

– ***du mouvement saisonnier*** $\boldsymbol{S(t)}$, mouvement de période fréquemment annuelle résultant de forces se reproduisant d'année en année (tel la baisse de la consommation d'électricité durant les mois d'été) ; en gestion la période peut également être hebdomadaire, journalière (arrivée d'un grand nombre de clients dans une grande surface entre 12 et 14 heures) ;

– ***de variations accidentelles*** $\boldsymbol{A(t)}$ résultant de causes habituellement imprévisibles (une consommation plus importante d'énergie suite à un hiver rigoureux, etc.)

© Dunod. La photocopie non autorisée est un délit.

Généralement, les mouvements séculaires Ω et cycliques ***C*** sont confondus (lorsque l'on ne dispose pas d'une série assez longue ou que l'existence du cycle paraisse contestable). On parle alors de mouvement extra-saisonnier ou de *trend* que l'on désigne par la lettre ***T***.

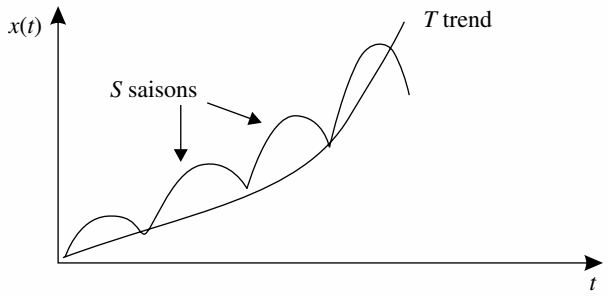


Figure 4.2 – *Composantes des variations temporelles*

Section 2

LES MODÈLES DE DÉCOMPOSITION

L'étude des chroniques consiste à estimer chaque composante à l'aide des données passées et d'en induire des prévisions.

Si les composantes sont indépendantes entre elles, on peut exprimer $\boldsymbol{x}(t)$ par un schéma additif :

$$\boldsymbol{x}(t) = T(t) + S(t) + A(t)$$

Si les composantes sont étroitement liées entre elles, $\boldsymbol{x}(t)$ peut apparaître comme une résultante multiplicative :

$$\boldsymbol{x}(t) = T(t) \times (S(t) \times (1 + A(t))$$

Le choix de l'un ou l'autre des modèles précités se fait par examen des graphes des courbes annuelles (ou selon, de périodicité autre) ou par étude du graphe de l'ensemble de la chronique.

Par examen des courbes annuelles on retiendra un modèle de décomposition de x

- de type additif lorsque les courbes annuelles superposées sont approximativement parallèles sur un papier à échelles arithmétiques ;
- de type multiplicatif si le même parallélisme est observé sur un papier dont l'abscisse est à l'échelle arithmétique et l'ordonnée à l'échelle logarithmique.

Utilisant le schéma de l'ensemble la chronique (*cf.* figure 4.2), il convient de relier les maxima annuels par une courbe C_1 et les minima par une courbe C_2 afin de déterminer le type de modèle, le plus approprié. Le modèle à retenir est de type :

– *additif* si les deux courbes C_1 et C_2 sont sensiblement parallèles,

– *multiplicatif* si les deux courbes C_1 et C_2 ne sont pas parallèles.

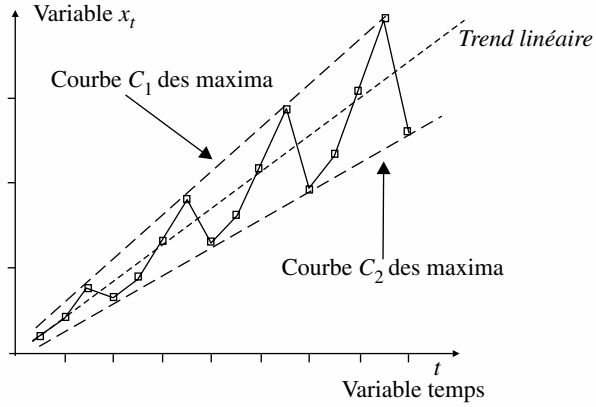


Figure 4.3 – *Schéma additif ou multiplicatif*

Remarque. Le modèle multiplicatif peut être aisément transformé en un modèle de type additif à l'aide d'un changement de variables. En effet, si on note $\boldsymbol{x}^*(t) = \ln(\boldsymbol{x}(t)), T^*(t) = \ln(T(t)), S^*(t) = \ln(S(t)), A^*(t) = \ln(1 + A(t))$ on a : $\boldsymbol{x}^*(t) = T^*(t) + S^*(t) + A^*(t)$.

Section 3

DÉTERMINATIONS DES COMPOSANTES TEMPORELLES PAR MÉTHODES EMPIRIQUES

Considérons n relevés périodiques successifs (hebdomadaires, mensuels, trimestriels...), notés de x_1 à x_n et supposons que le mouvement saisonnier comprend ν périodes successives d'observation. Alors la série peut être réécrite en distinguant le premier cycle d'observations comprenant ν périodes $(x_1, x_2, \ldots, x_\nu)$, du deuxième cycle d'observations comprenant les ν périodes $(x_{\nu+1}, x_{\nu+2}, \ldots, x_{2\nu})$...

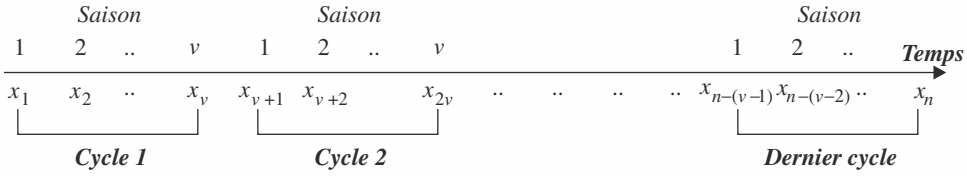


Figure 4.4

Ainsi dans le cas de mouvement saisonnier annuel avec observations trimestrielles : $\nu = 4$, le premier cycle est composé des relevés trimestriels (x_1, x_2, x_3, x_4), le second des relevés trimestriels (x_5, x_6, x_7, x_8),...

La h-ième observation x_h de la chronique n'est autre que la j-ème saison du i-ème cycle d'observation où

© Dunod. La photocopie non autorisée est un délit.

– $j = r(h)$ est le reste de la division de h par ν si h n'est pas un multiple de ν et dans le cas contraire on a $j = \nu$,

– $i = (h - j)/\nu + 1$ ou de façon équivalente $h = \nu \times (i - 1) + j$

Ainsi, dans le cas de mouvement saisonnier annuel avec fluctuations trimestrielles ($\nu = 4$), l'observation x_7 est la troisième saison ($j = 3$) du deuxième cycle annuel ($i = 2$). En effet, $7 = \nu \times 1 + 3 \Longrightarrow j = 3$ et $(i - 1) = 1$.

1 Méthode de la moyenne mobile

La méthode présentée ci-dessous dans le cadre d'un ***modèle de décomposition de type additif*** suppose que le mouvement extra-saisonnier est une fonction quelconque du temps, que le mouvement saisonnier est rigoureusement périodique et que le mouvement accidentel est de faible amplitude et de moyenne nulle.

L'observation x_h relative à la h-ième période se décompose alors sous la forme :

$$x_h = T_h + S_j + A_h$$

où T_h désigne le trend, S_j le facteur saisonnier, A_h le facteur accidentel.

Tableau 4.1 – *Série décomposée suivant l'année et la saison*

Saison $j = r(h) \rightarrow$ Cycle ou année $i\downarrow$	**1**	**2**	**...**	ν
1	$x_1 = T_1 + S_1 + A_1$	$x_2 = T_2 + S_2 + A_2$	**...**	$x_\nu = T_\nu + S_\nu + A_\nu$
2	$x_{\nu+1} = T_{\nu+1} + S_1 + A_{\nu+1}$	$x_{\nu+2} = T_{\nu+2} + S_2 + A_{\nu+2}$	**...**	$x_{2\nu} = T_{2\nu} + S_\nu + A_{2\nu}$
3	$x_{2\nu+1} = T_{2\nu+1} + S_1 + A_{2\nu+1}$	$x_{2\nu+2} = T_{2\nu+2} + S_2 + A_{2\nu+2}$		
...	**...**	**...**	**...**	**...**

Détermination du trend

La détermination du trend est différente selon que le mouvement saisonnier comprend un nombre de périodes d'observations ν impair ($\nu = 2p + 1, p$ entier) ou pair ($\nu = 2p$). (Par exemple dans le cas de fluctuations trimestrielles, $\nu = 4 = 2 \times p$ avec $p = 2$).

À chaque chronique $x(t)$, on peut associer sa moyenne mobile d'ordre p :

– lorsque le mouvement saisonnier comprend un *nombre impair* $\nu = (2p + 1)$ de saisons, on attribue pour trend

à la $(p + 1)$-ème période, la valeur $T_{p+1} = (x_1 + x_2 + ... + x_\nu)/\nu$,

à la $(p + 2)$-ème période, la valeur $T_{p+2} = (x_2 + x_3 + ... + x_{\nu+1})/\nu$,

à la $(p + 3)$-ème période, la valeur $T_{p+3} = (x_3 + x_4 + ... + x_{\nu+2})/\nu$,...

– lorsque le mouvement saisonnier a un *nombre pair* $\nu = 2p$ de saisons dans l'année on attribue pour trend

à la période $(p+1)$, la valeur $T_{p+1} = (0{,}5x_1 + x_2 + ... + x\nu + 0{,}5x_{\nu+1})/\nu$,
à la période $(p+2)$, la valeur $T_{p+2} = (0{,}5x_2 + x_3 + ... + x_{\nu+1} + 0{,}5x_{\nu+2})/\nu$,
à la période $(p+3)$, la valeur $T_{p+3} = (0{,}5x_3 + x_4 + ... + x_{\nu+2} + 0{,}5x_{\nu+3})/\nu$,...

Application. Une entreprise a relevé pendant quatre ans ses ventes trimestrielles ($\nu = 4$ et donc $p = 2$) et souhaite dégager le trend par un procédé de lissage.

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2004	102	100	104	108
2005	110	106	113	115
2006	117	115	120	122
2007	128	123	127	131
2008	134	130		

Le trend doit être calculé sur une période de quatre trimestres afin de lisser les mouvements saisonniers. Il est à noter que l'on ne dispose pas de mesure du trend pour les p premières périodes d'observations. À la $(p+1)$-ème période, soit ici la troisième période, on attribue au trend la valeur

$$T_{p+1} = T_3 = (0{,}5x_1 + x_2 + x_3 + x_4 + 0{,}5x_5)/4 = 418/4 = 104{,}5$$

Prenant en compte la première période d'observations d'une durée de quatre trimestres, il est naturel de vouloir attribuer la valeur moyenne $(x_1 + x_2 + ... + x_4)/4$ à la période médiane. Celle-ci n'existant pas puisque 4 est un nombre pair, on attribue la moyenne des deux moyennes : $(x_1 + x_2 + ... + x_4)/4$ et $(x_2 + x_3 + ... + x_5)/4$ au trimestre relatif à la $3^{\text{ème}}$ observation qui est la période médiane des cinq premières observations.

De même, à la $(p+2)$-ème période soit ici la quatrième période on attribue au trend la valeur

$$T_{p+2} = T_4 = (0{,}5x_2 + x_3 + x_4 + x_5 + 0{,}5x_6)/4 = 425/4 = 106{,}25\ldots$$

et l'on dresse le tableau suivant :

Tableau du trend T_h

Saisons *j* / Années *i*	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2004			104,50	106,25
2005	108,13	110,13	111,88	113,88
2006	115,88	117,63	119,88	122,25
2007	124,13	126,13	128,00	129,63

Détermination de la composante saisonnière S_j

Il convient au préalable

– de déterminer l'écart entre la valeur d'observation x_h et la valeur de son trend T_h précédemment calculé, soit $S'_h = x_h - T_h$ pour $h = p+1,\ p+2,..., n-p$;

© Dunod. La photocopie non autorisée est un délit.

– de calculer pour chacune des ν saisons (et donc pour chaque $j = 1,2,...,\nu$) la valeur moyenne $\overline{S}'_j$ des écarts constatés S'_h relatifs à cette saison :

Afin de respecter l'hypothèse selon laquelle le mouvement saisonnier est rigoureusement périodique, on soustrait à chaque $\bar{S}'_j$ la moyenne générale $\bar{S}' = (\bar{S}'_1 + \ldots + \bar{S}'_{p+1} + ... + \bar{S}'_\nu)/\nu$ de ces ν moyennes. Le coefficient saisonnier $S_j = \bar{S}'_j - \bar{S}'$.

Désaisonnalisation

Pour désaisonnaliser la série, il suffit alors de retrancher à chaque relevé x_h sa composante saisonnière afin de lui substituer soit $x'_h = x_h - S_j$, que l'on décompose sous la forme $x'_h = T_h + A_h$. Donc l'aléas $A_h = x_h - T_h - S_j$.

Application (suite). Pour terminer l'analyse de l'exemple précédent où ont été obtenues par procédé de lissage les valeurs T_h du trend, on doit déterminer les composantes saisonnières S_j. Au préalable, il faut calculer des différences $S'_h = (x_h - T_h)$, (ainsi $S'_3 = x_3 - T_3 = 105 - 104{,}5 = -0{,}5...$), en déduire la moyenne par saison de ces écarts, puis réaliser la moyenne générale $\bar{S}'$ de ces quatre moyennes $\bar{S}'_j$.

	Écarts S'_h			
Année ↓ Saison $j = r(h) \rightarrow$	**1**	**2**	**3**	**4**
2004			– 0,50	1,75
2005	1,88	– 4,13	1,13	1,13
2006	1,13	– 2,63	0,13	– 0,25
2007	3,88	– 3,13	– 1,00	1,38
2008				
$\bar{S}'_j$ **moyenne par saison**	2,292	– 3,292	– 0,063	1,000
$S_j = \bar{S}'_j - \bar{S}' = \bar{S}'_j - 0{,}0156$	2,307	– 3,276	– 0,047	1,016

Connaissant les valeurs x_{ij} et S_j, on détermine la série corrigée des variations saisonnières $x'_{ij} = x_{ij} - S_j$. Ainsi $x'_{11} = x_{11} - S_1 = 102 - 2{,}307...$

Tableau de la série désaisonnalisée x'_{ij}

	Trimestre 1	**Trimestre 2**	**Trimestre 3**	**Trimestre 4**
2004	99,693	103,276	104,047	106,984
2005	107,693	109,276	113,047	113,984
2006	114,693	118,276	120,047	120,984
2007	125,693	126,276	127,047	129,984
2008	131,693	133,276		

Puis on peut déterminer la suite des valeurs des aléas : $A_h = x_h - T_h - S_j$.

2 Méthode des pourcentages au trend

La méthode présentée ci-dessous entre dans le cadre d'un ***modèle de décomposition de type multiplicatif***. Le mouvement extra-saisonnier est une fonction quelconque du temps et le mouvement saisonnier est d'amplitude variable.
On détermine tout d'abord le trend sous la forme d'une fonction du temps $T = f(t)$ où $t = h = \nu(i-1) + j$. En général, on régresse la variable $x(h)$ sur h selon la méthode des moindres carrés, les valeurs du trend T correspondent alors aux valeurs estimées de x par l'ajustement considéré. On dispose donc des valeurs T_h ou de façon équivalente T_{ij}, valeur de T pour chaque saison j de chaque année i.

Coefficients saisonniers. On détermine les coefficients saisonniers en procédant successivement comme suit :

– on calcule les rapports S'_{ij} entre valeurs observées x_{ij} et les valeurs ajustées T_{ij} :

$$S'_{ij} = 100(x_{ij}/T_{ij}) \text{ (en base 100)}$$

– on fait la moyenne par saison de ces coefficients S'_{ij} soit

$\bar{S}'_{\bullet j} = (S'_{1j} + S'_{2j} + \ldots + S'_{kj})/k$ si il y a k observations pour la saison j ;

– on attribue finalement à chaque saison j le coefficient saisonnier

$S_j = \bar{S}'_{\bullet j} - (\bar{S}' - 100)$ où $\bar{S}' = (\bar{S}'_{\bullet 1} + \bar{S}'_{\bullet 2} + \ldots + \bar{S}'_{\bullet \nu})/\nu$ (*cf.* exercice 3).

Section 4

DÉSAISONNALISATION PAR LA MÉTHODE ANALYTIQUE DE BUYS-BALLOT

Considérons n relevés périodiques successifs, le mouvement saisonnier comprenant ν périodes.

1 Modèle additif

Dans un modèle additif, l'observation x_h relative à la h-ème période se décompose sous la forme : $x_h = T_h + S_j + A_h$ où rappelons-le, le nombre j caractérise la saison correspondant à la h-ième observation.

Si on considère que :

– le mouvement conjoncturel T_h est un trend linéaire : $T_h = \alpha\, h + \beta$;

– le mouvement saisonnier est rigoureusement périodique : $S_j = \gamma_j$ (où $\gamma_1, \gamma_2, \ldots$ sont des constantes) ;

© Dunod. La photocopie non autorisée est un délit.

– le mouvement accidentel A_h est un écart en moyenne nul dont les valeurs successives sont indépendantes ; on a

[1] $x_h = \alpha \times h + \beta + \gamma_j + A_h$ *ou de façon équivalente*

[1'] $x_h = x_{ij} = \alpha \times [(\nu \times (i-1) + j] + \beta + \gamma_j + A_{ij}$ *puisque* $h = \nu \times (i-1) + j$

Indexant la série suivant les deux dimensions numéro *i* du cycle et numéro *j* de la saison, on obtient la table de Buys-Ballot.

Tableau 4.2 – *de Buys-Ballot*

Valeurs de *j* (la saison) → Valeurs de *i* ↓ l'année ou cycle	**1**	**2**	**...**	ν	**Total en ligne**	**Moyenne annuelle $\bar{x}_{i\bullet}$**	$i \times \bar{x}_{i\bullet}$
1	x_{11}	x_{12}		$x_{1\nu}$	$\Sigma_j x_{1j}$	$(\Sigma_j x_{1j})/\nu$	
2	x_{21}	x_{22}		$x_{2\nu}$	$\Sigma_j x_{2j}$	$(\Sigma_j x_{2j})/\nu$	
...	...	...		...	...		
k	x_{k1}	x_{k2}		$x_{k\nu}$	$\Sigma_j x_{kj}$	$(\Sigma_j x_{kj})/\nu$	
Total en colonne	$\Sigma_j x_{i1}$	$\Sigma_j x_{i2}$		$\Sigma_j x_{i\nu}$	$\Sigma_{ij} x_{ij}$		$\Sigma i \times \bar{x}_{i\bullet}$
Moyenne pour la saison j : $\bar{x}_{\bullet j}$	$(\Sigma_i x_{i1})/k$	$(\Sigma_i x_{i2})/k$		$(\Sigma_i x_{i\nu})/k$			
Coefficients saisonniers γ_i							

La méthode d'estimation étant la méthode des moindres carrés ordinaires, on cherche les valeurs α^* et β_j^* ($\beta_j^* = \beta^* + \gamma_j^*$ où $j = 1,2,...,\nu$) qui minimisent la fonction de $(\nu + 1)$ variables $\mathbf{e}(\alpha, \beta_j) = \sum_{i=1}^{k} \sum_{j=1}^{\nu} (x_{ij} - \alpha \times (\nu \times (i-1) + j) - \beta_j)^2$.

$\bar{x}$ désignant la moyenne générale de la série : $\bar{x} = \sum_{i=1}^{k} \sum_{j=1}^{\nu} x_{ij}/(n \times \nu)$, on obtient :

$$\alpha^* = \left[\left(\sum_{i=1}^{k} i \times \bar{x}_{i\bullet} \right) - \bar{x} \times k \times (k+1)/2 \right] / k \times (k^2 - 1);$$

$$\beta^* = \bar{x} - \alpha^* \times (k \times \nu + 1)/2] \; ; \; \gamma_j^* = \bar{x}_{\bullet j} - \bar{x} - \alpha^* \times [j - (\nu + 1)/2]$$

Application. Considérons la série mensuelle des ventes d'un rayon de magasin

Année	**Mois**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total en ligne**	**Moyenne par an $\bar{x}_{i\bullet}$**	$i \times \bar{x}_{i\bullet}$
2001	1	353	332	318	318	364	399	368	334	335	375	419	504	4419	368,3	368,3
2002	2	372	352	338	337	383	419	387	353	354	394	438	523	4650	387,5	775
2003	3	392	371	357	356	403	438	406	373	374	413	458	543	4884	407	1221
2004	4	411	391	377	376	422	457	426	392	393	433	477	562	5117	426,4	1705,7
2005	5	431	410	396	395	441	477	445	411	412	452	496	582	5348	445,7	2228,3
2006	6	450	429	415	415	461	496	465	431	432	471	516	601	5582	465,2	2791
Total colonne		2409	2285	2201	2197	2474	2686	2497	2294	2300	2538	2804	3315	30000		9089,3
Moyenne par mois $\bar{x}_{\bullet j}$		401,5	380,8	366,8	366,2	412,3	447,7	416,2	382,3	383,3	423	467,3	552,5			
γ_i		– 6,3	– 28,6	– 44,2	– 46,5	– 1,9	31,8	– 1,3	– 36,8	– 37,4	0,7	43,4	126,9			

La moyenne $\bar{x}$ sur l'ensemble des relevés est égale à 416,666 : $\bar{x} = 416{,}666$.
La moyenne sur la première année $(i = 1)$ est $\bar{x}_{1\bullet} = 368{,}25$, le produit $i \times \bar{x}_{i\bullet} = 1 \times \bar{x}_{1\bullet} = 368{,}25$, etc.
On obtient $\sum_{i=1}^{6} i \times \bar{x}_{i\bullet} = 368{,}25 + 775 + ... + 2\,791 = 9\,089{,}3$. On en déduit que

$$\alpha^* = (9\,089{,}3 - 416{,}666 \times 3 \times 7)/(6 \times 35) = 1{,}6158,$$

$$\beta^* = 416{,}666 - 1{,}6158 \times (6 \times 12 + 1)/2 = 357{,}689 \text{ et}$$

$$\gamma_j^* = \bar{x}_{\bullet j} - 416{,}66 - 1{,}6158 \times (j - 13/2) \text{ pour } j = 1{,}2...{,}12.$$

2 Modèle multiplicatif à trend exponentiel

Le modèle à composantes multiplicatives et à trend exponentiel peut s'exprimer par la relation : $\boldsymbol{x}_h = x_1 \times (1 + \tau)^{h-1} \times (1 + S_j) \times (1 + A_h)$.
Passant aux logarithmes et posant $\boldsymbol{x}_h^* = \ln(\boldsymbol{x}_h)$, $\alpha^* = \ln(1 + \tau)$, $\beta^* = \ln(x_0)$, $\gamma_j^* = \ln(1 + S_j)$, $A_h^* = \ln(1 + A_h)$, on a $x_h^* = \alpha^* \times h + \beta^* + \gamma_j^* + A_h^*$.
L'estimation des paramètres α^*, β^* et γ_j^* se fait alors conformément à la méthode employée pour les modèles additifs.

Section **5**

MÉTHODE DE LISSAGE EXPONENTIEL

1 Lissage simple pour les séries sans trend

Le procédé s'inscrit dans un modèle prévisionnel. Connaissant les valeurs $x_i = \boldsymbol{x}(t_i)$ prises par une variable statistique $\boldsymbol{x}(t)$ aux dates équiréparties $t_0, t_1, \ldots, t_n$ ($t_h = \theta \times h \quad \forall\, h \in \mathbb{N}$, θ constante positive) le modèle a pour objectif d'obtenir à la date t_n une estimation de la valeur $\boldsymbol{x}(t_{n+1})$ qui sera observée à la date t_{n+1}. Cette estimation prévisionnelle notée $\boldsymbol{y}(t_{n+1})$ est réalisée à la date t_n à partir des valeurs connues $x_1, x_2, \ldots, x_n$.

Le modèle de lissage exponentiel prend en compte l'écart entre la valeur observée $x(t_n)$ et la valeur $\boldsymbol{y}(t_n)$ qui avait été prévue :

$\boldsymbol{y}(t_{n+1}) = \boldsymbol{y}(t_n) + \alpha(\boldsymbol{x}(t_n) - \boldsymbol{y}(t_n))$ avec α constante vérifiant $0 < \alpha < 1$.

ou de façon équivalente : « $\boldsymbol{y}(t_{n+1}) = \alpha \boldsymbol{x}(t_n) + (1 - \alpha)\boldsymbol{y}(t_n)$ » qui est la moyenne pondérée entre la valeur réellement observée et la valeur estimée. Connaissant les valeurs $\boldsymbol{x}(t_i)$ antérieures à la date t_{n+1} et ayant choisi une valeur pour α on connaît immédiatement la valeur prévisionnelle $\boldsymbol{y}(t_{n+1})$:

$$\boldsymbol{y}(t_{n+1}) = \alpha \boldsymbol{x}(t_n) + \alpha\,(1 - \alpha)\boldsymbol{x}(t_{n-1}) + \alpha(1 - \alpha)^2 \boldsymbol{x}(t_{n-2}) + \ldots + \alpha(1 - \alpha)^n \boldsymbol{x}(t_0)$$

© Dunod. La photocopie non autorisée est un délit.

Choix de la valeur α. On choisit la valeur α qui minimise la variance des écarts et donc qui minimise $\sum_{i=1}^{n}(\boldsymbol{y}(t_i) - \boldsymbol{x}(t_i))^2$. En pratique on peut donner à α différentes valeurs autour de 0,30, par exemple 0,15, 0,20, 0,25, 0,30, 0,35, 0,40, 0,45 puis choisir la valeur associée à la plus petite des sommes obtenues.

2 Lissage de Holt-Winters pour modèle avec trend et saisonnalité

Dans cette méthode, chaque nouvelle observation provoque une mise à jour du trend ainsi que des coefficients du mouvement saisonnier de période p. Connaissant les valeurs $x_i = \boldsymbol{x}(t_i)$ pour les dates $t_0, t_1, \ldots, t_n$, la valeur de $\boldsymbol{x}(t_{n+h})$ est estimée dans un modèle additif par $\boldsymbol{y}(t_{n+h}) = a_0(t_n) \times h + b_0(t_n) + s^*(t_{n+h})$ où $[a_0(t_n) \times h + b_0(t_n)]$ et $s^*(t_{n+h})$ sont respectivement les estimations du trend et du mouvement saisonnier réalisées à partir des données disponibles à la date t_n. Les valeurs estimées $a_0(t_n), b_0(t_n)$ et $s^*(t_{n+h})$ s'obtiennent par le procédé de lissage ci-après où $\boldsymbol{x}(t_i), \boldsymbol{y}(t_i), a_0(t_i), b_0(t_i)$ et $s^*(t_i)$ sont plus simplement notés x_i, y_i, a_i, b_i, s_i et où α, β et γ sont des constantes dont les valeurs sont comprises entre 0 et 1.

Pour un modèle additif : $y_{n+h} = a_n \times h + b_n + s_{n+h}$ où

$b_i = \alpha(x_i - s_{i-p}) + (1-\alpha)(b_{i-1} + a_{i-1})$; $s_i = \beta(x_i - b_i) + (1-\beta)s_{i-p}$;

$a_i = \gamma(b_i - b_{i-1}) + (1-\gamma)a_{i-1}$ pour $i > p$

avec les valeurs initiales : $a_p = \bar{x}_2 - \bar{x}_1$ où $\bar{x}_1 = (x_1 + \ldots + x_p)/p$ et

$\bar{x}_2 = (x_{p+1} + \ldots + x_{2p})/p$; $b_p = \bar{x}_1$; $s_i = x_i - \bar{x}_1$ pour $i = 1,2,\ldots,p$.

Pour un modèle multiplicatif : $y_{n+h} = (a_n \times h + b_n)s_{n+h}$ où

$b_i = \alpha(x_i/s_{i-p}) + (1-\alpha)(b_{i-1} + a_{i-1})$; $s_i = \beta(x_i/b_i) + (1-\beta)s_{i-p}$;

$a_i = \gamma(b_i - b_{i-1}) + (1-\gamma)a_{i-1}$ pour $i > p$ avec les valeurs initiales

$a_p = \bar{x}_2 - \bar{x}_1$, $b_p = \bar{x}_1$; $s_i = x_i/\bar{x}_1$ pour $i = 1,\ldots,p$.

Choix des valeurs α, β et γ. Comme dans le cas du lissage exponentiel simple, on choisit le triplet de valeurs (α, β, γ) qui minimise la variance des écarts et donc qui minimise $\sum_{i=1}^{n}(\boldsymbol{y}(t_i) - (\boldsymbol{x}(t_i))^2$.

Section 6

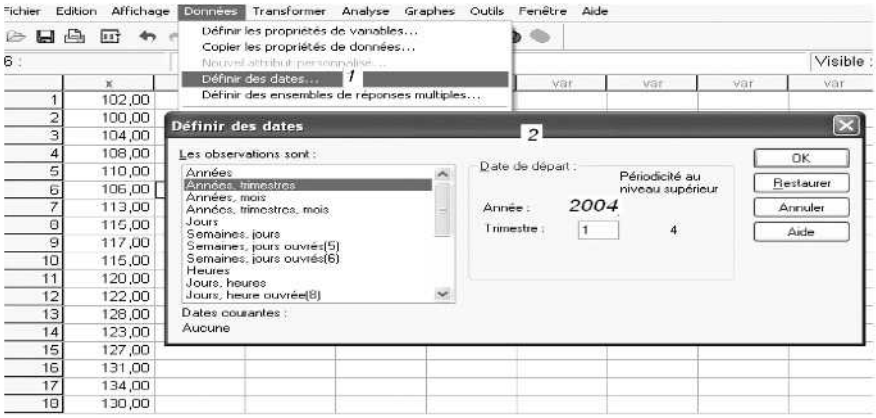
TRAITEMENTS AVEC LOGICIELS

1 Traitements avec SPSS

On reprend l'application de la section 3 1. où l'entreprise a relevé pendant quatre ans ses ventes trimestrielles ($\nu = 4$ et donc $p = 2$) et l'on souhaite dégager le trend

par la technique de la moyenne mobile. La série chronologique $x(t)$ se décomposant de façon additive et x_h désignant la h-ème observation, on a : $x_h = T_h + S_{r(h)} + A_h$ avec T_h le trend, $S_{r(h)}$ le facteur saisonnier, et A_h le facteur accidentel.

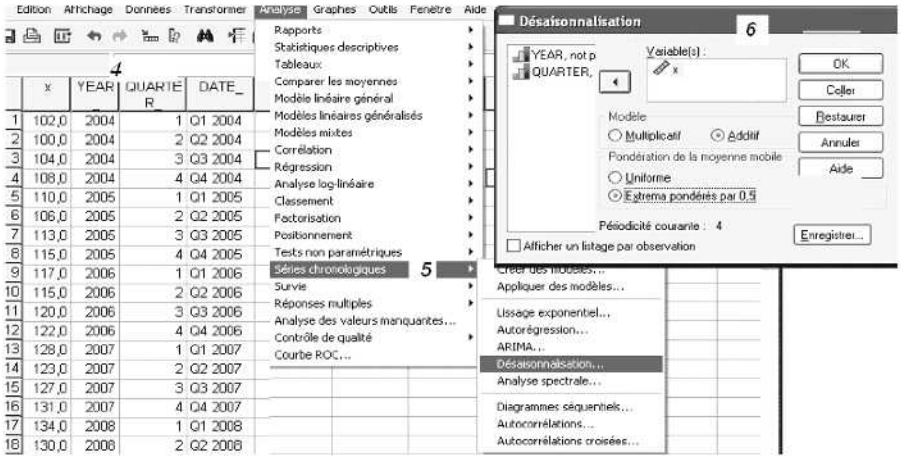
1.1 Obtention d'une série chronologique



Procédure. 1. On entre les chiffre d'affaires trimestriels (variable x) dans la première colonne. **2.** On clique sur (Données), (Définir des dates).

3. On sélectionne l'indexation par ((Années, trimestre)). Le premier relevé est celui du premier trimestre 2004 et la périodicité est 4 (nombre $\nu = 4$) Dans la feuille des données apparaissent les variables year_ (année) quarter_ (trimestre) et date.

1.2 Désaisonnalisation



© Dunod. La photocopie non autorisée est un délit.

5. On sélectionne (Analyse) puis (Série chronologique) et (Désaisonnalisation).

6. On sélectionne la variable ***x*** à désaisonnaliser et on opte pour le modèle additif. La périodicité étant paire $\nu = 2p$, on retient Extrema pondérés par (0.5).

On obtient alors les composantes de la série chronologique. *err*_1 désigne le facteur accidentel A_h, ***sas*_1 la série désaisonnalisée, *saf*_1 le coefficient saisonnier** et ***stc*_1 le trend.**

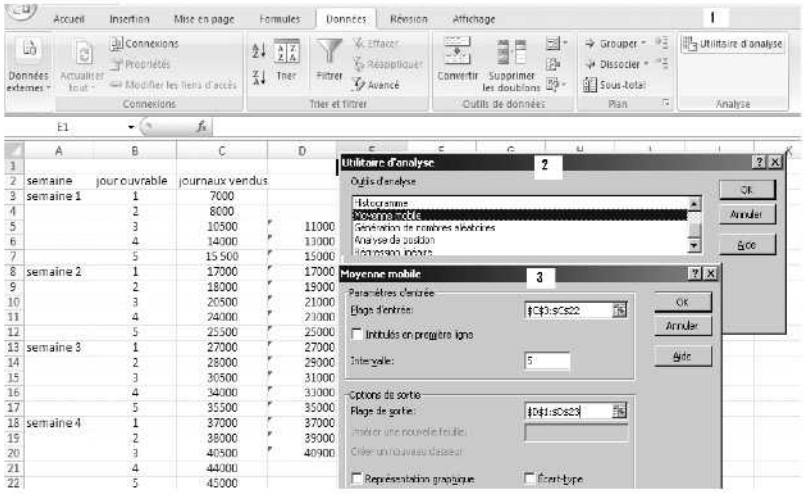
x	YEAR	QUARTE R	DATE_	ERR_1	SAS_1	SAF_1	STC_1
102,0	2004	1	Q1 2004	-1,59	99,69	2,31	101,28
100,0	2004	2	Q2 2004	,94	103,28	-3,28	102,34
104,0	2004	3	Q3 2004	,40	104,05	-,05	104,45
108,0	2004	4	Q4 2004	,65	106,98	1,02	106,33
110,0	2005	1	Q1 2005	-,38	107,69	2,31	108,08
106,0	2005	2	Q2 2005	-,75	109,28	-3,28	110,03
113,0	2005	3	Q3 2005	1,04	113,05	-,05	112,01
115,0	2005	4	Q4 2005	,10	113,98	1,02	113,89
117,0	2006	1	Q1 2006	-1,05	114,69	2,31	115,74
115,0	2006	2	Q2 2006	,58	118,28	-3,28	117,70
120,0	2006	3	Q3 2006	,15	120,05	-,05	119,89
122,0	2006	4	Q4 2006	-1,13	120,98	1,02	122,11
128,0	2007	1	Q1 2007	1,39	125,69	2,31	124,30
123,0	2007	2	Q2 2007	,13	126,28	-3,28	126,14
127,0	2007	3	Q3 2007	-,85	127,05	-,05	127,89
131,0	2007	4	Q4 2007	,32	129,98	1,02	129,66
134,0	2008	1	Q1 2008	,04	131,69	2,31	131,65
130,0	2008	2	Q2 2008	,63	133,28	-3,28	132,64

2 Traitements avec Excel

Est directement utilisable le module consacré aux moyennes mobiles dans le cas où le nombre de périodes constitutives du mouvement saisonnier est impair $\nu = 2 \times p + 1$.

Exemple

On s'intéresse aux fluctuations quotidiennes des ventes d'un nouveau journal régional vendu tous les jours ouvrés de la semaine. La périodicité du mouvement saisonnier est $\nu = 5$ (cinq jours ouvrables dans la semaine $\nu = 2 \times p + 1$ avec $p = 2$).



Procédure. 1. On entre les semaines, jours ouvrables et le nombre de journaux vendus.

2. On clique sur Données puis Utilitaire d'analyse. **3.** On sélectionne Moyenne Mobile. Le premier relevé est celui du premier jour ouvrable de la semaine 1 et la périodicité est 5 (nombre $\nu = 5$).

4. On sélectionne les cellules correspondant à la série chronologique \$C\$3:\$C\$22 puis on sélectionne pour plage de sortie \$D\$1: \$D\$23 (soit 2 lignes plus haut ou d'une façon générale p lignes plus hauts).

Les fonctions habituelles du tableur permettent de calculer la moyenne mobile d'une série dans le cas où le nombre de périodes du mouvement saisonnier est pair.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 340)

Exercice 1

La série chronologique suivante représente le chiffre d'affaires mensuel y d'un rayon d'une grande surface en milliers d'euros de l'année $N-2$ à l'année N.

Année \ Mois	1	2	3	4	5	6	7	8	9	10	11	12
$N-2$	300	195	685	790	475	475	405	756	720	1 350	1 000	1 280
$N-1$	1 560	1 455	1 945	2 050	1 735	1 735	1 665	2 015	1 985	2 610	2 260	2 540
N	2 820	2 715	3 204	3 310	2 995	2 995	2 925	3 277	3240	3 870	3 520	3 800

1. Représenter graphiquement la série y. Pourquoi peut-on penser à un modèle additif ?
2. Désaisonnaliser la série par la technique de la moyenne mobile.

Exercice 2

La série chronologique suivante représente le chiffre d'affaires trimestriel y_i d'une entreprise en millions d'euros de 2003 à 2007.

Années \ Trimestres	1	2	3	4	Moyenne annuelle	Écart-type
2003	1 650	2 038	2 430	1 904	2 005,5	281,93
2004	1 704	2 292	2 746	2 188	2 232,5	370,30
2005	1 904	2 566	3 044	2 626	2 535	408,19
2006	2 354	3 186	3 656	2 994	3 047,5	467,24
2007	2 550	3 196	4 022	2 912	3 170	542,57

1. Pourquoi faut-il décomposer cette série chronologique suivant un modèle multiplicatif ?
2. Désaisonnaliser la série par la méthode de la moyenne mobile.
3. Dégager le trend selon la méthode des moindres carrés.

Exercice 3

La SND a développé depuis la fin de l'année $N-4$ une division de légumes en semi-conserve. Les ventes de ces produits ne sont pas régulières tout au long de l'année et le directeur commercial souhaite étudier la structure des ventes, à la lumière des statistiques disponibles données ci-après, afin d'établir une prévision pour l'exercice $N+1$.

Statistique de vente de centaines de barquettes de $N-3$ à N

Année \ Trimestre	Premier trimestre	Deuxième trimestre	Troisième trimestre	Quatrième trimestre	Moyenne annuelle
$N-3$	152	175	126	140	148,25
$N-2$	185	207	151	172	178,75
$N-1$	218	238	180	205	210,25
N	249	280	215	243	246,75

Calculer les coefficients saisonniers trimestriels par la méthode des rapports au trend.

QCM 1. On dispose des données semestrielles de 3 années consécutives. Afin de lisser les variations saisonnières, le trend est calculé par la méthode de la moyenne mobile portant sur deux semestres. Le trend est alors caractérisé par ν valeurs consécutives : ① $\nu = 6$; ② $\nu = 5$; ③ $\nu = 4$; ④ $\nu = 3$; ⑤ aucune réponse proposée ne convient.

QCM 2. Des relevés statistiques trimestriels ont été réalisés au cours de 4 années consécutives : $t = 1, 2, 3, 4$. La série chronologique est correctement ajustée par le modèle $Y = 1{,}05t + 820 + (0{,}1) \times \cos(\omega t + 0{,}6) + A_t$ où $(0{,}1) \times \cos(\omega\pi t + 0{,}6)$ est la composante saisonnière et A_t est le facteur aléas. La constante positive ω est égale à : ① $\pi/6$; ② $\pi/4$; ③ $\pi/3$; ④ $\pi/2$; ⑤ aucune réponse proposée ne convient.

© Dunod. La photocopie non autorisée est un délit.

5 ÉLABORATION ET FIABILITÉ D'UN QUESTIONNAIRE

L'élaboration et l'administration d'un questionnaire permet d'étudier des *comportements* (acheter ou non des actions d'une entreprise privatisée), des *opinions* (opinions sur une campagne publicitaire réalisée par telle banque), des *attitudes* (attitudes vis-à-vis de l'argent), des *intentions* (intention ou non d'épargner) et sur les *caractéristiques* ou identité de la personne interrogée (sexe, âge...).

La démarche de construction d'un questionnaire comprend quatre phases :

– répertorier les éléments d'informations souhaités,
– définir le type de questions à poser et leur forme (question ouverte, fermée...), leur nombre, l'ordre des questions...
– réaliser le prétest du questionnaire auprès d'un petit échantillon pour vérifier le degré de compréhension des questions, l'absence de biais...
– évaluer la fiabilité du test.

Section 1	■	Construction d'un questionnaire
Section 2	■	Ordre des questions et pré-test
Section 3	■	Fiabilité et élaboration des échelles de mesures
Section 4	■	Procédure sous SPSS

Section 1

CONSTRUCTION D'UN QUESTIONNAIRE

1 Analyse du problème étudié

Dans une première étape, il convient de réaliser une analyse précise du phénomène étudié (opinions, intentions, comportements, attitudes) avant d'élaborer les questions. Les questionnaires peuvent servir dans un but exploratoire à décrire des nouvelles attitudes, de nouveaux comportements mais ils sont également utilisés pour (in)valider des hypothèses.

2 Nature des questions

Dans une seconde étape il convient de définir les questions à poser et la forme qu'elles doivent prendre. Il convient notamment de se demander si la question est strictement nécessaire, si une seule question suffit à obtenir l'information recherchée et si l'interlocuteur voudra et pourra répondre à celles-ci. Il convient, lors de la formulation des questions, d'éviter l'emploi de termes mal définis (la réponse doit être sans ambiguïté pour éviter qu'une même réponse recouvre des idées différentes), de termes trop techniques (qu'il convient alors de définir ou d'illustrer par un exemple), de questions complexes ou suggérant la réponse.

Concernant la forme des questions posées, il peut s'agir de *questions ouvertes* où la personne interrogée formule sa propre réponse ou de *questions fermées* où le sondé effectue un choix entre un nombre de réponses prédéterminées.

Une question est dite ouverte lorsqu'aucune réponse n'est préétablie. Elle permet d'obtenir des réponses spontanées, le sondé pouvant formuler la réponse dans son propre langage. Ce type de questions est à privilégier dans le cadre d'études où l'on dispose de très peu d'éléments d'informations sur le phénomène étudié ou que celui-ci est difficilement mesurable. Dans le cadre d'un entretien, la personne se sent davantage valorisée par ce type de question mais le traitement de questions ouvertes se révèle souvent plus long.

Exemple

Question 1 Nom du répondant ————————————

Résidence du répondant————————

Question 2 À quoi vous fait penser la couleur bleu du logo de la banque ?

Question 3 Quels noms de produits financiers offerts par la banque connaissez-vous ? :

————————————

© Dunod. La photocopie non autorisée est un délit.

La question 3 est en fait celui d'une question ouverte pré-codifiée, la réponse est libre mais pré-codifiée (le ou les produits financiers appartiennent à une liste de placements que connaît l'enquêteur)

Dans les questions fermées les réponses sont préétablies, il y a une liste exhaustive, exclusive, catégorisée de réponses possibles. La standardisation des questions et des réponses possibles permet de réaliser facilement de bons traitements statistiques.

3 Catégories de questions fermées

Parmi les questions fermées peuvent être distinguées : les questions dichotomiques offrant une seule alternative au répondant, les questions à choix simple ou multiples et enfin les questions correspondant à des échelles mesurant un phénomène (attitudes, opinions, comportements...). Sont présentées et classées ci-après les principales variantes de questions fermées qui sont utilisées.

3.1 La question dichotomique

Elle n'offre qu'une seule alternative de réponse au sondé.

Exemple

Question 1 Sexe du répondant Homme ❑ Femme ❑ (à cocher)

Question 2 Connaissez-vous les produits suivants offerts par la banque (à cocher) :

	Oui	**Non**
– assurance auto	❑	❑
– assurance habitation	❑	❑
– assurance santé complémentaire	❑	❑

À chaque question peut être associée une variable binaire prenant par exemple la valeur 1 si oui ou 0 dans le cas contraire. Ce type de question sert à caractériser un comportement ($Y = 1$ si achat ou $Y = 0$ si non achat), une identité ($X = 1$ si homme, $X = 2$ si femme).

On peut, pour décrire et expliquer ce genre de variable, utiliser les diagrammes en barre (*cf.* chapitre 1) et recourir à l'inférence statistique à partir d'un échantillon ou sous échantillon : intervalle de confiance de proportion (de ceux qui connaissent l'assurance auto de la banque, *cf.* chapitre 10), test d'indépendance du khi-deux (entre ceux qui connaissent l'assurance auto de la banque et le sexe de la personne

interrogée, *cf.* chapitre 13)...., régression logistique (avec pour variable explicative *Y* variable associée par exemple à l'achat d'un produit, *cf.* chapitre 17...

Ce type de question permet de scinder l'échantillon en deux sous-échantillons (les hommes d'un côté, les femmes de l'autre) et offre alors la possibilité des tests de comparaison (entre hommes et femmes, ceux qui connaissent l'assurance auto et les autres, *cf.* chapitre 12).

3.2 La réponse nominale

Dans ce type de question à chaque réponse possible est associé un chiffre mais l'attribution de ce chiffre ne correspond à aucune relation d'ordre. Ce type de question permet de prendre en compte une classification La réponse peut être unique comme dans la question 1 ci-dessous ou multiple comme dans la question 2.

Exemples

Question 1. Quelle est votre activité professionnelle actuelle ? (entourer la réponse)

Agriculteur	1
Artisan, commerçant et chef d'entreprise	2
Cadres, profession libérale et intellectuelle supérieure	3
Profession intermédiaire	4
Employé	5
Ouvrier	6
Retraité	7
Étudiant, lycéen	8
Chômeur	9
Autres Inactifs	10

Question 2. J'ai connaissance du produit financier par (entourer la ou les réponses) :

– par mon conseiller financier..... 1
– par la publicité dans la presse quotidienne régionale 2
– par les prospectus disponibles à l'agence 3
– par la publicité à la radio 4

Dans le cas de réponses nominales exclusives, il y a une seule variable par question (comme *question 1*). Les éléments de statistique descriptive pouvant être employés sont les diagrammes circulaires ou en barres, les tableaux avec les modalités et les fréquences relatives, le mode (*cf.* chapitre 1). Ce type de variables couplé à d'autres variables peut être utilisé pour les tests d'indépendance du Khi-deux (*cf.* chapitre 13).

Ce type de question sert aussi à scinder l'échantillon en plusieurs sous-échantillons correspondant à différentes populations (agriculteurs, artisans, cadres...) et

© Dunod. La photocopie non autorisée est un délit.

offre alors la possibilité de tests de comparaison d'un même caractère sur deux sous-populations ou d'analyse de variances sur plusieurs sous-populations (*cf.* chapitre 15).

Dans le cas de réponses nominales multiples (*question 2*), il faut coder chaque réponse par une variable binaire prenant la valeur 0 ou 1 si la réponse est entourée et l'on retrouve les traitements proposés pour les variables dichotomiques.

3.3 La réponse ordinale

Il est demandé au sondé de classer les réponses possibles dans l'ordre de ses préférences. Autrement dit, il lui est demandé d'établir une relation d'ordre entre les réponses possibles.

Exemple

> *Question 1* – Classez par ordre d'importance vos critères de choix en terme de produit financier (par ordre décroissant mettre la valeur 1 pour le critère le plus important..., 3 pour le critère le moins important)
>
> sa sécurité .... son rendement .... sa liquidité ....

Dans cet exemple où figurent 3 critères, à chaque réponse est associé un triplet d'entiers naturels (r_1, r_2, r_3) où r_1, r_2 et r_3 sont les rangs respectifs attribués par le répondant aux critères *sécurité*, *rendement* et *liquidité* : $1 \leqslant r_i \leqslant 3$, $r_i \neq r_j$, $r_1 + r_2 + r_3 = n(n+1)/2$ où ici $n =$ nombre de critères à classer $= 3$.

Afin d'établir une hiérarchie et d'éventuels liens entre les critères proposés, on peut déterminer le rang médian de chaque critère puis utiliser des tests non paramétriques avec départition des ex-eaquo (*cf.* par exemple tests des rangs signés de wilcoxon).

3.4 Les échelles de notation

Les échelles de notation traduisent l'intensité du jugement exprimé et sont souvent utilisées pour mesurer les comportements, attitudes et phénomènes cognitifs. L'échelle est dite simple lorsque le phénomène est observé à partir d'une question (ou item), elle est dite multiple lorsqu'elle est élaborée avec plusieurs items. Les échelles les plus courantes emploient 4, 5, 6, 7 ou 9 modalités de réponses, la plus fréquente étant l'échelle à 7 points. Les échelles à nombre impair de questions permettent une position neutre alors que les échelles paires obligent le sondé à prendre position de façon plus tranchée. On distingue essentiellement :

① **l'échelle dite verbale** lorsque toutes les catégories de réponses sont repérées par des mots.

Exemple

Question 1. Lorsque vous achetez un produit financier la sécurité est-elle un critère

Très important	Important	Assez important	Moyennement important	Assez peu important	Peu important	Très peu important

② l'échelle numérique

L'échelle est dite numérique lorsque toutes les catégories de réponses sont repérées par des chiffres.

Exemple

Question 2. Notez sur une échelle de 1 à 10 la pertinence des informations transmises par votre conseiller financier (1 pour pas du tout pertinentes..., 10 pour tout à fait pertinente)

10	9	8	7	6	5	4	3	2	1

Dans ce type d'échelle figure l'échelle de *Likert* qui consiste à faire évaluer plusieurs énoncés (ou *items*) correspondant à une attitude, au moyen d'opinions reflétant le degré d'accord avec ces énoncés. Le nombre d'échelons de l'échelle est en général impair soit généralement sept ou cinq.

Exemple

Question 1 – Indiquez votre degré d'accord avec les propositions suivantes en entourant la réponse (1 pour pas du tout d'accord et 7 pour tout à fait d'accord) pour chaque item proposé.

Je suis satisfait de l'accueil à l'agence 1 2 3 4 5 6 7

Je suis satisfait de l'information donnée à l'agence 1 2 3 4 5 6 7

Dans ce type d'échelle peut également figurer l'*échelle sémantique différentielle d'Osgood* qui a été crée pour analyser la structure sémantique sous-jacente des mots qui traduisent des évaluations et des impressions. C'est une échelle bipolaire en 7 points, notée de –3 à +3 avec 0 pour point neutre avec aux extrémités deux antonymes soigneusement choisis.

Exemple.

Question 2. Que pensez-vous du service au guichet :

Rapide	+3	+2	+1	0	–1	–2	–3	Lent

© Dunod. La photocopie non autorisée est un délit.

③ **L'échelle mixte** lorsque les catégories de réponses sont repérées par des chiffres mais que les extrémités de l'échelle ont leur correspondance verbale.

Exemple

> *Question 3*. L'information transmise par la banque sur le crédit à la consommation vous paraît-elle complète pour prendre votre décision ? (entourer le chiffre correspondant)
>
> Pas complète du tout 1 2 3 4 5 6 7 Très complète

④ **L'échelle graphique ou illustrée** (visages souriants).

Ces échelles permettent des études sur des populations dont les capacités de traitement sont limitées (notamment les enfants) ou auprès de populations dont les capacités de traitement sont différentes des nôtres.

⑤ **L'échelle peut être unidirectionnelle** allant du pôle négatif au pôle positif ou être bidirectionnelle et dans ce dernier cas elle peut être symétrique ou asymétrique avec ou sans point neutre

Les échelles de notation permettent d'établir une relation d'ordre et autorise dans une certaine mesure le calcul des distances entre les objets évalués sur une même échelle. Les échelles métriques permettent des représentations sous forme de diagramme en bâtons, elles se prêtent aux intervalles de confiance de la moyenne, de la médiane, de l'écart-type... Elles peuvent faire l'objet de tests paramétriques et non paramétriques.

Section 2 ORDRE DES QUESTIONS ET PRÉ-TEST

Le questionnaire doit commencer par des questions générales (relativement faciles) et progresser vers des questions plus précises et difficiles. Il convient lorsque l'on formule les questions d'éviter les effets de *halo* et de *contamination*.

L'effet de halo correspond à la situation où la personne interrogée a tendance à répondre toujours de la même manière lorsque plusieurs questions sont posées sous formes d'échelles orientées dans le même sens. Pour éviter ce biais il est judicieux de changer de temps en temps le sens des échelles. Il y a effet de contamination lorsqu'une réponse à une question influence la réponse à une autre question.

Il est possible de poser des *questions filtres* qui permettent de détecter les personnes qui sont concernées par un ensemble de questions (parce que par exemple, elles ont acheté le produit ou le service pour lequel on souhaite savoir sa satisfaction) et qui vont pouvoir répondre aux questions suivantes et d'autres qui ne sont pas concernées (exemple : elles n'ont pas acheté le produit).

Exemple

On s'intéresse à la perception par le client des services délivrés par une banque. Après avoir abordé les questions sur l'identité de la personne, on souhaite poser une partie des questions (*n* questions) qu'aux seules personnes ayant demandé ou souscrit un crédit à l'habitat. Dans ce cas là on peut recourir à une question filtre

la question n° 4 : avez-vous demandé ou souscrit un crédit à l'habitat **Oui** ❑ **Non** ❑

si réponse non, passez à la question $4 + n$

Dans une dernière étape le pré-test du questionnaire auprès d'un petit échantillon permet de vérifier le degré de compréhension des questions, l'absence de biais et conduit à la rédaction finale du questionnaire

Section 3

FIABILITÉ ET ÉLABORATION DES ÉCHELLES DE MESURE

L'élaboration d'échelles de mesure multiples nécessite une grande expertise.

1 Processus d'élaboration

REPÈRES

Selon le paradigme de Churchill G.A., la démarche d'analyse et de recherche pour construire des échelles de mesure est la suivante :

1/ spécifier le *construit* ou caractéristique à évaluer (comportement, attitude, phénomène) après examen de la théorie sous-jacente au construit et des éléments d'observation du phénomène à mesurer. Ainsi pour construire une échelle il convient d'examiner d'un point de vue théorique la définition du construit, les items retenus pouvant provenir de la revue de la littérature, du corpus théorique ou d'études qualitatives. S'il n'y a pas de travaux antérieurs on peut préalablement réaliser des séries d'entretiens au moyen de question ouvertes à partir du concept étudié ;

2/ créer l'ensemble d'items possibles mesurant le comportement, l'attitude, le phénomène étudié. Il convient de s'interroger « les items donnent-ils une vision représentative et exhaustive du phénomène étudié ? » ;

3/ constitution d'un premier échantillon et recherche des données permettant d'évaluer la pertinence des échelles ;

4/ s'assurer de la purification des échelles et estimer leur fiabilité par l'alpha de Cronbach (noté α) et l'analyse factorielle. On peut procéder à une analyse factorielle pour identifier les dimensions de l'échelle et éliminer les items qui contribuent peu aux axes retenus dans l'analyse factorielle. De même lorsque l'on obtient une faible valeur du coefficient α sur l'ensemble des items d'une échelle on peut être amené à supprimer les items qui sont faiblement corrélés aux autres et réduisent la valeur de l'alpha de Cronbach.

5/ collecter des données finales sur un deuxième échantillon et étudier la fiabilité des échelles (l'alpha de Cronbach...), leur validité convergente et discriminante.

© Dunod. La photocopie non autorisée est un délit.

2 Mesures de fiabilité

La fiabilité correspond au degré de cohérence des résultats d'une échelle lorsque l'on répète des mesures.

2.1 Méthode du test/retest

La méthode du test/retest consiste à proposer à des sondés des séries d'items identiques d'une même échelle à deux moments différents. Le degré de similitude entre les deux mesures est déterminé par le calcul d'un coefficient de corrélation entre les deux mesures aux sondés, plus la corrélation est forte meilleure est la fiabilité. Evidemment l'effet du temps (évolution du comportement avec le temps), l'existence d'un effet de report (surtout si le second test est réalisé à une date approchée du premier), l'effet de l'administration du questionnaire précédent sont autant d'éléments pouvant biaiser cette analyse.

Exemple

On demande avec un intervalle de deux semaines aux dirigeants d'une PME d'indiquer leur degré d'accord concernant la qualité de l'information dont ils disposent sur les progiciels intégrés en répondant à cinq items d'une échelle (1 pour pas du tout d'accord et 7 pour tout à fait d'accord).

L'information est exhaustive	1	2	3	4	5	6	7
L'information est précise	1	2	3	4	5	6	7
L'information est fiable	1	2	3	4	5	6	7
L'information est claire	1	2	3	4	5	6	7
L'information est riche	1	2	3	4	5	6	7

2.2 Cohérence interne des items

Les items d'une même échelle devraient être corrélées puisqu'ils sont censés mesurer le même phénomène. Plusieurs techniques statistiques (le *split-half*, les coefficients alpha de Cronbach, le coefficient de Spearman-Brown) permettent d'évaluer la cohérence interne d'une échelle de mesure.

La mesure la plus simple de cohérence interne consiste à partager l'ensemble des items en deux parties (méthode *split-half*) puis à élaborer un score total pour chacune des deux parties et calculer le coefficient de corrélation entre les deux scores sur l'échantillon.

La méthode la plus usitée, est le « *coefficient alpha* (α)» encore appelé « *alpha de Conbrach* ». Il synthétise le degré de corrélations entre réponses aux questions dif-

férentes d'un même questionnaire. Le coefficient α permet de mesurer la fiabilité d'un ensemble de questions ou items (I_1, I_2...,I_k) censées contribuer à la mesure d'un phénomène. La valeur du coefficient α est donnée par la formule suivante :

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_i \sigma_i^2}{\sigma_s^2}\right) = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_i \sigma_i^2}{\sum_i \sigma_i^2 + 2\sum_{i,j} \sigma_{i,j}}\right)$$

où k est le nombre de questions ou items ; σ_i^2 la variance de l'item I_i, $\sigma_{i,j}$ la covariance entre l'item I_i et l'item I_j et σ_S^2 est la variance observée du score total $(I_1 + I_2 + ... + I_k)$.

La valeur du coefficient α est comprise entre 0 et 1. Plus la valeur du coefficient α est proche de 1, plus les items qui composent l'énoncé sont corrélés entre eux. Lorsque la valeur de α est proche de 1 on dira qu'il existe une bonne cohérence interne entre ces items qui participent à expliquer le même phénomène. Par contre, on conclut le contraire lorsque la valeur de α est proche de 0. *L'American Psychological Association* considère un questionnaire comme acceptable quand le coefficient alpha est au dessus de 0,7.

– On peut également utiliser un coefficient de Spearman-Brown

$r_k = \dfrac{k\bar{r}}{1 + (k-1)\bar{r}}$ avec k le nombre d'items et $\bar{r}$ le coefficient de corrélation moyen entre items qui représente une mesure de l'homogénéité des énoncés. Le score idéal est fixé à la valeur 1.

REPÈRES : Fondements de la théorie de la fiabilité des questionnaires

Si les individus sont pris au hasard dans une population donnée, le score aléatoire obtenu par chacun d'eux à un item est une variable aléatoire X qui se décompose en la somme d'un score vrai aléatoire T et d'une erreur de mesure ε soit $X = T + \varepsilon$ où ε est une variable centrée non corrélée avec T.

On a donc évidemment $E(X) = E(T)$.

La théorie classique des tests s'intéresse aux relations entre ces trois variables qui permettent d'établir la fiabilité des tests.

La fiabilité

La fiabilité d'une mesure X, notée ρ_X^2 est évaluée par le quotient des variances $V(T)/V(X)$: $\rho_X^2 = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / (\sigma_T^2 + \sigma_\varepsilon^2)$.

Lorsque le questionnaire comporte k items It_1, It_2, ..., It_k, à It_j est associée la variable X_j et sa décomposition $X_j = T_j + \varepsilon_j$. On considère alors la fiabilité de la variable composite

$$S = \sum_{j=1}^{k} X_j : \ \rho_S^2 = \sigma^2\left(\sum_{j=1}^{k} T_j\right) / \sigma^2(S)$$

© Dunod. La photocopie non autorisée est un délit.

L'estimation de cette fiabilité peut se faire par la corrélation entre deux mesures lorsque les variables sont parallèles.

Mesure de la fiabilité lorsque les variables sont parallèles

Deux mesures X et X' sont dites strictement parallèles si $X = T + \varepsilon$ et $X' = T + \varepsilon'$ avec $V(\varepsilon) = V(\varepsilon')$ et $\text{Cov}(\varepsilon, \varepsilon') = 0$. Sous ces simples hypothèses on constate que le coefficient de corrélation linéaire $r(X, X')$ est égal à leur fiabilité : $r(X, X') = \text{Cov}(X, X')/\sigma(X) \times \sigma(X') = \rho^2_X = \rho^2_{X'}$. En effet $\text{Cov}(X, X') = \text{Cov}(T + \varepsilon, T + \varepsilon') = V(T)$ car $\text{Cov}(T, \varepsilon) = \text{Cov}(T, \varepsilon') = \text{Cov}(\varepsilon, \varepsilon') = 0$ et $\sigma(X) = \sigma(X')$

Si l'on dispose de k scores parallèles $X_1,..., X_k$ alors la fiabilité ρ^2_S de leur somme $S = X_1 + ... + X_k$ est estimée par $\rho^2_S = k\rho^2 / (1 + (k - 1)\rho^2)$ où ρ^2 est le coefficient de corrélation linéaire moyen entre les variables $X_1,..., X_k$.

En effet, $\rho^2_S = V(kT + \varepsilon_1 + ... \varepsilon_k)/V(S)$. Or $V(kT + \varepsilon_1 + ... \varepsilon_k) = k^2 V(T) = k^2 \rho^2_X V(X)$ et

$V(S) = kV(X) + k(k - 1)Cov(X, X') = kV(X) + k(k - 1)\rho^2_X V(X)$ implique

« $\rho^2_S = k\rho^2_X / (1 + (k - 1)\rho^2_X)$ » où $\rho^2_X = r(X, X')$

Cette formule, appelée formule de Spearman-Brown, montre que dans le cas de variables parallèles, la fiabilité augmente avec le nombre d'items et tend vers 1.

Les mesures tau équivalentes

Deux mesures X et X' sont dites *tau équivalentes* lorsque $X = T + \varepsilon$ et $X' = T + \varepsilon'$ et $\text{cov}(\varepsilon, \varepsilon') = 0$ mais à la différence de mesures parallèles, les variances des erreurs aléatoires σ^2_ε et $\sigma^2_{\varepsilon'}$ peuvent être différentes. Sous l'hypothèse que les k variables $X_1, X_2,..., X_k$ associées aux k items $It_1, It_2, ..., It_k$ soient *tau* équivalentes, soit « $X_j = T_j + \varepsilon_j$ avec $T_j = T$ et $\forall\, h \neq l$, $\text{Cov}(\varepsilon_h, \varepsilon_l) = 0$ », la fiabilité du score total $S = \sum_{j=1}^{k} X_j$ est égal à l'alpha de Cronbach :

$$\rho^2_S = \frac{\sigma^2\left(\sum_{j=1}^{k} T_j\right)}{\sigma^2\left(\sum_{j=1}^{k} X_j\right)} = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k}\sigma^2(X_j)}{\sigma^2\left(\sum_{j=1}^{k} X_j\right)}\right) = \alpha \text{ de Cronbach}$$

En effet, $V\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} V(X_i) + 2\sum_{1 \leqslant i \leqslant j \leqslant k} \text{Cov}(X_i, X_j) = \sum_{j=1}^{k} V(X_j) + k(k-1)V(T)$ car

$\text{Cov}(X_i, X_j) = Cov(T + \varepsilon_i, T + \varepsilon_j) = V(T)$ et

$$V\left(\sum_{j=1}^{k} T_j\right) = V(kT) = k^2 V(T) = (k/(k-1))\left[V\left(\sum_{i=1}^{k} X_i\right) - \sum_{j=1}^{k} V(X_j)\right]$$

Le coefficient alpha de Cronbach permet d'estimer la fidélité de la somme des variables lorsqu'elles possèdent la propriété de *tau-équivalence essentielle* (TEE).

3 Validité

L'étude de validité consiste à apprécier dans quelle mesure l'échelle proposée permet d'observer le *construit*.

La validité de contenu d'une échelle est sa capacité de refléter de façon exhaustive le construit, autrement dit l'échelle doit représenter tous les aspects du construit.

L'étude de validité de la mesure d'un construit vise à vérifier que la mesure retenue mesure parfaitement et uniquement le construit.

Une échelle a une bonne validité convergente lorsqu'elle est fortement corrélée avec des échelles appréhendant un même construit. Cette validité peut être appréciée à l'aide du *W* de Kendall, on retient généralement la valeur 0,8 (*cf*. chapitre 15 pour l'expression de *W*).

Une échelle a une bonne valeur discriminante lorsqu'elle fait bien la différence entre le construit mesuré et tout autre construit. Autrement dit l'échelle proposée pour mesurer un construit doit être faiblement corrélé avec les autres échelles correspondant aux autres construits (autrement dit les items de l'échelle mesurant un construit sont plus fortement corrélés entre eux qu'avec les mesures de tout autre construit).

Section 4

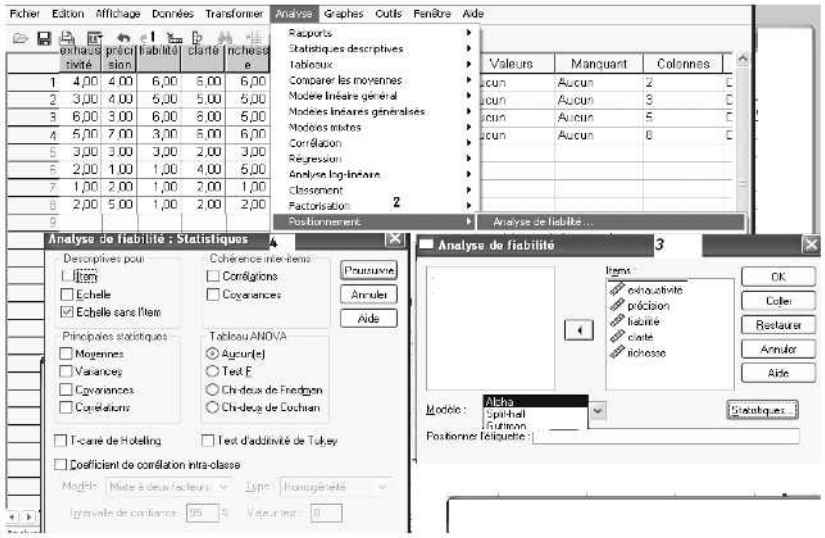
PROCÉDURE SOUS SPSS

On interroge huit individus et on leur demande d'indiquer leur degré d'accord avec les propositions suivantes concernant la qualité de l'information dont ils disposent sur les progiciels intégrés (1 pour pas du tout d'accord et 7 pour tout à fait d'accord)

L'information est exhaustive	1	2	3	4	5	6	7
L'information est précise	1	2	3	4	5	6	7
L'information est fiable	1	2	3	4	5	6	7
L'information est claire	1	2	3	4	5	6	7
L'information est riche	1	2	3	4	5	6	7

Les réponses sont consignées ci-dessous.

© Dunod. La photocopie non autorisée est un délit.



Procédure. 1. On entre les données correspondant aux 8 huit individus et cinq variables.

2. On clique sur (Analyse), (Positionnement) et on sélectionne (Analyse de fiabilité)

3. Dans le menu analyse de fiabilité on sélectionne le modèle méthode alpha (de Cronbach) ou split-half **4**. On saisit les éléments souhaités soit ici échelle sans item

On obtient les résultats suivants

Statistique de fiabilité

Alpha de Cronbach'	Nombre d'éléments
,875	5

L'alpha de cronbach étant supérieur à 0,7 on en conclut que l'échelle a une bonne cohérence interne. Le tableau de valeur de l'alpha de cronbach en cas de suppression de l'item permet de savoir, lorsque l'échelle n'est pas très bonne, s'il faut ou non maintenir l'item dans l'échelle.

6 ANALYSE COMBINATOIRE

Partant d'un ensemble F constitué de n éléments distincts b_1, b_2, …, b_n nous définirons les notions de permutations, d'arrangements, de combinaisons…

Section 1 ■ Permutations

Section 2 ■ Arrangements

Section 3 ■ Combinaisons

Section 4 ■ Répartition d'éléments non différentiables

Section 5 ■ Formule de Poincaré

© Dunod. La photocopie non autorisée est un délit.

Section 1

PERMUTATIONS

Soit un ensemble F constitué de n éléments distincts $F = \{b_1, b_2, \ldots, b_n\}$.

Une permutation de n éléments est un ensemble rangé de ces n éléments, soit par exemple : $(b_1, b_2, \ldots, b_n)$ ou $(b_2, \ldots, b_n, b_1)$ ou …

Les permutations se distinguent les unes des autres par l'ordre de rangement de ces n éléments. Le nombre de permutations que l'on peut réaliser avec les n éléments de F est égal à $n!$ (où $n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$).

Exemples

1/ Avec l'ensemble F constitué des 3 lettres b,c,d on peut réaliser $3! = 6$ permutations distinctes : $\{d,c,b\}$, (d,b,c), (c,b,d), (c,d,b), (b,c,d), (b,d,c).

2/ Un DRH doit attribuer des primes aux employés les plus méritants du service comptable qui comporte 20 employés $\{e_1,e_2,\ldots,e_{20}\}$. Décidant de classer ces employés selon un critère qui lui est propre, son classement est un élément parmi les $n! = 20!$ classements théoriquement possibles.

Section 2

ARRANGEMENTS

1 Arrangements avec répétition

Soit un ensemble F constitué de n éléments $b_1,b_2,\ldots,b_n$ distincts : $F = \{b_1,b_2,\ldots,b_n\}$. Pour tout entier $p (p \geqslant 2)$ on définit l'ensemble F^p des p-uples constitués d'éléments non nécessairement distincts de F, chaque p-uple étant appelé p-arrangements avec répétition.

Ainsi F^3 désigne l'ensemble des triplets constitués avec des éléments de F et l'on a par exemple (si $n \geqslant 5$) : $(b_3, b_5, b_3) \in F^3$, $(b_2,b_4,b_1) \in F^3$, $(b_3,b_3,b_3) \in F^3$, etc.

F^p est constitué de n^p p-uples. Autrement dit : card F^p = (card $F)^p = n^p$.

Rappelons que l'expression card F désigne le nombre d'éléments constituant l'ensemble F.

Exemples

1/ L'ensemble F étant constitué des 3 lettres b,c et d : $F = \{b,c,d\}$, l'ensemble F^2 contient $3^2 = 9$ couples de lettres :

$F^2 = \{(b,b),(b,c),(b,d),(c,b),(c,c),(c,d),(d,b),(d,b),(d,d)\}$.

2/ Le DRH dispose de trois primes d'un montant différent qu'il peut attribuer à sa guise soit à un seul employé soit à deux employés soit à trois des 20 employés. Le nombre de façons de répartir ces trois primes est égal au nombre d'arrangements de trois éléments avec répétition : (e_1,e_1,e_1) ou (e_1,e_1,e_2) ou (e_1,e_2,e_1) ou $(e_1,e_2,e_4)\ldots$ ou $(e_1,e_2,e_3)\ldots$ ou (e_3,e_1,e_2) ou … et est donc égal $3^{20} = 3\,486\,784\,401$.

2 Arrangements sans répétition

Soit un ensemble $F = \{b_1,b_2,\ldots,b_n\}$ constitué de n éléments distincts. On appelle arrangement de ces n éléments pris p à p, toute suite de p éléments distincts

prélevés dans F, l'ordre de rangement des p éléments distinguant un arrangement d'un autre arrangement. Le nombre de ces arrangements est égal à

$$A_n^p = \frac{n!}{(n-p)!}$$

où $n! = n \times (n-1) \times (n-2) \times ... \times 2 \times 1$

et $(n-p)! = (n-p) \times (n-p-1) \times (n-p-2) \times \ldots \times 2 \times 1$

Autrement dit, H_p désignant le sous-ensemble de F^p constitué des p-uples dont les p éléments de F sont distincts, on a card $H_p = A_n^p = \dfrac{n!}{(n-p)!}$

Exemples

1/ L'ensemble $F = \{a,b,c,d,e\}$ a cinq éléments $(n = 5)$. $H_3 = \{(a,b,c),(a,b,d),(a,b,e),(b,a,c),\ldots,\}$ est constitué de $A_5^3 = 60$ triplets dont les éléments constitutifs sont distincts. Remarquer que (b,a,c) et (c,b,a) sont deux éléments distincts de H_3. Par contre le triplet (b,a,b) n'appartient pas à H_3 car il y a répétition de la lettre b.
2/ Le DRH dispose de trois primes de montants différents et décide d'en faire profiter trois employés. Le nombre de façons de répartir ces trois primes est égal aux nombres d'arrangements sans répétition de trois éléments (trois individus) prélevés parmi les 20 employés du service : (e_1,e_2,e_3) ou (e_2,e_1,e_3) ou (e_1,e_3,e_4) … soit
$A_{20}^3 = \dfrac{20!}{(20-3)!} = 6\,460.$

Section 3

COMBINAISONS

Soit un ensemble fini $F = \{b_1,b_2,\ldots,b_n\}$ constitué de n éléments distincts. On appelle combinaison de p éléments pris parmi n, tout sous-ensemble de F contenant p éléments, l'ordre étant donc indifférent et les éléments constituant le sous-ensemble étant tous distincts (éléments ne pouvant être répétés).

Le nombre de combinaisons de p éléments de l'ensemble F est égal $C_n^p = \dfrac{n!}{p!(n-p)!}$. Autrement dit si J_p désigne l'ensemble des p-uples $(b_{i_1},b_{i_2},\ldots b_{i_p})$ tels que $i_1 < i_2 < \ldots < i_p$ on a card $J_p = C_n^p$

Exemples

1/ L'ensemble $F = \{a,b,c,d,e\}$ étant constitué de cinq éléments $(n = 5)$, le nombre de combinaisons de 3 éléments de cet ensemble F est égal $C_5^3 = \dfrac{5!}{3!(5-3)!} = 10.$
On constate que J_3 a bien 10 éléments :
$J_3 = \{(a,b,c),(a,b,d),(a,b,e),(a,c,d),(a,c,e),(a,d,e),(b,c,d),(b,c,e),(b,d,e),(c,d,e)\}$.

© Dunod. La photocopie non autorisée est un délit.

2/ Le DRH après consultation du chef du service comptable décide d'attribuer une prime d'un même montant à trois des 20 employés selon un critère qui reste à définir. Aussi y a t-il *a priori* C_{20}^{3} combinaisons possibles ou façons d'attribuer les primes dans le service, soit les combinaisons (e_1,e_2,e_3), (e_1,e_2,e_4), …., (e_1,e_2,e_{20}), (e_2,e_3,e_4), …

Section 4

RÉPARTITION D'ÉLÉMENTS NON DIFFÉRENTIABLES

– Il existe C_{n+p-1}^{p-1} p-uples d'entiers naturels $(n_1,n_2,\ldots,n_p)$ distincts tels que :

$$n_i \geqslant 0 \quad \forall\, i = 1,2,\ldots,p \quad \text{et} \sum_{i=1}^{p} n_i = n.$$

On peut donc répartir n objets identiques dans p casiers de C_{n+p-1}^{p-1} façons distinctes.

– Il existe C_{n-1}^{p-1} p-uples d'entiers naturels $(n_1,n_2,...,n_p)$ tels que :

$$n_i \geqslant 1 \quad \forall i = 1,2,\ldots,p \quad \text{et} \sum_{i=1}^{p} n_i = n.$$

Section 5

FORMULE DE POINCARÉ

Soit un ensemble fini $F = \{b_1,b_2,\ldots,b_n\}$.

$A_1,A_2,\ldots,A_h$ désignent des parties de F. On a les égalités suivantes :

a) $\text{card}\,(A_1 \cup A_2) = \text{card}\,A_1 + \text{card}\,A_2 - \text{card}\,(A_1 \cap A_2)$

b) $\text{card}(A_1 \cup A_2 \cup A_3) = \text{card}\,A_1 + \text{card}\,A_2 + \text{card}\,A_3 - \text{card}\,(A_1 \cap A_2)$
$-\text{card}\,(A_1 \cap A_3) - \text{card}\,(A_2 \cap A_3) + \text{card}(A_1 \cap A_2 \cap A_3)$

c) $\text{card}\,(A_1 \cup A_2 \cup A_3 \cup A_4) = \Sigma_i\,\text{card}\,A_i - \Sigma_{i<j}\,\text{card}(A_i \cap A_j)$
$+\Sigma_{i<j<k}\text{card}\,(A_i \cap A_j \cap A_k) - \text{card}\,(A_1 \cap A_2 \cap A_3 \cap A_4)$

Exercices (*cf.* corrections page 344)

Exercice 1

Un DRH présente à 20 cadres quatre stages de nature différente.

1. Déterminer le nombre de combinaisons distinctes possibles si les cadres doivent suivre un seul des quatre stages proposés.

2. Déterminer le nombre de combinaisons distinctes possibles si les cadres doivent suivre deux des quatre stages.

Exercice 2

Un conseiller commercial est absent, ses 20 clients, 11 hommes et 9 femmes, doivent être répartis entre ses 3 collègues en fonction du nombre de rendez-vous ou plages horaires disponibles : 8 pour M. X premier collègue, 7 pour M. Y deuxième collègue et 5 pour Mme Z troisième collègue.

Déterminer le nombre n de répartitions distinctes qui peuvent être réalisées :

a) si on ne tient pas compte du sexe des clients ;

b) si on tient compte du sexe dans les répartitions possibles, après décision d'attribuer 3 femmes clientes à chaque conseiller.

Exercice 3

De combien de façons distinctes peut-on répartir 100 euros entre quatre personnes chacune recevant au moins 5 euros ?

Exercice 4

Trois véhicules sans chauffeur sont loués pour transporter 14 personnes dont 5 ont le permis de conduire. Dans chaque véhicule il y a une place pour le conducteur et 4 places pour les passagers. De combien de façons distinctes peut-on répartir entre les 3 véhicules les 14 personnes en formant 3 groupes, chaque groupe étant caractérisé par la désignation du conducteur et des passagers ?

QCM 1. Un opérateur doit faire une « checklist » comprenant cinq opérations successives qui doivent être effectuées dans un ordre bien établi. Quel est le nombre de façons de faire cette « checklist » ?

① $5!/2!$ ② $5!$ ③ $(5!)/(3!\times 2!)$ ④ 5^5 ⑤ aucune réponse ne convient.

QCM 2. On dispose d'une urne contenant 3 boules blanches et 2 boules noires. On tire successivement 3 boules (on remet la boule tirée dans l'urne avant le tirage suivant). Quelle est la probabilité de tirer exactement 3 boules blanches ?

① $3^3/5^3$ ② $3^3 \times 2^2/5^5$ ③ $(3!)\times(2!)/(5!)$ ④ $3 \times 3/15$ ⑤ aucune réponse ne convient.

QCM 3. On choisit au hasard un nombre entier de 8 chiffres. Quelle est la probabilité que tous les chiffres soient différents : ① $8!/10^8$ ② $(8!) \times (2!)/10^8$ ③ $10!/10^8$ ④ $10!/(2!) \times 10^8$ ⑤ aucune réponse ne convient.

© Dunod. La photocopie non autorisée est un délit.

7 NOTIONS DE PROBABILITÉS

Les statistiques descriptives concernent un ensemble de données immédiatement disponibles. *A contrario*, le calcul des probabilités permet d'évaluer les chances de réalisation d'un évènement dans le futur.

Section 1 ■ Notions essentielles

Section 2 ■ Probabilité définie sur un référentiel

Section 3 ■ Composition d'évènements

Section 4 ■ Probabilités conditionnelles

Section 1

NOTIONS ESSENTIELLES

1 Référentiel

Il convient de différencier l'expérience, appelée également *épreuve*, du résultat de l'expérience qui, *a priori*, n'est pas connu mais appartient à un ensemble E appelé *référentiel*.

2 Événements

Toute partie A_i d'un référentiel E est appelée événement. Les éléments de E sont appelés événement simple. Un *événement composé* est une partie de E comportant plus d'un élément. E considéré comme une partie de lui-même est appelé *événement certain*. Un événement impossible ou irréalisable est représenté par le symbole $\varnothing$ de l'ensemble vide.

Exemple

1/ On se propose de jeter une pièce de monnaie puis de regarder le résultat. L'épreuve est le jet de la pièce, le référentiel E_1 est l'ensemble des résultats possibles, c'est-à-dire P « pile » ou F « face » : $E_1 = \{P; F\}$.

2/ L'expérience consiste à jeter une pièce de monnaie puis un dé dont les 6 faces sont numérotées de 1 à 6. Le résultat de l'expérimentation consiste en la lecture des faces supérieures de la pièce puis du dé. Le référentiel E_2 est constitué de 12 résultats possibles : $E_2 = \{(P,1),(P,2),\ldots,(P,6),(F,1),(F,2),\ldots,(F,6)\}$. Considérons l'événement A « obtenir pile puis un nombre pair sur la face supérieure du dé ». A est constitué des 3 événements simples $(P,2),(P,4),(P,6)$. Autrement dit, $A = \{(P,2),(P,4),(P,6)\}$.

Section 2

PROBABILITÉ DÉFINIE SUR UN RÉFÉRENTIEL

À chaque événement ou sous-ensemble A_i du référentiel E, on associe sa probabilité de réalisation $p(A_i)$ qui est un nombre réel positif compris entre 0 et 1, soit $0 \leqslant p(A_i) \leqslant 1$. De plus, $p(E) = 1$.

Lorsque E est constitué d'un nombre fini ou dénombrable d'événements, p est une application définie sur le référentiel E et à valeurs réelles dans [0 ; 1] soit

$$p : E \longmapsto [0; 1].$$

Remarque. Lorsque le référentiel $E = \{e_1,e_2,\ldots,e_N\}$ est constitué de N événements simples e_i, la connaissance des probabilités de réalisation de chaque événement simple : $p(e_1) = p_1, p(e_2) = p_2, \ldots, p(e_N) = p_N$ permet de déterminer la probabilité de réalisation de tout événement de E. Si par exemple, $A = \{e_1,e_3,e_5\}$ on a $p(A) = p(e_1) + p(e_3) + p(e_5) = p_1 + p_3 + p_5$.

La condition $p(E) = 1$ implique $p_1 + p_2 + ... + p_N = 1$.

Exemples

L'expérience consiste à jeter une pièce de monnaie homogène puis à lancer dans un deuxième temps un dé homogène si et seulement si on a obtenu pile « P » au jet de la

© Dunod. La photocopie non autorisée est un délit.

pièce de monnaie. On ne jette donc pas le dé si l'on a obtenu face « F ». Le référentiel E est constitué de 7 événements simples qui ne sont pas équiprobables :

$$E = \{F,(P,1),(P,2),(P,3),(P,4),(P,5),(P,6)\}.$$

Soit A l'événement « obtenir pile » : $A = \{(P,1),(P,2),(P,3),(P,4),(P,5),(P,6)\}$. On a

$$p(F) = 1/2 \text{ et } p(A) = 1/2 = p[(P,1)] + p[(P,2)] + \ldots + p[(P,6)].$$

$$\text{Donc } p[(P,1)] = p[(P,2)] = \ldots = p[(P,6)] = 1/12.$$

Le référentiel associé à une épreuve peut contenir soit un nombre fini d'éléments $E = \{e_1,e_2,\ldots,e_N\}$, soit un nombre infini dénombrable d'éléments $E = \{e_1,e_2,\ldots,e_N,\ldots\}$, soit un nombre infini et non dénombrable d'éléments (ainsi s'intéressant à la distribution de probabilité de la durée de vie X d'un appareil, on peut avoir pour référentiel $E = \mathbb{R}^+ = [0,\infty[$.

Section 3

COMPOSITION D'ÉVÉNEMENTS

1 Réalisation simultanée ou conjointe de deux événements

Exemple introductif

L'expérience consiste à jeter un dé 2 fois de suite et à relever le couple de nombres ainsi obtenu sur la face supérieure du dé. Le référentiel E est constitué de 36 événements simples : $E = \{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6),\ldots,(6,6)\}$.

Considérons les deux événements : A « le couple de nombres ainsi obtenu est tel que la somme des deux est $\leqslant 4$ », B « le premier nombre obtenu est égal à 1 et le second est un nombre impair ».

Les ensemble A et B sont constitués des couples

$$A = \{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)\} \ ; \ B = \{(1,1),(1,3),(1,5)\}.$$

A et B seront réalisés simultanément si et seulement si on obtient les couples (1, 1) ou (1, 3) : l'événement « A et B simultanément réalisés » $= \{(1,1),(1,3)\}$.

Plus généralement, l'événement « A **et** B simultanément réalisés » est constitué de l'ensemble des événements simples communs à A et à B. Pour cette raison on le note $A \cap B$ (où $\cap$ est le symbole de l'intersection des 2 parties A et B d'un même ensemble E). Par conséquent, la probabilité de réalisation conjointe est notée $p(A \cap B)$.

Dans l'exemple ci-dessus, on a $p(A \cap B) = p[(1,1)] + p[(1,3)]$.

Si le dé utilisé est homogène, on a $p[(1,1)] = p[(1,3)] = 1/36$ et $p(A \cap B) = 2 \times (1/36) \cong 0{,}055$.

Événements indépendants. Deux événements A et B d'un même référentiel E sont dits indépendants lorsque $p(A \cap B) = p(A) \times p(B)$. Dans ce cas, la réalisation de B ne dépend pas de A et vice-versa.

2 Réalisation d'un événement au moins

Soit deux événements A et B d'un même référentiel E. Au moins l'un de ces deux événements sera réalisé si et seulement si le résultat de l'expérience est un événement simple qui appartient à A **ou** à B, voire aux deux. L'événement « A **ou** B » est constitué de la réunion des événements simples qui appartiennent soit à A soit à B ou aux deux. Il est noté $A \cup B$ ($\cup$ étant le symbole de la réunion) et la probabilité correspondante est notée $p(A \cup B)$.

Dans l'exemple introductif : $A \cup B = \{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1),(1,5)\}$. Autrement dit, « A ou B sera réalisé » si et seulement si le résultat de l'expérience est un des 7 événements simples ci-dessus.

3 Événement contraire

Soit un événement A de E. L'événement contraire $\bar{A}$ est la *non réalisation* de A. $\bar{A}$ est constitué de l'ensemble d'éléments de E qui n'appartiennent pas à A.

On a : $A \cup \bar{A} = E$ et $A \cap \bar{A} = \varnothing$.

4 Formule de Poincaré

Théorème. Soit un référentiel E dont chaque partie A_i est munie d'une *probabilité* (dite aussi *poids*) $p(A_i)$. On a les propriétés suivantes :

i) $p(A) + p(\overline{A}) = 1$.

ii) $p(A_1 \cup A_2) = p(A_1) + p(A_2) - p(A_1 \cap A_2)$

iii) $p(A_1 \cup A_2 \cup A_3) = p(A_1) + p(A_2) + p(A_3) - p(A_1 \cap A_2) - p(A_1 \cap A_3) - p(A_2 \cap A_3) + p(A_1 \cap A_2 \cap A_3)$ etc.

Ainsi dans l'exemple introductif où le dé est supposé homogène on a : $p(A) = 6/36$, $p(B) = 3/36$ et $p(A \cap B) = 2/36$. Donc $p(A \cup B) = p(A) + p(B) - p(A \cap B) = 6/36 + 3/36 - 2/36 = 7/36$.

© Dunod. La photocopie non autorisée est un délit.

5 Implication, *i.e.* inclusion

A et B désignant des événements d'un même référentiel E, écrire $A \subset B$ signifie que la *réalisation de l'événement A entraîne automatiquement la réalisation de l'événement B*. En effet *A réalisé* signifie que la réalisation de l'expérience est un élément e_i de A. Pour que B soit automatiquement réalisé, il faut donc que cet élément e_i appartienne à B. Par suite, au sens de la théorie des ensembles, A est effectivement *inclus dans* B : $A \subset B$.

Section 4

PROBABILITÉS CONDITIONNELLES

1 Définition

Soit deux événements A et B d'un même référentiel E muni d'une distribution de probabilité. On nomme *probabilité conditionnelle de B relativement à A*, la probabilité pour que B se réalise sachant que A est réalisé. Elle se note $p_A(B)$ ou $p(B/A)$ et peut se calculer à partir de la relation

$$p(B/A) = \frac{p(A \cap B)}{p(A)}. \quad \textit{Remarquer que } p(A \cap B) = p(A) \times p(B/A).$$

Pour 3 évènements A, B et C on a $p(A \cap B \cap C) = p(A) \times p(B/A) \times p(C/A \cap B)$.

Exemple

Une urne contient 4 boules de même dimension portant respectivement les numéros 1, 2, 3 et 4. L'épreuve consiste à extraire successivement 2 boules (au hasard et sans remise de la première dans l'urne) puis à lire les numéros extraits. Les couples (1, 1), (2, 2), (3, 3) et (4, 4) ne pouvant apparaître, le référentiel E est constitué de 12 couples d'entiers équiprobables :

$$E = \{(1,2),(1,3),(1,4),(2,1),(2,3),(2,4),(3,1),(3,2),(3,4),(4,1),(4,2),(4,3)\}.$$

Considérons les événements : A « extraire la boule n°1 à l'issue du premier tirage », B « la somme des numéros des deux boules extraites est $\leqslant 4$ » puis calculons la probabilité pour que B se réalise sachant que A est réalisé, soit $p(B/A)$. On a

$$A = \{(1,2),(1,3),(1,4)\}, \quad B = \{(1,2),(1,3),(2,1),(3,1)\}, \quad A \cap B = \{(1,2),(1,3)\}.$$

donc $p(A) = 3 \times (1/12)$, $p(B) = 4 \times (1/12)$, $p(A \cap B) = 2 \times (1/12)$ et par suite $p(B/A) = p(A \cap B)/p(A) = 2/3$.

2 Formule de Bayes

Soit n événements $E_1, E_2, \ldots, E_n$ constituant une partition du référentiel E c'est-à-dire $E_1 \cup E_2 \cup \ldots \cup E_n = E$ et $E_i \cap E_j = \varnothing \; \forall \, i, j$ vérifiant $1 \leqslant i < j \leqslant n$ et soit A un événement quelconque de probabilité non nulle, on a alors :

$$p(E_{i_0}/A) = \frac{p(A/E_{i_0}) \times p(E_{i_0})}{\sum_{i=1}^{n} p(A/E_i) \times p(E_i)}$$

Justification.

On sait que $p(E_{i_0}/A) = p(E_{i_0} \cap A)/p(A)$. Or $p(E_{i_0} \cap A) = p(E_{i_0}) \times p(A/E_{i_0})$ et $p(A) = \sum_{i=1}^{n} p(A \cap E_i) = \sum_{i=1}^{n} p(A/E_i) \times p(E_i)$

Cas particulier. A et B désignant 2 événements d'un même référentiel E tels que $p(A) \neq 0$, on a $p(B/A) = \dfrac{p(A/B) \times p(B)}{p(A/B) \times p(B) + p(A/\bar{B}) \times p(\bar{B})}$.

REPÈRES : Espace probabilisable et probabilité

1 Espace probabilisable

Considérons un référentiel E (appelé aussi *univers des possibles*) et $P^*(E)$ l'ensemble des parties de E. Soit T est un sous-ensemble de $P^*(E)$ qui satisfait aux axiomes suivants :

(i) $E \in T$;

(ii) $\forall A \in T \Rightarrow \overline{A} \in T$; ($\overline{A}$ complémentaire de $\overline{A}$ dans E).

(iii) pour toute suite $\{A_n\}$ de parties de E appartenant à T, la réunion appartient à T : $\cup_n A_n \in T$.

Ces propriétés impliquent que $\cap_n A_n \in T$ et ce sous-ensemble T est appelé une tribu de parties de E. Le couple (E, T) s'appelle espace probabilisable.

2 Espace probabilisé

(E, T) étant un espace probabilisable, on appelle probabilité, toute application p de T dans [0, 1] qui satisfait aux axiomes suivants :

(j) $p(E) = 1$

(jj) $A \in T$, $B \in T$ et $A \cap B = \varnothing \Rightarrow p(A \cup B) = p(A) + p(B)$,

et plus généralement pour toute suite $\{A_n\}$ de parties de E appartenant à T et deux à deux disjointes on a $p\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} p(A_n)$.

Le triplet (E, T, p) est appelé espace probabilisé.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 345)

Exercice 1

Une compagnie d'assurance automobile a classé ses assurés en 4 classes d'âge :

classe 1 : moins de 26 ans ; classe 2 : de 26 à 45 ans, classe 3 : de 46 à 65 ans, classe 4 : plus de 65 ans.

Le tableau ci-dessous donne deux informations pour chaque classe :

– la proportion d'assurés appartenant à la classe,

– la probabilité qu'un assuré de la classe déclare au moins un accident au cours d'une année (probabilité obtenue par estimation à partir de l'étude statistique des accidents déclarés au cours des années précédentes).

Classe	Proportion d'assurés	Probabilité de déclarer au moins un accident
1	0,15	0,09
2	0,35	0,04
3	0,34	0,06
4	0,16	0,08

1. On prend un assuré au hasard *e* dans le fichier de la compagnie, quelle est la probabilité qu'il ait déclaré au moins un accident au cours de l'année ?

2. Quelle est la probabilité qu'un assuré de plus de 45 ans déclare au moins un accident en cours d'année ?

3. Quelle est la probabilité qu'un assuré qui a déclaré un accident au moins en cours d'année ait moins de 26 ans ?

4. Quelle est la probabilité qu'un assuré qui n'a pas déclaré d'accident ait plus de 45 ans ?

Exercice 2

Pour essayer de prévoir la défaillance des entreprises clientes de la banque *F*, l'économiste W. B. introduit le ratio *Z* défini pour chaque entreprise par le quotient de la marge brute d'autofinancement et des dettes totales. Les entreprises sont supposées courir de graves risques de défaillance lorsque le ratio est inférieur à une valeur critique c. Consignant les résultats portant sur un nombre très important d'entreprises, il observe que 5 % des entreprises sont défaillantes et que 95 % sont saines.

Dans la sous-population des entreprises défaillantes, 80 % des entreprises avaient au moment de l'octroi du crédit un ratio inférieur à c, 20 % un ratio supérieur à c. Dans la sous-population des entreprises saines, 10 % des entreprises avaient au moment de l'octroi du crédit un ratio inférieur à c, 90 % un ratio supérieur à c.

Quels sont les types d'erreurs que l'on peut commettre en situant une entreprise par référence à cette valeur critique c ? Calculer leurs probabilités respectives.

8 VARIABLES ALÉATOIRES RÉELLES

Une variable aléatoire réelle *v.a.*, souvent notée X, est le résultat numérique d'une expérience envisagée, expérience à laquelle est associée une probabilité de réalisation. Ces variables aléatoires continues ou discrètes, possèdent certaines propriétés générales exposées dans ce chapitre. Cette notion de *v.a.* s'applique notamment aux résultats de n expériences $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ réalisées dans des conditions identiques et constituant un échantillon.

Section 1 ■ Définition

Section 2 ■ Variables aléatoires discrètes

Section 3 ■ Variables aléatoires continues

Section 4 ■ Variables aléatoires du type $Y = \phi(X)$

Section 5 ■ Inégalités de Markoff et de Bienaymé-Tchébycheff

Section 6 ■ Variables aléatoires réelles indépendantes

Section 7 ■ Convergence en probabilité et en loi d'une suite Z_n de *v.a.* réelles

Section 8 ■ Échantillon iid d'une loi de probabilité

© Dunod. La photocopie non autorisée est un délit.

Section 1

DÉFINITION

Une variable aléatoire (en abrégé $v.a.$) réelle X est une application d'un référentiel E vers le domaine des réels $\mathbb{R}$, soit $X : E \longmapsto \mathbb{R}$, qui, à chaque résultat possible de l'expérience, associe un nombre réel. Autrement dit à chaque évènement simple e_i de E on fait correspondre sa « mesure numérique » $X(e_i) = x_i$.

Soit Δ l'image de E par $X : \Delta = X(E)$. Le référentiel E étant muni de la loi de probabilité p, X induit sur Δ une loi de probabilité qui sera notée P.

Exemple

On jette une pièce de monnaie, les résultats possibles étant pile P ou face F, le référentiel $E = \{P, F\}$. On peut définir la *v.a.* X qui prend la valeur 0 si on obtient face ou bien la valeur 1 si on obtient pile. Le domaine Δ_X des valeurs possibles de X est constitué des entiers 0 et 1: $\Delta_X = \{0,1\}$.
Lorsque la pièce est équilibrée on a $P(X = 1) = p(P) = 0{,}5$ et $P(X = 0) = p(F) = 0{,}5$.

Parmi les *v.a.* on distingue selon la forme prise par leur domaine des valeurs Δ_X, les variables discrètes des variables continues. Lorsque la *v.a.* X prend un nombre fini ou dénombrable de valeurs, ce qui est le cas dans les exemples précédents, on dit que X suit une loi discrète ou que X est une *v.a. discrète*.

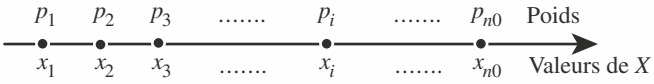
A contrario une variable aléatoire X qui peut prendre n'importe quelle valeur d'un certain intervalle et ne prend une valeur fixée *a priori* qu'avec une probabilité nulle est dite continue. Ainsi, dans l'exemple suivant, X suit une loi continue. On s'intéresse alors à la durée de vie X d'un certain type d'appareil. Si la durée de vie maximale est de 10 ans pour ce type d'appareil, X peut prendre n'importe quelle valeur comprise entre 0 et 10, soit $\Delta_X = [0; 10]$.

Section 2

VARIABLES ALÉATOIRES DISCRÈTES

1 Lois de probabilité discrètes

On appelle loi de probabilité d'une *v.a.* discrète X, la liste rangée par ordre croissant $(x_1, x_2, \ldots, x_{n_0})$ des valeurs qu'elle est susceptible de prendre accompagnée de la liste $(p_1, p_2, \ldots, p_{n_0})$ des probabilités correspondantes, nommée aussi *distribution de poids*, où $p_i \geqslant 0$ et $p_1 + p_2 + ... + p_{n_0} = 1$.



Autrement dit, X a pour domaine des valeurs possibles $\Delta_X = \{x_1, x_2, \ldots, x_{n_0}\}$ et $P(X = x_1) = p_1$, $P(X = x_2) = p_2, \ldots, P(X = x_i) = p_i, \ldots, P(X = x_{n_0}) = p_{n_0}$ (lire : la probabilité pour que X prenne la valeur x_i a pour valeur p_i).

Repères

Une loi de probabilité discrète est caractérisée par son support $\Delta = \{x_1, x_2, \cdots, x_{n_0}\}$ et sa distribution de poids $p_1, p_2, \cdots, p_{n_0}$, le poids de valeur p_i étant placé au point x_i.

Sa distribution est représentée par un diagramme en bâtons où la hauteur des bâtons associés à chaque valeur x_i est proportionnelle au poids p_i.

Le tableau ci-dessous permet de caractériser la loi de probabilité de X et de déterminer successivement $P(X \leqslant x_1), P(X \leqslant x_2), \ldots, P(X \leqslant x_{n_0})$.

Tableau 8.1 – *Distribution de probabilité*

Valeurs de X	x_1	x_2	x_3	...	x_{n_0}
$P(X = x_i)$	p_1	p_2	p_3	...	p_{n_0}
$P(X \leqslant x_i)$	p_1	$p_1 + p_2$	$p_1 + p_2 + p_3$	...	1

En effet, $(X \leqslant x_1) \Longleftrightarrow (X = x_1)$ et donc $P(X \leqslant x_1) = P(X = x_1) = p_1$. De même on a $(X \leqslant x_2) \Longleftrightarrow (X = x_1)$ *ou* $(X = x_2)$ et donc $P(X \leqslant x_2) = P(X = x_1) + P(X = x_2) = p_1 + p_2$, etc.

2 Fonction de répartition de X

La fonction de répartition $F(t)$, associée à la distribution de probabilité de la *v.a.* réelle X, est la fonction qui à chaque réel t, associe la valeur $P(X \leqslant t)$:

$$F(t) = P(X \leqslant t) \quad \forall\, t \in \mathbb{R}.$$

Autrement dit, $F(t)$ est égal au poids porté par l'intervalle $]-\infty, t]$:

$$F(t) = p(]-\infty, t])$$

Dans le cas d'une loi discrète, $F(t)$ est une *fonction en escalier* puisque :

$$F(t) = \begin{cases} 0 & \text{si} \quad t < x_1 \\ \displaystyle\sum_{i=1}^{h} p_i & \text{si} \quad x_h \leqslant t < x_{h+1} \ \forall h = 1,2,\ldots,(n_0 - 1) \\ 1 & \text{si} \quad t \geqslant x_{n_0} \end{cases}$$

© Dunod. La photocopie non autorisée est un délit.

Exemple

Soit X le nombre aléatoire de véhicules de tourisme loués dans une journée par une petite agence de centre ville. X est une variable aléatoire qui peut prendre les valeurs 0, 1, 2, 3 ou 4 avec les probabilités correspondantes indiquées ci-après :

Valeurs x_i	0	1	2	3	4
Probabilités individuelles p_i	0,1	0,25	0,3	0,2	0,15
Probabilités cumulées $P(X \leqslant x_i)$	0,1	0,35	0,65	0,85	1

Sa fonction de répartition qui est une fonction en escalier, est définie comme suit :

$F(t) = 0 \quad \forall t < 0$; $F(t) = 0{,}1 \quad \forall t$ vérifiant $0 \leqslant t < 1$; $F(t) = 0{,}35 \quad \forall t$ vérifiant $1 \leqslant t < 2$; $F(t) = 0{,}65 \quad \forall t$ vérifiant $2 \leqslant t < 3$, $F(t) = 0{,}85 \quad \forall t$ vérifiant $3 \leqslant t < 4$, $F(t) = 1 \quad \forall t \geqslant 4$.

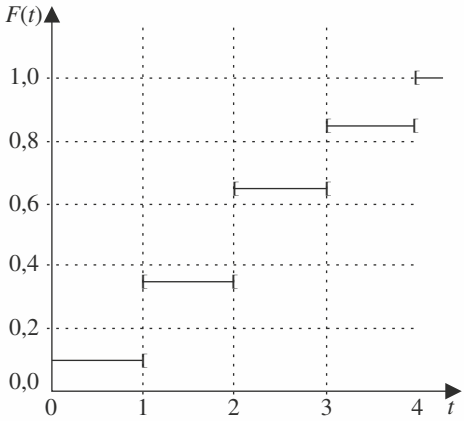


Figure 8.1 – *Diagramme en escalier de la fonction de répartition*

3 Valeur moyenne *m* ou espérance mathématique *E*(*X*)

La moyenne de la loi que suit X, généralement désignée par la lettre m, est égale à la moyenne des valeurs possibles de X pondérées par leur poids respectifs. Elle est également appelée espérance mathématique de X et est alors notée $E(X)$:

$$m = E(X) = \sum_{i=1}^{n_0} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_{n_0} p_{n_0}$$

Plus généralement, r désignant un entier naturel quelconque, on définit le *moment d'ordre r* noté $m_r(X)$, expression également notée $E(X^r)$:

$$m_r(X) = E(X^r) = \sum_{i=1}^{n_0} x_i^r p_i = (x_1)^r p_1 + (x_2)^r p_2 + \ldots + (x_{n_0})^r p_{n_0}$$

Ainsi, $E(X^2) = \sum_{i=1}^{n_0} x_i^2 p_i = (x_1)^2 p_1 + (x_2)^2 p_2 + \ldots + (x_{n_0})^2 p_{n_0}$.

4 Variance $V(X)$ et autres moments centrés μ_r

La variance de X notée $V(X)$ mesure la dispersion de la distribution autour de sa valeur moyenne m :

$$V(X) = E[(X-m)^2] = \sum_{i=1}^{n_0}(x_i - m)^2 p_i$$
$$= (x_1 - m)^2 p_1 + (x_2 - m)^2 p_2 + \ldots + (x_{n_0} - m)^2 p_{n_0}$$

Plus généralement, r désignant un entier naturel quelconque, on définit le *moment centré d'ordre r* noté μ_r :

$$\mu_r = E[(X-m)^r] = \sum_{i=1}^{n_0}(x_i - m)^r p_i. \quad \text{Ainsi } V(X) = \mu_2.$$

Pour calculer la variance de X, on utilise le plus souvent la formule de *Koenig* :

$$V(X) = E(X^2) - [E(X)]^2.$$

La racine carrée de la variance est appelée ***écart-type*** et est notée $\sigma(X)$: $\sigma(X) = \sqrt{V(X)}$.

Exemple précédent

On a $m = E(X) = 0 \times 0{,}1 + 1 \times 0{,}25 + 2 \times 0{,}3 + 3 \times 0{,}2 + 4 \times 0{,}15 = 2{,}05$
et $E(X^2) = 0^2 \times 0{,}1 + 1^2 \times 0{,}25 + 2^2 \times 0{,}3 + 3^2 \times 0{,}2 + 4^2 \times 0{,}15 = 5{,}65$

$$V(X) = E(X^2) - [E(X)]^2 = 5{,}65 - 2{,}05^2 = 1{,}4475$$

Valeur du moment centré d'ordre 4 : $\mu_4 = E[(X-m)^4] = \sum_{i=1}^{n_0}(x_i - m)^4 p_i = (0-2{,}05)^4$
$\times 0{,}1 + (1-2{,}05)^4 \times 0{,}25 + (2-2{,}05)^4 \times 0{,}3 + (3-2{,}05)^4 \times 0{,}2 + (4-2{,}05)^4 \times 0{,}15 = 4{,}40.$

5 Médiane, p-quantiles d'une loi discrète

Notons X une *v.a.* qui suit la loi discrète $\mathcal{L}$ caractérisée par son support

$$\Delta = \{x_1, x_2, \ldots, x_{n_0}\} \text{ (où } x_1 < x_2 < \ldots < x_{n_0})$$

et par la distribution de poids : $p(x_1) = p_1, p(x_2) = p_2, \ldots, p(x_{n_0}) = p_{n_0}$.

Médiane. Grosso modo, la valeur médiane m_e de la loi $\mathcal{L}$ partage $\mathbb{R}$ en deux parties de même poids : $p(]-\infty, m_e]) \cong 0{,}5$ et $p([m_e, \infty[) \cong 0{,}5$; autrement dit

$$P(X \leqslant m_e) \cong 0{,}5 \text{ et } P(X \geqslant m_e) \cong 0{,}5.$$

© Dunod. La photocopie non autorisée est un délit.

Deux possibilités se présentent :

– s'il existe une valeur x_{i_0} de Δ telle que « $P(X \leqslant x_{i_0}) = 0{,}5$ et conséquemment $P(X \geqslant x_{i_0+1}) = 0{,}5$ » alors $(x_{i_0} + x_{i_0+1})/2$ est considérée comme valeur médiane ;

– s'il existe une valeur x_{i_0} de Δ telle que « $P(X \leqslant x_{i_0}) > 0{,}5$ et $P(X \geqslant x_{i_0}) > 0{,}5$ » alors x_{i_0} est appelée valeur médiane de la distribution.

Exemple

Dans l'exemple précédent, $x_{i_0} = 2$ est la valeur médiane de la distribution car $P(X \leqslant 2) = 0{,}65 > 0{,}5$ et $P(X \geqslant 2) = 0{,}65 > 0{,}5$.

p-quantiles ξ_p. Le raisonnement s'applique plus généralement aux *p-quantiles* (où p désigne un nombre réel vérifiant $0 < p < 1$).

Grosso modo, la valeur du *p-quantile* ξ_p de la loi discrète $\mathcal{L}$ partage $\mathbb{R}$ en deux parties dont les poids sont respectivement p et $(1 - p)$:

$p(]-\infty, \xi_p]) \cong p$ et $p([\xi_p, \infty[) \cong 1 - p$; autrement dit $P(X \leqslant \xi_p) \cong p$ et $P(X \geqslant \xi_p) \cong 1 - p$.

Deux cas sont à envisager :

– s'il existe une valeur x_{i_0} de Δ telle que « $P(X \leqslant x_{i_0}) = p$ » alors $\xi_p = (x_{i_0} + x_{i_0+1})/2$;

– s'il existe une valeur x_{i_0} telle que « $P(X \leqslant x_{i_0}) > p$ et $P(X \geqslant x_{i_0}) > 1 - p$ alors $\xi_p = x_{i_0}$.

$\xi_{0,25}$ et $\xi_{0,75}$ sont respectivement appelés premier et troisième quartile puisque $0{,}25 = 1/4$ et $0{,}75 = 3/4$. Ainsi dans l'exemple précédent, on constate que le premier quartile $\xi_{0,25}$ de la distribution de X est $x_1 = 1$. En effet $P(X \leqslant 1) = 0{,}35 > 0{,}25$ et $P(X \geqslant 1) = 0{,}9 > (1 - 0{,}25)$

Section 3

VARIABLES ALÉATOIRES CONTINUES

1 Lois de probabilité continues dont le support est un intervalle

Une *v.a. X* est dite *continue* lorsqu'elle peut prendre n'importe quelle valeur d'un intervalle donné (a,b) appelé domaine des valeurs possibles : $\Delta_X = (a,b)$.

La distribution de probabilité associée est caractérisée par une fonction $f(x)$ appelée « *densité de probabilité* » :

quelque soit l'intervalle $[\alpha,\beta]$ inclus dans l'intervalle (a,b) (avec $\alpha < \beta$) on a $P(\alpha \leqslant X \leqslant \beta) = \int_{\alpha}^{\beta} f(x)\mathrm{d}x =$ « *aire du domaine D hachuré délimité par les droites* $x = \alpha$, $x = \beta$ *et la courbe* $y = f(x)$ » lorsque le repère est orthonormé. Voir figure 8.2.

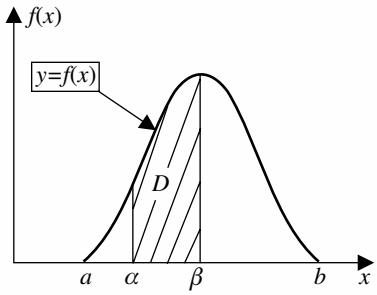


Figure 8.2 – *Représentation d'une fonction de densité* $f(x)$

Cette fonction de densité $f(x)$ doit satisfaire aux conditions suivantes :

i) $f(x) \geqslant 0 \quad \forall\, x \in (a,b)$ [éventuellement nulle en des points isolés] ;

ii) $\int_{a}^{b} f(x)\mathrm{d}x = 1$. Cette condition résulte du fait que l'événement $(a < X < b)$ étant un *événement certain*, on doit nécessairement avoir $1 = P(a < X < b)$.

Remarque. δ désignant un nombre quelconque du domaine des valeurs possibles $\Delta_X = (a,b)$ on a $P(X = \delta) = 0$. Autrement dit, la probabilité pour qu'une *v.a.* continue X prenne une valeur donnée, fixée *a priori*, est nulle. Par suite $P(\alpha \leqslant X \leqslant \beta) = P(\alpha < X \leqslant \beta) = P(\alpha \leqslant X < \beta) = P(\alpha < X < \beta)$.

En effet, $(\alpha \leqslant X \leqslant \beta) = (\alpha \leqslant X < \beta) \cup (X = \beta)$ implique $P(\alpha \leqslant X \leqslant \beta) = P(\alpha \leqslant X < \beta) + P(X = \beta) = P(\alpha \leqslant X < \beta) + 0$. Etc.

Exemple

Soit une *v.a.* réelle X qui peut prendre n'importe quelle valeur de l'intervalle [0, 2] avec une densité de probabilité constante : $f(x) = k \quad \forall\, x \in [0,2]$. Sachant que $\int_{a}^{b} f(x)\mathrm{d}x = 1$ lorsque $\Delta_X = [a,b]$ on en déduit la valeur de la constante k : $1 = \int_{0}^{2} k\mathrm{d}x = 2k$. Donc $k = 0{,}5$ et par suite $f(x) = 0{,}5 \quad \forall\, x \in [0,2]$.

2 Valeur moyenne *m* ou espérance mathématique *E*(*X*)

Soit une *v.a.* X qui suit la loi continue $\mathcal{L}$ caractérisée par le domaine des valeurs $\Delta =]a,b[$ et la densité de probabilité $f(x)$. Selon la même terminologie que celle présentée dans le cas discret on a :

$$m = E(X) = \int_{a}^{b} xf(x)\mathrm{d}x.$$

© Dunod. La photocopie non autorisée est un délit.

De même, est défini le moment d'ordre r de la loi que suit X, également appelé espérance mathématique de X^r :

$$m_r = E(X^r) = \int_a^b x^r f(x)\mathrm{d}x. \text{ Ainsi, } E(X^2) = \int_a^b x^2 f(x)\mathrm{d}x.$$

3 Variance $V(X)$ et moments centrés μ_r

$V(X) = \int_a^b (x-m)^2 f(x)\mathrm{d}x = E(X^2) - [E(X)]^2$ est la variance de la *v.a.* continue X, $\sigma(X) = \sqrt{V(X)}$ est l'écart-type de X,

$\mu_r = E[(X-m)^r] = \int_a^b (x-m)^r f(x)\mathrm{d}x$ est le moment centré d'ordre r.

Exemple

Dans l'exemple précédent on a : $E(X) = \int_0^2 x \times (0{,}5)\mathrm{d}x = 1$; $E(X^2) = \int_0^2 x^2 \times (0{,}5)\mathrm{d}x = 4/3$. Utilisant la formule de Koenig l'on en déduit que $V(X) = E(X^2) - [E(X)]^2 = 4/3 - 1^2 = 1/3$.

4 Fonction de répartition

Notons F la fonction de répartition de la loi que suit X :

$$F(t) = P(-\infty < X \leqslant t) \quad \forall t \in \mathbb{R}.$$

Autrement dit, la répartition de la masse de poids total égal à 1 étant continûment étalée sur l'intervalle $\Delta = (a,b)$ avec une densité de répartition $f(x)$, la fonction de répartition $F(t)$ est égale au poids porté par l'intervalle $]-\infty,t]$:

$$F(t) = p(]-\infty,t])$$

REPÈRES : Propriétés

La fonction de répartition F de la loi continue satisfait aux propriétés suivantes :

i) elle est strictement croissante sur $\Delta =]a, b[$,

ii) $F(t) = \begin{cases} 0 & \forall\, t \leqslant a \text{ lorsque } a \text{ a une valeur finie (ou } F(t) \xrightarrow[t \to -\infty]{} 0 \text{ si } a = -\infty) \\ \int_a^t f(x)\mathrm{d}x & \forall\, t \in]a, b[\\ 1 & \forall t \geqslant b \text{ lorsque } b \text{ a une valeur finie (ou } F(t) \xrightarrow[t \to +\infty]{} 1 \text{ si } b = +\infty) \end{cases}$

iii) sur $]a; b[$ la densité de probabilité f est la dérivée de la fonction de répartition : $F'(t) = f(t)\ \forall t \in]a; b[$.

Remarque. La distribution des probabilités de la loi que suit X est entièrement caractérisée par la connaissance de la fonction de répartition F.

En effet, $\forall\, u$ et v réels tels que $u < v$, on a $P(u < X \leqslant v) = F(v) - F(u)$.

Cette propriété remarquable, valable pour les lois discrètes ou continues est à la base de l'utilisation des tables numériques.

Exemple

Ci-après la représentation graphique de la fonction de répartition de la loi précédemment étudiée et qui est caractérisée par « son support $\Delta = [0,2]$ et sa densité de probabilité (c'est-à-dire de répartition massique) $f(x) = 0{,}5 \quad \forall x \in \Delta$ ». En effet, on a

$$F(t) = 0 \quad \forall t < 0,\ F(t) = \int_0^t (0{,}5)\mathrm{d}x = t/2 \quad \forall t \in [0,2],\ \ F(t) = 1\ \forall\, t > 2.$$

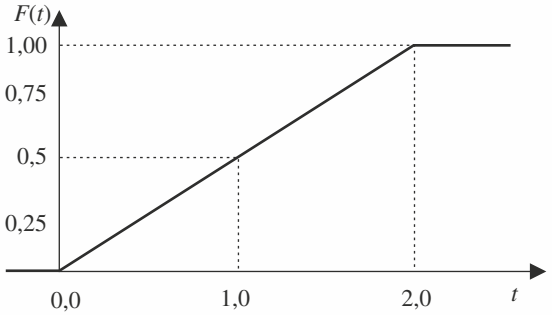


Figure 8.3 – *Représentation de la fonction de répartition* $F(t)$

5 Médiane, *p*-quantiles d'une loi continue

Notons X une *v.a.* qui suit la loi continue $\mathcal{L}$ caractérisée par son support Δ qui est un intervalle (a,b) et par la fonction de répartition F.

Médiane. La médiane m_e de la loi $\mathcal{L}$ est la solution de l'équation $F(m_e) = 0{,}5$.

Elle partage $\mathbb{R}$ en deux parties de même poids : $p(]-\infty, m_e]) = p([m_e, \infty[) = 0{,}5$; autrement dit, $P(X \leqslant m_e) = 0{,}5$ et $P(X \geqslant m_e) = 0{,}5$.

p-quantile ξ_p. La valeur du *p-quantile* ξ_p de la loi $\mathcal{L}$ est la solution de l'équation « $F(\xi_p) = p$ ». Elle partage $\mathbb{R}$ en deux parties dont les poids sont respectivement p et $1 - p$: $p(]-\infty, \xi_p]) = p$ et $p([\xi_p, \infty[) = 1 - p$.

Autrement dit $P(X \leqslant \xi_p) = p$ et $P(X \geqslant \xi_p) = 1 - p$.

Exemple

Dans l'exemple précédent, la valeur médiane est 1 car cherchant t_0 tel que $0{,}5 = F(t_0)$ on a $0{,}5 = t_0/2$ soit $t_0 = 1$; le premier quartile $\xi_{0,25}$ est égal à 0,5 car cherchant t_0 tel que $0{,}25 = F(t_0)$ on a $0{,}25 = t_0/2$ soit $t_0 = 0{,}5$ (*cf.* figure 8.3).

© Dunod. La photocopie non autorisée est un délit.

Section 4

VARIABLES ALÉATOIRES DU TYPE $Y = \phi(X)$

Soient une *v.a.* réelle X discrète ou continue et une fonction ϕ définie sur Δ_X et à valeur réelle. À chaque valeur x que peut prendre X à l'issue de l'expérimentation on peut faire correspondre la valeur $y = \phi(x)$ que peut prendre la *v.a.* $Y = \phi(X)$. La loi de Y se déduit de celle de X.

Exemples

1/ Soit X une *v.a.* prenant les valeurs – 1, 0 ou 1 avec les probabilités respectives 0,2, 0,3 et 0,5. La *v.a.* $Y = 2X + 1$ peut prendre les valeurs – 1, 1 ou 3 avec les probabilités respectives 0,2 , 0,3 et 0,5.

2/ Soit X une *v.a.* qui peut prendre n'importe quelle valeur réelle positive, autrement dit $\Delta_X = [0,\infty[$ et dont la distribution de probabilités est caractérisée par la fonction de répartition F définie ci-après : $F(t) = 0 \quad \forall\, t < 0$, $F(t) = 1 - \mathrm{e}^{-t}\ \forall t \geqslant 0$.

Dans ce cas la *v.a.* $Y = 2X + 1$ peut prendre n'importe quelle valeur de l'intervalle $[1,\infty[$: $\Delta_Y = [1,\infty[$.

G désignant la fonction de répartition de Y, on a pour $t \in \Delta_Y = [1,\infty[$:

$G(t) = P(Y \leqslant t) = P(1 \leqslant Y \leqslant t) = P(1 \leqslant 2X + 1 \leqslant t) = P(0 \leqslant X \leqslant (t-1)/2) = F\left(\dfrac{t-1}{2}\right) - F(0) = 1 - \mathrm{e}^{-(t-1)/2}$. La fonction de densité $g(t)$ de la loi que suit Y s'obtient en dérivant $G(t)$: $g(t) = (1/2)\mathrm{e}^{-(t-1)/2}$.

– Lorsque X suit une loi discrète caractérisée par son domaine des valeurs $\Delta_X = \{x_1, x_2, \ldots, x_{n_0}\}$ (éventuellement $n_0 = \infty$) et $P(X = x_i) = p_i$, on a

$$E(Y) = E[\phi(X)] = \Sigma_i \phi(x_i)\, p_i.$$

– Lorsque X suit une loi continue caractérisée par son domaine de valeurs $\Delta_X = (a,b)$ et la densité de probabilité $f(x)$, on a

$$E(Y) = E[\phi(X)] = \int_a^b \phi(x)\, f(x)\mathrm{d}x$$

REPÈRES : Propriété

Si à une *v.a.* réelle X discrète ou continue on associe la *v.a.* $Y = cX + d$ où c et d sont deux réels donnés on a alors :

i) $E(Y) = E(cX + d) = cE(X) + d$ *ii)* $V(Y) = V(cX + d) = V(cX) = c^2 V(X)$.

Section 5

INÉGALITÉS DE MARKOFF ET DE BIENAYMÉ-TCHEBYCHEFF

1 Inégalité de Markoff

Soit une *v.a. X* qui ne peut prendre que des valeurs positives ou nulles c'est-à-dire dont le domaine des valeurs Δ_X est inclus dans l'intervalle $[0,\infty[$. Alors quelque soit la constante λ vérifiant $\lambda > 1$, on a : $P[X \geqslant \lambda E(X)] \leqslant 1/\lambda$.

2 Inégalité de Bienaymé-Tchébycheff

Soit une *v.a. X* qui suit une loi de valeur moyenne m et d'écart type σ. Quelque soit le choix d'une constante λ vérifiant $\lambda > 1$, on a les inégalités suivantes :

i) $P(|X - m| \geqslant \lambda\sigma) \leqslant 1/\lambda^2$,

ii) $P(|X - m| < \lambda\sigma) \geqslant 1 - 1/\lambda^2 \iff P(m - \lambda\sigma < X < m + \lambda\sigma) \geqslant 1 - 1/\lambda^2$.

Prenant par exemple $\lambda = 10$, l'inégalité *ii)* montre que la probabilité pour que X prenne une valeur appartenant à l'intervalle $]m - 10\sigma, m + 10\sigma[$ est supérieure à 99 %. ***Justification.*** Appliquer l'inégalité de Markoff à la *v.a.* $Y = (X - m)^2$ et remarquer que $E(Y) = V(X)$.

Section 6

VARIABLES ALÉATOIRES INDÉPENDANTES

1 Couple de *v.a.* réelles indépendantes

Deux *v.a.* réelles X et Y sont dites indépendantes lorsque quelque soient les réels α_1, α_2, β_1, β_2 tels que $\alpha_i \leqslant \beta_i$, on a (*cf.* chapitre 7, page 105) :

$$P(\alpha_1 < X \leqslant \beta_1 \text{ et } \alpha_2 < Y \leqslant \beta_2) = P(\alpha_1 < X \leqslant \beta_1) \times P(\alpha_2 < Y \leqslant \beta_2).$$

De façon équivalente, X et Y indépendantes

$\iff$ quelque soient les réels x et y, $P(X \leqslant x \text{ et } Y \leqslant y) = P(X \leqslant x) \times P(Y \leqslant y)$

2 *n*-uple de *v.a.* réelles indépendantes

n v.a. réelles $X_1, X_2, \ldots, X_n$ sont dites indépendantes lorsque quelque soient les réels $\alpha_1, \ldots, \alpha_n$, $\beta_1, \ldots, \beta_n$ tels que $\alpha_i \leqslant \beta_i$, on a :

$$P(\alpha_1 < X_1 \leqslant \beta_1 \text{ et } \alpha_2 < X_2 \leqslant \beta_2 \text{ et } \ldots \text{ et } \alpha_n < X_n \leqslant \beta_n) = P(\alpha_1 < X_1 \leqslant \beta_1)$$
$$\times P(\alpha_2 < X_2 \leqslant \beta_2) \times \ldots \times P(\alpha_n < X_n \leqslant \beta_n)$$

© Dunod. La photocopie non autorisée est un délit.

REPÈRES : Propriétés

$X_1, X_2, \cdots, X_n$ désignant *n v.a.* réelles quelconques on a les propriétés suivantes

i) $E(X_1 + X_2 + \cdots X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$

ii) $\sigma(X_1 + X_2 + \cdots X_n) \leqslant \sigma(X_1) + \sigma(X_2) + \cdots + \sigma(X_n)$

iii) $X_1, X_2, \cdots, X_n$ indépendants $\Rightarrow V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n)$ autrement dit, « la variance de la somme est égale à la somme des variances ».

Section 7

CONVERGENCE EN PROBABILITÉ ET EN LOI D'UNE SUITE Z_n DE V.A. RÉELLES

Convergence en probabilité. On dit que la suite de *v.a.* réelles Z_n converge en probabilité vers un nombre réel donné k lorsque, quelque soit le choix de l'intervalle $[k-\varepsilon, k+\varepsilon]$ et quelque soit le choix d'une forte probabilité θ on a :

$$P(k-\varepsilon \leqslant Z_n \leqslant k+\varepsilon) \geqslant \theta \text{ dès que } n \text{ est assez grand, soit } \forall n \geqslant n_0(\varepsilon,\theta).$$

« Z_n converge en probabilité vers le nombre k lorsque $n \longrightarrow \infty$ » s'écrit : $Z_n \xrightarrow[n\to\infty]{pr} k$. En choisissant par exemple $\varepsilon = 10^{-3}$ et $\theta = 0{,}99$, on a $P(k-0{,}001 \leqslant Z_n \leqslant k+0{,}001) \geqslant 0{,}99$ dès que n est assez grand.

Théorème. Si $E(Z_n) \longrightarrow k$ et $V(Z_n) \longrightarrow 0$ lorsque $n \longrightarrow \infty$, alors $Z_n \xrightarrow[n\to\infty]{pr} k$.

Notamment si $E(Z_n) = k\ \forall n$ et $V(Z_n) \longrightarrow 0$ lorsque $n \longrightarrow \infty$, alors $Z_n \xrightarrow[n\to\infty]{pr} k$.

Convergence en loi. Soit une suite de *v.a* réelles Z_n et soit $F_n(t) = P(Z_n \leqslant t)$ la fonction de répartition de Z_n. On dit que la suite Z_n converge en loi vers la *v.a.* Z lorsque pour $n \to \infty$, on a $F_n(t) \to F(t) = P(Z \leqslant t)$ en tout point t où F est continue. On écrit : $Z_n \xrightarrow{loi} Z$.

Soit un domaine δ constitué d'une réunion finie ou dénombrable de points et d'intervalles de $\mathbb{R}$. Alors, $P(Z_n \in \delta) \to P(Z \in \delta)$ lorsque $n \to \infty$.

Théorème de Rao. Soit une suite de *v.a.* réelles Z_n telles que $Z_n \xrightarrow{loi} Z$ et soit une fonction g continue sur $\mathbb{R}$. Alors, $g(Z_n) \xrightarrow{loi} g(Z)$

Théorème de Slutsky. Soit une suite de *v.a.* réelles X_n telles que $X_n \xrightarrow{\text{loi}} X$ et soit une suite de *v.a.* réelles Y_n telles que $Y_n \xrightarrow{\text{pr}} a$ (a réel). Alors, *i)* $X_n Y_n \xrightarrow{\text{loi}} aX$, *ii)* $X_n + Y_n \xrightarrow{\text{loi}} X + a$. *iii)* Si $a \neq 0$, « $1/Y_n \xrightarrow{\text{pr}} 1/a$ ». X_n et Y_n ne sont pas nécessairement indépendants.

Section 8

ÉCHANTILLON *iid*

1 Définition

À n épreuves $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ que l'on se propose de réaliser dans des conditions identiques, sont respectivement associées les *v.a.* $X_1, X_2, \ldots, X_n$. En raison du fait que les expériences sont réalisées dans des conditions identiques, il en résulte que les résultats X_i sont indépendants et suivent la même loi de probabilité, autrement dit, sont ***i***ndépendants et ***i***dentiquement ***d***istribués (en abrégé, ***iid***)[1]

2 Moyenne $\bar{X}$ d'un échantillon *iid*

Considérons la moyenne $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$ des n *v.a.* $X_1, X_2, \ldots, X_n$ associées à n épreuves que l'on se propose de réaliser dans des conditions identiques. Les n *v.a.* $X_1, \ldots, X_n$ suivant une même loi de valeur moyenne m_0 et d'écart-type σ_0, on a $E(X_i) = m_0$ et $V(X_i) = \sigma_0^2 \quad \forall i = 1, \ldots, n,$ et par suite

$$E(\bar{X}) = m_0 \text{ et } V(\bar{X}) = \sigma_0^2/n.$$

Justifications. Soit $Z = X_1 + X_2 + \ldots X_n$. On a $E(Z) = E(X_1 + \ldots X_n) = E(X_1) + \ldots E(X_n) = n \times m_0$, donc $E(\bar{X}) = E[(1/n)Z] = (1/n) \times E(Z) = m_0$. De même, $V(Z) = V(X_1 + \ldots X_n) = V(X_1) + \ldots + V(X_n) = n \times \sigma_0^2$, car $X_1, X_2, \ldots, X_n$ indépendantes donc $V(\bar{X}) = V[(1/n)Z] = (1/n)^2 V(Z) = \sigma_0^2/n$.

Propriété. Lorsque $n \longrightarrow \infty$, la moyenne aléatoire de l'échantillon $\{X_1, X_2, \ldots, X_n\}$ converge en probabilité vers la moyenne m_0 de la loi suivie par les *v.a.* X_i : $\bar{X} \xrightarrow{\text{pr}} m_0$.

Justification. *Cf.* théorème p. 120 et remarquer que $E(\bar{X}) = m_0$ et $V(\bar{X}) = \sigma_0^2/n \longrightarrow 0$ si $n \longrightarrow \infty$.

Remarque. En appliquant l'inégalité *ii)* de Bienaymé-Tchébycheff à la *v.a.* $\bar{X}$ avec $\lambda = 10$ on obtient $P(|\bar{X} - m_0| < 10 \times \sigma_0/\sqrt{n}) \geqslant 0{,}99$. Lorsque le nombre n d'expériences est grand, le nombre $10\sigma_0/\sqrt{n}$ est petit, autrement dit il y a une probabilité supérieure à 99 % pour que $\bar{X}$ prenne une valeur proche de m_0.

1. Il convient de ne pas confondre ces n mesures $X_1, \ldots, X_n$ d'un même caractère, associées à n expériences réalisées dans des conditions identiques et un n-uple $(X_1, \ldots, X_n)$ associé à une seule expérience et qui mesure n caractères distincts.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 347)

Exercice 1

Le responsable d'un rayon de bricolage dans une grande surface estime que la demande aléatoire journalière X de perceuses PB se situe entre 0 et 5 avec les probabilités suivantes

Demande	0	1	2	3	4	5
Probabilités	0,05	0,15	0,30	0,30	0,15	0,05

1. Représenter la distribution de probabilité et déterminer sa fonction de répartition.

2. Calculer $E(X)$ et $V(X)$ puis déterminer la valeur médiane.

3. Si le responsable du rayon a en stock en début de journée 2 perceuses, quelle est la probabilité qu'il y ait rupture de stock ?

4. Pour être sûr (à plus de 90 %) de servir tous les clients, combien doit-il avoir au minimum de perceuses en stock en début de journée ?

Exercice 2

Soit X une variable aléatoire réelle qui peut prendre n'importe quelle valeur de l'intervalle (0, 3) avec une densité de probabilité constante k.

1. Représenter le graphe de la densité de probabilité de la loi que suit X puis celui de la fonction de répartition F. Calculer $E(X), V(X)$ et $F(2)$.

2. Soit $Y = 2X + 3$. Déterminer la fonction de répartition G de Y puis calculer $E(Y)$ et $V(Y)$.

QCM 1. Une *v.a.* X prend les valeurs – 1 ; 0 ; 1 avec les probabilités respectives 0,2, 0,6, 0,2. Alors sa variance vaut : ① 0 ② 1 ③ 0,4 ④ 0,8 ⑤ aucune réponse ne convient.

QCM 2. Soit X une variable aléatoire, V désigne la variance. Alors $V(-2X + 4) =$
① $-2V(X) + 4$ ② $4V(X) + 16$ ③ $4V(X)$ ④ $-2V(X)$ ⑤ aucune réponse ne convient.

QCM 3. Une variable aléatoire X prend les valeurs 1, 2 et 3, chacune avec la probabilité 1/3. Alors sa variance vaut : ① 2 ② 5/4 ③14/3 ④ 2/3 ⑤ aucune réponse ne convient.

9 LES PRINCIPALES LOIS DE PROBABILITÉS

L'objet de ce chapitre est de présenter les principales lois de probabilité permettant une modélisation mathématique du problème étudié.

Par la suite X désignera une *v.a.* réelle qui suit la loi de probabilité considérée.

Chacune des *lois discrètes* présentées ci-après sera caractérisée par :

– son support $\Delta = \{x_1, x_2, \ldots, x_{n_0}\}$ (éventuellement, $n_0 = \infty$),

– la distribution de poids $p(x_i) = p_i$, autrement dit la connaissance des nombres $p_i = P(X = x_i)$.

Les *lois continues* étudiées sont quant à elles caractérisées par :

– un intervalle $\Delta_X = (a,b)$ qui est le domaine des valeurs de X

– la définition de la fonction de densité $f(x)$ ou de répartition $F(x)$ pour $x \in \Delta_X$.

En raison de son importance dans le domaine des applications, l'étude de la loi normale centrée est abordée prioritairement.

Section 1	■	Lois Normales (qui sont des lois continues)
Section 2	■	Lois discrètes : de Bernoulli, Binomiales, Géométriques, Hypergéométriques, de Poisson, des rangs signés de Wilcoxon
Section 3	■	Lois continues (suite) : Khi-deux, Student-Fisher, Gamma, exponentielles, uniformes, Fisher-Snédécor, Cauchy, Bêta, Logistique, Log-Normale, Weibull
Section 4	■	Lois multinomiales
Section 5	■	Procédures avec Excel, SPSS

© Dunod. La photocopie non autorisée est un délit.

Section 1

LOIS NORMALES

1 Loi normale centrée réduite $\mathcal{N}(0;1)$

Une *v.a.* X qui suit la loi normale centrée réduite, « en abrégé, $X \sim> \mathcal{N}(0\,;1)$ » est souvent notée $\mathcal{N}_{0;1}$.

1.1 Caractérisation

Son support ou domaine des valeurs est $\Delta =]-\infty,\infty[$; sa densité de probabilité est $f(x) = \dfrac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\ \forall\, x \in \mathbb{R}$.

La fonction de densité est paire puisque $f(-x) = f(x)\ \forall\, x$, aussi il existe une symétrie par rapport à l'axe des ordonnées.

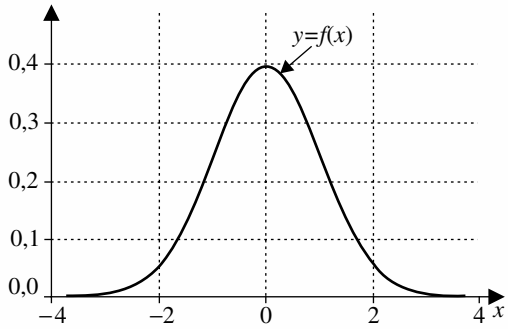


Figure 9.1 – *Fonction de densité de la loi normale centrée réduite* $\mathcal{N}(0;\,1)$

La valeur moyenne m de cette distribution est nulle et sa variance σ^2 est égale à 1. On a :

$$E(X) = m = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = 0,\ V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-m)^2 f(x)\mathrm{d}x = 1\,.$$

1.2 Fonction de répartition

Sa fonction de répartition $\Phi(t) = P(X \leqslant t) = \displaystyle\int_{-\infty}^{t} f(x)\mathrm{d}x$ est tabulée page 374 pour diverses valeurs de t positives. Par suite, $\forall\, \alpha$ et β réels vérifiant $\alpha < \beta$, on a : $P(\alpha < X \leqslant \beta) = \Phi(\beta) - \Phi(\alpha)$.

Cet expression est valable pour $\alpha = -\infty$ en posant $\Phi(\alpha) = 0$ et pour $\beta = +\infty$ en posant $\Phi(\infty) = 1$. On a aussi $P(\alpha < X < \beta) = \Phi(\beta) - \Phi(\alpha)$ puisque $P(X = \beta) = 0$.

Propriété. En raison de la symétrie de la fonction de densité $f(x)$ autour de 0 on a pour tout t, $P(\mathcal{N}_{0;1} > t) = P(\mathcal{N}_{0;1} < -t)$ donc $1 - P(\mathcal{N}_{0;1} \leqslant t) = P(\mathcal{N}_{0;1} < -t)$ et par suite $\Phi(t) + \Phi(-t) = 1 \quad \forall t \in \mathbb{R}$.

Cette relation permet de déterminer $\Phi(t)$ pour des valeurs de t négatives à partir de la table statistique de la page 374. Ainsi $\Phi(-1{,}96) = 1 - \Phi(1{,}96) \cong 1 - 0{,}975 = 0{,}025$.

Remarque. $P(-a \leqslant \mathcal{N}_{0;1} \leqslant a) = 1 - \alpha \Longrightarrow \Phi(a) = 1 - \alpha/2$ expression déterminant a lorsqu'on se fixe α.

2 Loi normale $\mathcal{N}(m\ ;\ \sigma^2)$ de moyenne m et d'écart-type σ

Une *v.a.* X qui suit la loi normale $\mathcal{N}(m,\sigma^2)$, « en abrégé $X \sim> \mathcal{N}(m,\sigma^2)$ », est souvent notée $\mathcal{N}_{m;v}$ où $v = \sigma^2$.

2.1 Caractérisation

m et σ désignent des nombres réels donnés avec σ positif. La loi normale de valeur moyenne m et de variance σ^2 est caractérisée

– par son support $\Delta =]-\infty,\infty[$

– sa densité de probabilité $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} \ \forall x \in \mathbb{R}$

La valeur moyenne de la distribution est égale à m et l'écart-type est égal à σ :

$$E(X) = m = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = m,\ V(X) = \int_{-\infty}^{\infty} (x-m)^2 f(x)\mathrm{d}x = \sigma^2$$

le skewness $\mu_3(X) = E[(X-m)^3]/\sigma^3(X) = 0$,

le kurtosis $\mu_4(X) = E[(X-m)^4]\ /\sigma^4(X) = 3$.

2.2 Propriétés

Soit une *v.a.* X qui suit la loi normale $\mathcal{N}(m,\sigma^2)$. On a les propriétés suivantes :

i) la *v.a.* $Y = (X-m)/\sigma$ suit la loi normale centrée réduite :

$$(X-m)/\sigma = \mathcal{N}_{0;1}\ ;$$

ii) désignant par $F(t)$ la fonction de répartition de la *v.a.* X qui suit la loi normale $\mathcal{N}(m,\sigma^2)$ et par Φ la fonction de répartition de la loi normale centrée réduite

$$F(t) = P(X \leqslant t) = \Phi[(t-m)/\sigma]$$

car $P(X \leqslant t) = P((X-m)/\sigma \leqslant (t-m)/\sigma) = P(\mathcal{N}_{0;1} \leqslant (t-m)/\sigma) = \Phi[(t-m)/\sigma]$

iii) la *v.a.* $Z = cX + d$ (où c et d sont des réels donnés) suit la loi normale $\mathcal{N}(cm + d;\ (c\sigma)^2)$.

© Dunod. La photocopie non autorisée est un délit.

Exemple

Une presse façonne des plaques de chocolat dont le poids en g suit sensiblement une loi normale de valeur moyenne $m = 100$ et d'écart-type $\sigma = 4$ grammes. On cherche à déterminer la probabilité pour qu'une plaque ait un poids inférieur à 96 g :

$$P(X < 96) = \Phi[(96 - 100)/4] = \Phi(-1) = 1 - \Phi(1) = 1 - 0{,}8413 = 15{,}87\,\%.$$

2.3 Somme de variables normales indépendantes

Soit n *v.a.* normales indépendantes $X_1, X_2, \ldots, X_n$: $X_1 \sim> \mathcal{N}(m_1, \sigma_1^2), \ldots, X_n \sim> \mathcal{N}(m_n, \sigma_n^2)$ alors la somme $Z = X_1 + X_2 + \ldots + X_n$ suit la loi $\mathcal{N}(m_0; \sigma_0^2)$ où $m_0 = \sum_{i=1}^{n} m_i$ et $\sigma_0^2 = \sum_{i=1}^{n} \sigma_i^2$.

Théorème. Soit n *v.a.* indépendantes $X_1, X_2, \ldots, X_n$ qui suivent une même loi normale $\mathcal{N}(m; \sigma^2)$. Alors $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$ suit la loi $\mathcal{N}(m; \sigma^2/n)$ et de façon équivalente : $\sqrt{n}(\overline{X} - m)/\sigma = \mathcal{N}_{0;1}$

Exemple (suite)

Reprenant le dernier exemple et supposant que les plaques sont vendues par lot de 5 on se demande quelle est la probabilité pour que le poids moyen des plaques sur le lot soit inférieur à 96 g. Les poids $X_1, X_2, \ldots, X_5$ suivent la loi normale $\mathcal{N}(100; 4^2)$ donc $\overline{X} = (X_1 + X_2, \ldots + X_5)/5$ suit la loi $\mathcal{N}(100; 16/5)$. Aussi $P(\overline{X} < 96) = \Phi[(96 - 100)/\sqrt{16/5}] = \Phi(-2{,}45) = 1 - \Phi(2{,}45) = 1 - 0{,}993 = 0{,}7\ \%$

Repères : Convergence en loi, théorème central limite (forme réduite)

Soit une suite de variables aléatoires indépendantes $X_1, X_2, \cdots, X_n, \cdots$ qui suivent une même loi discrète ou continue $\mathcal{L}(m;\sigma)$ de valeur moyenne m et d'écart-type σ.

La suite des moyennes aléatoires $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i = (X_1 + X_2 + \cdots + X_n)/n$ a la propriété suivante : « $\sqrt{n}(\overline{X}_n - m)/\sigma$ converge en loi vers $\mathcal{N}(0; 1)$ lorsque $n \to \infty$ »,

autrement dit, $P(\sqrt{n}(\overline{X}_n - m)/\sigma \leqslant t) \xrightarrow[n\to\infty]{} P(\mathcal{N}_{0;1} \leqslant t) = \Phi(t)\ \forall\, t \in \mathbb{R}$.

En pratique, lorsque n est grand, on peut approximer la fonction de répartition $F_n(t)$ de $\sqrt{n}(\overline{X}_n - m)/\sigma$ par $\Phi(t)$ qui est la fonction de répartition de $\mathcal{N}(0; 1)$: $F_n(t) \cong \Phi(t)\ \forall\, t \in \mathbb{R}$.

L'approximation « $P(\sqrt{n}(\overline{X}_n - m)/\sigma \leqslant t) \cong P(\mathcal{N}_{0;1} \leqslant t)\ \forall\, t \in \mathbb{R}$ » s'écrit $\sqrt{n}(\overline{X}_n - m)/\sigma \cong \mathcal{N}_{0;1}$[1]. La notation $\cong$ signifie que, $\forall\, t \in \mathbb{R}$, l'écart entre les valeurs prises par les fonctions de répartition des deux variables est négligeable.

1. En fait, $\nu_3 = E[|X - m|^3]$ désignant le moment centré absolu d'ordre 3 de la loi $\mathcal{L}$, on a $|F_n(t) - \Phi(t)| \leqslant 0{,}7975 \times \nu_3/(\sqrt{n} \times \sigma^3)$ (van Beeck).

Généralisation. Théorème de Liapounoff.

Soit une suite de *v.a.* indépendantes $X_1, X_2, \cdots, X_n, \cdots$ où, pour chaque i, la valeur moyenne $m_i = E(X_i)$, la variance $V(X_i)$ et le moment centré absolu d'ordre trois $\nu_3(X_i) = E[|X_i - m_i|^3]$ sont définis et pour laquelle le quotient

$q_n = \sqrt[3]{\sum_{i=1}^{n} \nu_3(X_i)} \Big/ \sqrt[2]{\sum_{i=1}^{n} V(X_i)}$ tend vers zéro lorsque $n \to \infty$ (condition notamment satisfaite quand les *v.a.* suivent la même loi).

La suite des sommes partielles $T_n = X_1 + X_2 + \cdots + X_n$ satisfait à la propriété de convergence vers la loi normale : $\dfrac{T_n - E(T_n)}{\sigma(T_n)} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}_{0;1}$,

Théorème. X_n et Y_n désignent deux suites de *v.a.* réelles indépendantes telles que $X_n \xrightarrow{\text{loi}} \mathcal{N}_{m_1;\nu_1}$ et $Y_n \xrightarrow{\text{loi}} \mathcal{N}_{m_2;\nu_2}$.; a et b désignent deux réels donnés. Alors $Z_n = aX_n + bY_n \xrightarrow{\text{loi}} \mathcal{N}_{m;\nu}$ où $m = am_1 + bm_2$ et $\nu = a^2\nu_1 + b^2\nu_2$.

Section **2**

LOIS DISCRÈTES

1 Loi de Bernoulli $\mathcal{B}(p)$

1.1 Caractérisation

La loi est caractérisée par

– son support constitué des entiers 0 et 1 : $\Delta = \{0,1\}$

– sa distribution de poids : $p(1) = p$ et $p(0) = 1 - p$ où p est un nombre donné vérifiant $0 < p < 1$.

De façon équivalente, on dit que la *v.a.* X suit la loi de Bernoulli $\mathcal{B}(p)$ de paramètre p « en abrégé $X \sim> \mathcal{B}(p)$ » lorsque :

$$\Delta_X = \{0,1\},\ P(X = 1) = p \text{ et } P(X = 0) = (1 - p) \text{ (noté } q).$$

On a $E(X) = p$; $V(X) = p(1 - p)$.

Justification. $E(X) = 0 \times q + 1 \times p = p$; $E(X^2) = 0^2 \times q + 1^2 \times p = p$ donc $V(X) = E(X^2) - [E(X)]^2 = pq$.

© Dunod. La photocopie non autorisée est un délit.

1.2 Utilisation des variables de Bernoulli

À une expérience élémentaire dont l'issue est un succès $\boldsymbol{S}$ ou un échec $\boldsymbol{E}$ avec les probabilités respectives p et $q = 1 - p$, on associe la *v.a.* X qui prend la valeur 1 en cas de succès ou 0 en cas d'échec. X a pour domaine des valeurs $\Delta_X = \{0,1\}$ et $P(X = 1) = p(\boldsymbol{S}) = p$, $P(X = 0) = p(\boldsymbol{E}) = q$.

2 Loi binomiale $\mathcal{B}(m_0, p)$

Une *v.a.* X qui suit la loi binomiale de paramètres (m_0, p) « en abrégé $X \sim> \mathcal{B}(m_0, p)$ » est souvent notée $\mathcal{B}_{m_0, p}$.

2.1 Caractérisation

La loi binomiale de paramètres (m_0, p) où m_0 est un entier naturel et p un nombre réel vérifiant $0 < p < 1$, a

– pour support $\Delta = \{0,1,2,\ldots,m_0\}$

– la distribution de poids : $p(h) = C_{m_0}^h p^h (1 - p)^{m_0 - h} \quad \forall\, h \in \Delta$

avec $C_{m_0}^h = \dfrac{m_0!}{h!(m_0 - h)!}$.

Cette distribution a pour valeur moyenne $m = m_0 p$ et pour variance $\sigma^2 = m_0 p(1 - p)$.

De façon équivalente, on dit que le résultat X d'une expérience envisagée suit la loi binomiale de paramètres (m_0, p) lorsque X peut prendre l'une des valeurs entières appartenant à $\Delta_X = \{0,1,2,\ldots,m_0\}$ et que $P(X = h) = C_{m_0}^h p^h (1 - p)^{m_0 - h}\ \forall\, h \in \Delta_X$.

On a $E(X) = m_0 p$; $V(X) = m_0 pq$ où $q = (1 - p)$. [*Cf.* § 2.3].

2.2 Fonction de répartition

Propriétés

i) Sur la table numérique de la page 375 on peut lire, pour divers choix de m_0 et de p, les valeurs de $P(\mathcal{B}_{m_0, p} \leqslant h)$ ou utiliser $P(\mathcal{B}_{m_0, q} \leqslant h) = 1 - P(\mathcal{B}_{m_0, p} \leqslant m_0 - h - 1)$.

ii) Pour m_0 grand [ordre de grandeur $m_0 pq > 10$ ou $(m_0 > 50, m_0 p > 5$ et $m_0 q > 5)$], on a l'approximation suivante au sens des distributions :

$$\frac{\mathcal{B}_{m_0, p} - m_0 p}{\sqrt{m_0 pq}} \cong \mathcal{N}_{0;1} \text{ (théorème de Moivre).}$$

Cette approximation doit être utilisée avec *correction de continuité* car on substitue à la fonction de répartition d'une loi discrète celle d'une loi continue :

$\forall h_1$ et h_2 vérifiant $0 \leqslant h_1 \leqslant h_2 \leqslant m_0$ on a

$$P(h_1 \leqslant \mathcal{B}_{m_0,p} \leqslant h_2) \cong \Phi\left(\frac{h_2 - m_0 p + 0{,}5}{\sqrt{m_0 pq}}\right) - \Phi\left(\frac{h_1 - m_0 p - 0{,}5}{\sqrt{m_0 pq}}\right)$$

où $\Phi(t) = P(\mathcal{N}_{0;1} \leqslant t)$

Justification technique. $(h_1 \leqslant \mathcal{B}_{m_0,p} \leqslant h_2) \Leftrightarrow (h_1 - 0{,}5 \leqslant \mathcal{B}_{m_0,p} \leqslant h_2 + 0{,}5)$ puisque $\mathcal{B}_{m_0,p}$ ne peut prendre que des valeurs entières. Donc

$$P(h_1 \leqslant \mathcal{B}_{m_0,p} \leqslant h_2) = P(h_1 - 0{,}5 \leqslant \mathcal{B}_{m_0,p} \leqslant h_2 + 0{,}5)$$

$$= P\left(\frac{h_1 - m_0 p - 0{,}5}{\sqrt{m_0 pq}} \leqslant \frac{\mathcal{B}_{m_0 p} - mp}{\sqrt{m_0 pq}} \leqslant \frac{h_2 - m_0 p + 0{,}5}{\sqrt{m_0 pq}}\right)$$

$$\cong P\left(\frac{h_1 - m_0 p - 0{,}5}{\sqrt{m_0 pq}} \leqslant \mathcal{N}_{0;1} \leqslant \frac{h_2 - m_0 p + 0{,}5}{\sqrt{m_0 pq}}\right).$$

2.3 Utilisation des variables binomiales

Soit une suite de m_0 épreuves réalisées dans des conditions identiques, la réalisation de chaque épreuve étant soit un succès ***S*** soit un échec ***E*** avec les probabilités respectives p et $q = 1 - p$. Alors, le nombre aléatoire X de succès obtenus à l'issue des m_0 épreuves suit la loi binomiale $\mathcal{B}(m_0, p)$.

Les m_0 épreuves étant notées $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m_0}$, associons à la i-ème épreuve ε_i la variable de Bernoulli X_i qui vaut 1 si on obtient un succès ou 0 sinon. On a évidemment $X = X_1 + \ldots + X_{m_0}$, autrement dit X est la somme de m_0 variables indépendantes $X_1, X_2, \ldots, X_{m_0}$ qui suivent la loi de Bernoulli $\mathcal{B}(p)$ et donc $E(X_i) = p$; $V(X_i) = p(1-p)$ quelque soit $i = 1,2,\ldots,m_0$. On en déduit

$$E(X) = E(X_1 + \ldots + X_{m_0}) = E(X_1) + \ldots + E(X_{m_0}) = m_o p$$

puis

$$V(X) = V(X_1 + \ldots + X_{m_0}) = V(X_1) + \ldots + V(X_{m_0}) = m_0 pq$$

car les *v.a.* $X_1, X_2, \ldots, X_{m_0}$ sont indépendantes.

D'une façon générale $\mathcal{B}_p^{(1)}, \ldots, \mathcal{B}_p^{(m_0)}$ désignant m_0 variables de Bernoulli indépendantes de même paramètre p, on a : $\mathcal{B}_p^{(1)} + \mathcal{B}_p^{(2)} + \ldots \mathcal{B}_p^{(mo)} = \mathcal{B}_{m_0;p}$.

Exemple

Une machine M produit en série des pièces d'un même type dont 5 % sont défectueuses. On constitue des lots de 10 pièces et l'on s'intéresse au nombre X de pièces défectueuses

© Dunod. La photocopie non autorisée est un délit.

dans un lot choisi au hasard. Le nombre aléatoire X de pièces défectueuses suit la loi binomiale $\mathcal{B}(10; 0{,}05)$. En effet il suffit de considérer qu'à l'issue de chaque extraction l'événement A « obtenir une pièce défectueuse » est un succès $\boldsymbol{S}$ et de constater que $p(\boldsymbol{S}) = p(A) = \ 0{,}05 \cdot p(\overline{\boldsymbol{S}}) = p(\overline{A}) = 0{,}95$.

La probabilité pour qu'il y ait au plus un élément défectueux dans le lot est donnée par le nombre $P(X \leqslant 1) = P(X = 0) + P(X = 1)$
$= C_{10}^0(0{,}05)^0(0.95)^{10} + C_{10}^1(0{,}05)^1(0{,}95)^9 = 0{,}914.$

2.4 Somme de variables binomiales indépendantes

La somme de variables binomiales indépendantes $\mathcal{B}_{m_1;p}$ et $\mathcal{B}_{m_2;p}$ de même paramètre p, suit la loi binomiale de paramètres $(m_1 + m_2; p)$: $\mathcal{B}_{m_1;p} + \mathcal{B}_{m_2;p} = \mathcal{B}_{m_1+m_2;p}$. ***Justification.*** $\mathcal{B}_{m_1;p}$ [resp. $\mathcal{B}_{m_2;p}$] pouvant être décomposé en une somme de m_1 [resp. m_2] variables de Bernoulli indépendantes de paramètre p, $\mathcal{B}_{m_1;p} + \mathcal{B}_{m_2;p}$ est la somme des $m_1 + m_2$ variables de Bernoulli de paramètre p.

3 Loi géométrique $\mathcal{G}(p)$

Une *v.a.* X qui suit la loi géométrique de paramètre p (en abrégé $X \sim> \mathcal{G}(p)$) est souvent notée $\mathcal{G}_p$.

3.1 Caractérisation

La loi géométrique $\mathcal{G}(p)$, où p désigne un nombre vérifiant $0 < p < 1$, a :

– pour support $\Delta = \{0,1,2,\ldots,n,\ldots\} = \mathbb{N}$

– pour distribution de poids : $p(n) = pq^n \ \forall\, n \in \Delta$ avec $q = 1 - p$.

On dit que X suit la loi géométrique de paramètre p lorsque $\Delta_X = \mathbb{N}$ et $P(X = n) = pq^n$. On a $E(X) = q/p$; $V(X) = q/p^2$.

3.2 Utilisation des variables géométriques

Soit une expérience élémentaire dont l'issue est un succès $\boldsymbol{S}$ ou un échec $\boldsymbol{E}$ avec des probabilités respectives p et $q = 1 - p$. On renouvelle autant de fois qu'il est nécessaire ce type d'expérience dans des conditions identiques jusqu'à l'obtention d'un premier succès. Le nombre aléatoire X d'échecs précèdant le 1er succès suit la loi géométrique $\mathcal{G}(p)$.

En effet, $\Delta_X = \{0,1,2,\ldots,n,\ldots\} = \mathbb{N}$ et

– $P\{X = 0\} =$ Proba(avoir un succès à la 1re expérience) $= p$;

– $P\{X = 1\} =$ Proba(avoir un échec à la 1re et un succès à la 2e)
$= p(\boldsymbol{E}_1 \text{ et } \boldsymbol{S}_2) = p(\boldsymbol{E}_1) \times p(\boldsymbol{S}_2) = q \times p$

(car les événements qui résultent des expériences successives sont indépendants les uns des autres) ;

– $P\{X = 2\} = P$(avoir un échec à la 1re et un échec à la 2e et 1 succès à la 3e) = $p(\boldsymbol{E}_1$ et $\boldsymbol{E}_2$ et $\boldsymbol{S}_3) = p(\boldsymbol{E}_1) \times p(\boldsymbol{E}_2) \times p(\boldsymbol{S}_3) = q \times q \times p = q^2 p$; etc.

Remarque. Certains auteurs considèrent que c'est le nombre aléatoire Y d'expériences nécessaires pour obtenir un succès qui suit une loi géométrique. On a $Y = X + 1$ et par suite $\Delta_Y = \{1,2,\ldots,n,\ldots\} = \mathbb{N}^*$; $P(Y=n)=P(X+1=n)=P(X = n-1) = pq^{n-1}\ \forall n \in \Delta_Y$; $E(Y) = 1/p$; $V(Y) = q/p^2$.

Exemple

En faisant du marketing téléphonique on sait que la probabilité de réussir à obtenir un rendez-vous en face à face est égale à 0,25. Le nombre aléatoire X d'appels précédant le premier appel concrétisé par une prise de rendez-vous, suit la loi géométrique de paramètre $p = 0{,}25$. La probabilité pour qu'un vendeur n'obtienne un rendez-vous qu'au quatrième appel est égale à la probabilité d'avoir un échec aux trois premiers appels et un succès au quatrième appel : $P(X = 3) = 0{,}25 \times 0{,}75^3 = 0{,}105$. La distribution de X est présentée ci-dessous.

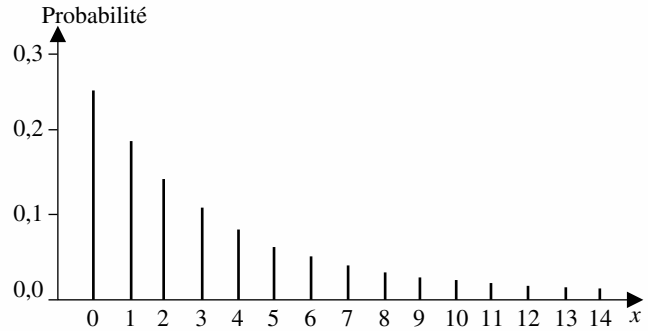


Figure 9.2 – *Distribution de probabilité de la loi géométrique de paramètre 0,25*

4 Loi hypergéométrique notée $\mathcal{H}(\boldsymbol{m_0};\boldsymbol{n_1},\boldsymbol{n_2})$ ou $\mathcal{H}(\boldsymbol{N}\ ;\ \boldsymbol{m_0}, \boldsymbol{p})$

La loi hypergéométrique $\mathcal{H}(m_0; n_1,n_2)$ est parfois notée $\mathcal{H}(N; m_0, p)$ où $N = n_1 + n_2$ et $p = n_1/N$ (autrement dit $n_1 = N \times p$ et $n_2 = N \times (1 - p)$). Une *v.a.* X qui suit la loi hypergéométrique de paramètres m_0 ; n_1, n_2 est notée $\mathcal{H}_{N;m_0,p}$.

4.1 Caractérisation

m_0, n_1 et n_2 désignant trois entiers naturels tels que $m_0 \leqslant n_1 + n_2$, la loi hypergéométrique $\mathcal{H}(m_0; n_1,n_2)$ est caractérisée par

© Dunod. La photocopie non autorisée est un délit.

– son support $\Delta_X = \{h \in \mathbb{N} | \, \mathrm{Sup}(0, m_0 - n_2) \leqslant h \leqslant \mathrm{Inf}(n_1, m_0)\}$.

– sa distribution de poids : $p(h) = \dfrac{C_{n_1}^h \times C_{n_2}^{m_0-h}}{C_{n_1+n_2}^{m_0}} \quad \forall\, h \in \Delta$.

Exemple

La distribution des valeurs d'une v.a. X qui suit la loi $\mathcal{H}(12; 7,13)$ est caractérisée par

$\Delta_X = \{0,1,2,\ldots,7\}$ car $\mathrm{Sup}(0, m_0 - n_2) = \mathrm{Sup}(0, 12 - 13) = 0$ et $\mathrm{Inf}(n_1, m_0) = \mathrm{Inf}(7, 12) = 7$

et par $\mathrm{P}(X = h) = C_7^h \times C_{13}^{12-h} / C_{20}^{12} \; \forall\, h \in \Delta_X$.

Sa représentation graphique est la suivante

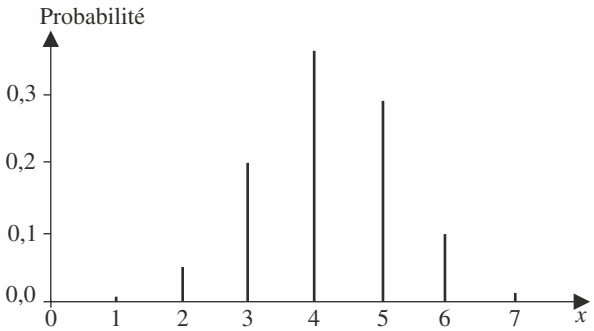


Figure 9.3 – *Distribution de probabilité de la loi* $\mathcal{H}(12; 7,13)$

4.2 Interprétation et propriétés

i) La loi $\mathcal{H}(m_0; n_1, n_2)$ ou $\mathcal{H}(N; m_0, p)$ avec $N = n_1 + n_2$ et $p = n_1/N$ peut être associée à un tirage exhaustif (sans remise) de m_0 boules dans une urne qui contient n_1 boules blanches et n_2 boules noires. En effet, le nombre X de boules blanches tirées obéit à la loi hypergéométrique $\mathcal{H}(N; m_0, p)$

ii) $E(X) = m_0 p$; $V(X) = m_0 pq \dfrac{N - m_0}{N - 1}$ (où $q = 1 - p$).

Le coefficient $\dfrac{N - m_0}{N - 1}$ est appelé « facteur d'exhaustivité » par référence au cas où le tirage ayant lieu avec remise on aurait $V(X) = m_0 pq$.

iii) Lorsque $m_0/N < 0,1$ on peut utiliser l'approximation : $\mathcal{H}_{N;m_0,p} \cong \mathcal{B}_{m_0;p}$, c'est-à-dire $P(\mathcal{H}_{N;m_0,p} \leqslant h) \cong P(\mathcal{B}_{m_0;p} \leqslant h) \; \forall\, h$ vérifiant $0 \leqslant h \leqslant m_0$.

5 Loi de Poisson $\mathcal{P}(\lambda)$

Une *v.a.* qui suit la loi $\mathcal{P}(\lambda)$ est souvent notée $\mathcal{P}_\lambda$.

5.1 Caractérisation

La loi de Poisson de paramètre λ (où λ est un réel positif donné) est caractérisée par

– son support $\Delta = \{0,1,2,\ldots\} = \mathbb{N}$

– sa distribution de poids : $p(n) = \mathrm{e}^{-\lambda}\lambda^n/n! \quad \forall n \in \Delta$.

On dit qu'une *v.a.* X suit la loi de Poisson de paramètre λ (en abrégé, $X \sim> \mathcal{P}(\lambda)$) lorsqu'elle peut prendre n'importe quelle valeur entière naturelle n avec la probabilité $\mathrm{e}^{-\lambda}\lambda^n/n!$: $\Delta_X = \mathbb{N}$, $P(X = n) = \mathrm{e}^{-\lambda}\lambda^n/n!$.

On a $E(X) = \lambda$, $V(X) = \lambda$.

Application. Une entreprise de location de véhicules estime que la demande journalière X suit une loi de Poisson de moyenne égale à 4,5 soit $E(X) = \lambda = 4{,}5$. La distribution de probabilité répond à l'expression suivante : $P(X = n) = \mathrm{e}^{-4{,}5}4{,}5^n/n!$ avec $\Delta_X = \mathbb{N}$. Ainsi $P(X = 0) = \mathrm{e}^{-4{,}5}4{,}5^0/0! = 0{,}011$, $P(X = 1) = \mathrm{e}^{-4{,}5}4{,}5^1/1! = 0{,}05\ldots$

5.2 Fonction de répartition

– La distribution de probabilité et la fonction de répartition sont tabulées pour certaines valeurs de λ. Ainsi pour certains choix de λ on peut lire sur la table de la page 376 la valeur de $P(\mathcal{P}_\lambda \leqslant h)$ où h est un entier naturel.

– Pour λ grand [ordre de grandeur $\lambda > 25$] on a l'approximation suivante au sens des distributions : $(\mathcal{P}_\lambda - \lambda)/\sqrt{\lambda} \cong \mathcal{N}_{0;1}$. Cette approximation doit être utilisée avec correction de continuité car on substitue à la fonction de répartition d'une loi discrète celle d'une loi continue :

$\forall\, h_1$ et h_2 entiers vérifiant $0 \leqslant h_1 \leqslant h_2$ on a

$$P(h_1 \leqslant \mathcal{P}_\lambda \leqslant h_2) \cong \Phi\left(\frac{h_2 - \lambda + 0{,}5}{\sqrt{\lambda}}\right) - \Phi\left(\frac{h_1 - \lambda - 0{,}5}{\sqrt{\lambda}}\right)$$

Pour $\lambda > 10$ on peut également utiliser l'approximation normale $2(\sqrt{\mathcal{P}_\lambda} - \sqrt{\lambda}) \cong \mathcal{N}_{0;1}$ après avoir remarqué que $P(\mathcal{P}_\lambda \leqslant h) = P(\mathcal{P}_\lambda \leqslant h + 0{,}5) = P(2(\sqrt{\mathcal{P}_\lambda} - \sqrt{\lambda}) \leqslant 2(\sqrt{h + 0{,}5} - \sqrt{\lambda}) \cong \Phi[2(\sqrt{h + 0{,}5} - \sqrt{\lambda})]$.

– Pour m_0 grand et $m_0 p$ petit (ordre de grandeur : $m_0 > 50$ et $m_0 p < 3$), on a l'approximation suivante au sens des distributions : $\mathcal{B}_{m_0,\,p} \cong \mathcal{P}_{m_0 p}$.

Application. Reprenant l'exemple précédent et voulant savoir quelle est la probabilité pour que la demande soit comprise entre 4 et 8 véhicules il convient de calculer $P(4 \leqslant \mathcal{P}_{4,5} \leqslant 8)$.

La décomposition « $(\mathcal{P}_{4,5} \leqslant 8) = (\mathcal{P}_{4,5} \leqslant 3) \cup (4 \leqslant \mathcal{P}_{4,5} \leqslant 8)$ » implique $P(\mathcal{P}_{4,5} \leqslant 8) = P(\mathcal{P}_{4,5} \leqslant 3) + P(4 \leqslant \mathcal{P}_{4,5} \leqslant 8)$. Après lecture de table p. 376 où l'on constate que $P(\mathcal{P}_{4,5} \leqslant 8) = 0{,}9597$ et $P(\mathcal{P}_{4,5} \leqslant 3) = 0{,}3423$ on obtient $P(4 \leqslant \mathcal{P}_{4,5} \leqslant 8) = P(\mathcal{P}_{4,5} \leqslant 8) - P(\mathcal{P}_{4,5} \leqslant 3) = 0{,}9597 - 0{,}3423 = 0{,}6174$.

© Dunod. La photocopie non autorisée est un délit.

De façon équivalente, remarquons que
$P(4 \leqslant \mathcal{P}_{4,5} \leqslant 8) = P(\mathcal{P}_{4,5} = 4) + P(\mathcal{P}_{4,5} = 5) + \ldots P(\mathcal{P}_{4,5} = 8) = e^{-4,5}(4,5)^4/4!$
$+ e^{-4,5}(4,5)^5/5! + \ldots e^{-4,5}(4,5)^8/8! \ = 0,6174.$

5.3 Somme de variables de Poisson indépendantes

Soient deux variables de Poisson indépendantes X_1 et X_2 : $X_1 \sim> \mathcal{P}(\lambda_1)$, $X_2 \sim> \mathcal{P}(\lambda_2)$. Alors la *v.a.* $Z = X_1 + X_2$ sur la loi de Poisson de paramètres $\lambda_1 + \lambda_2$: $Z \sim> \mathcal{P}(\lambda_1 + \lambda_2)$. Cette propriété peut être généralisée à une somme de n variables de Poisson indépendantes. Ainsi, si X_1, X_2 et X_3 sont trois *v.a.* indépendantes qui suivent la même loi de Poisson $\mathcal{P}(1,5)$, alors $Z = X_1 + X_2 + X_3$ suit la loi de Poisson $\mathcal{P}(4,5)$.

6. Loi des rangs signés de Wilcoxon $W^+(n)$

À chaque n-uple$(B^1_{0.5}, B^2_{0.5}, \ldots, B^n_{0.5})$ de variables de Bernoulli indépendantes de paramètre 0.5 associons la somme pondérée $W_n^+ = \sum_{i=1}^{n} i\, B^i_{0.5}$. La variable W_n^+ suit la loi caractérisée par son domaine des valeurs $\Delta = \{0,1,2,3 \ldots, n(n+1)/2\}$ et la distribution $P(W_n^+ = h) =$ [nombre de n-uples $(b_1, b_2 \ldots, b_n)$ tels que $b_i = 1$ ou 0 et $\sum_{i=1}^{n} i b_i = h]/2^n$. *Cf.* table page 384.

Cette distribution a pour valeur moyenne $m = E(W_n^+) = n(n+1)/4$ et pour variance $V(W_n^+) = n(n+1)(2n+1)/24$.

Elle est symétrique : $P(W_n^+ \leqslant h) = P(W_n^+ \geqslant n(n+1)/2 - h)\ \forall h \in \Delta$.

Pour $n > 15$ on utilise l'approximation normale avec correction de continuité :
$P(W_n^+ \leqslant h) \cong \Phi[(h + 0.5 - n(n+1)/4)/\sqrt{n(n+1)(2n+1)/24}]$

Propriété fondamentale. Soit n *v.a.* indépendantes $Y_1, Y_2 \ldots, Y_n$ qui suivent une même loi continue symétrique autour de 0 et dont la fonction de répartition F vérifie pour tout t : $F(t) + F(-t) = 1\ \forall t$. Les n valeurs numériques $y_1, y_2, \ldots, y_n$ prises par les n *v.a.* Y_i étant rangées par ordre de croissance des valeurs absolues : $|y_1| < |y_2| < \ldots < |y_n|$, la somme t^+ des rangs des valeurs y_i positives est la valeur prise par une variable T_n^+ qui suit la loi $W^+(n)$.

Exemple

On dispose d'un échantillon de cinq valeurs : $-0,8$; $2,1$; $-3,5$; $2,6$; $4,7$. On a $|-0,8| < 2,1 < 2,6 < -|3,5| < 4,7$ donc $t^+ = 2 + 3 + 5 = 10$

Application. Lorsque la valeur t^+ prise par T_n^+ est trop proche d'une des extrémités 0 ou $n(n+1)/2$ du domaine Δ des valeurs possibles, on peut supputer que T_n^+ ne suit pas la loi des rangs signés de Wilcoxon et donc que la distribution F n'est pas symétrique autour de 0.

Section 3

LOIS CONTINUES (SUITE)

1 Loi de Student-Fisher $\mathcal{S}t(n)$

Une *v.a.* qui suit la loi $\mathcal{S}t(n)$ est souvent notée $\boldsymbol{t}_n$.

1.1 Caractérisation

La loi de Student-Fisher à n degrés de libertés (*en abrégé* $\mathcal{S}t(n)$) où n est un entier positif est caractérisée :

– par son support $\Delta =]-\infty, \infty[$

– sa densité de probabilité $f(x) = \dfrac{k(n)}{(1+\dfrac{x^2}{n})^{\frac{n+1}{2}}} \; \forall x \in \mathbb{R}$

où $$k(n) = \frac{(n/2-0{,}5)!}{\sqrt{n\pi} \times (n/2-1)!}$$

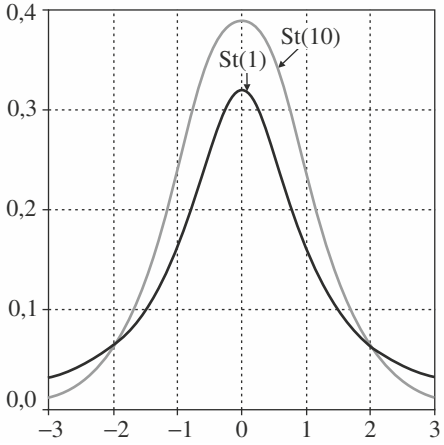


Figure 9.4 – *Fonctions de densité de la loi de Student-Fisher (pour n = 1 et n = 10)*

© Dunod. La photocopie non autorisée est un délit.

Lorsque α n'est pas un entier naturel, $\alpha! = \int_0^\infty x^\alpha e^{-x} dx$].

$$\alpha! = \alpha(\alpha-1)!;\ (1/2)! = \sqrt{\pi/2};\ (3/2)! = 3/2 \times (1/2)! = 3/2 \times \sqrt{\pi/2}$$

La moyenne de la loi est égale à 0 soit $E(\boldsymbol{t}_n) = 0$. La variance n'est pas définie pour $n = 1$ ou 2 et pour $n > 2$: $V(\boldsymbol{t}_n) = n/(n-2)$.

Une propriété fondamentale. Soit deux *v.a.* indépendantes : une variable $\mathcal{N}_{0;1}$ et une variable Khi-deux χ^2_n (*cf.* § 2 ci-après). Alors le quotient $\mathcal{N}_{0;1}/\sqrt{\chi^2_n/n} = \boldsymbol{t}_n$.

1.2 Propriétés de la fonction de répartition

– La fonction de répartition $F_n(t) = P(\boldsymbol{t}_n \leqslant t)$ est tabulée (*cf.* page 378), le lecteur trouvera la valeur de $F_n(t)$ pour diverses valeurs de t positif. De plus,

$$F_n(t) + F_n(-t) = 1\ \forall t \in \mathbb{R}$$

car la densité de probabilité est symétrique autour de 0.

La lecture de la table permet de déterminer directement la valeur de t telle que $P(\boldsymbol{t}_n \leqslant t) = \theta$ (θ valeur donnée $\geqslant 0{,}5$). Par exemple la valeur de t pour laquelle $P(\boldsymbol{t}_{10} \leqslant t) = 0{,}95$ se lit directement : $t = 1{,}812$. *A contrario* pour trouver la valeur de t pour laquelle $P(\boldsymbol{t}_n \leqslant t) = \theta < 0{,}5$, remarquons que $P(\boldsymbol{t}_n \leqslant -t) = 1 - \theta$. Par exemple pour déterminer la valeur de t pour laquelle $P(\boldsymbol{t}_{10} \leqslant t) = 0{,}05$ remarquons que $P(\boldsymbol{t}_{10} \leqslant -t) = 0{,}95$. Sur la table on lit $-t = 1{,}812$ soit $t = -1{,}812$.

– Pour $n > 120$ on utilise l'approximation $\boldsymbol{t}_n \cong \mathcal{N}_{0,1}$, autrement dit pour tout t réel :

$$P(\boldsymbol{t}_n \leqslant t) \cong P(\mathcal{N}_{0,1} \leqslant t) = \Phi(t).$$

1.3 Échantillon issu d'une distribution $\mathcal{N}(m, \sigma^2)$ et loi de Student-Fisher

Les principales utilisations des lois de Student-Fisher s'appuient sur le théorème suivant :

Théorème de la moyenne. Soit un échantillon de n *v.a* indépendantes X_1, $X_2, \ldots, X_n$ qui suivent une même loi normale $\mathcal{N}(m, \sigma^2)$. Alors,

i) la moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et la déviation standard $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ de cet échantillon sont deux *v.a.* indépendantes,

ii) la *v.a.* $\frac{\overline{X} - m}{S/\sqrt{n}}$ suit la loi $\mathcal{S}t(n-1)$; autrement dit $\frac{\overline{X} - m}{S/\sqrt{n}} = \boldsymbol{t}_{n-1}$.

2 Loi du $\chi^2(n)$ (lire khi-deux) à n degrés de liberté

Une *v.a.* qui suit la loi $\chi^2(n)$ est appelée variable khi-deux, on la note χ^2_n.

2.1 Caractérisation

La loi khi-deux à n degrés de libertés (notée $\chi^2(n)$) où n est un entier naturel non nul est caractérisée

– par son support $\Delta = [0,\infty[$

– sa densité de probabilité $f(x) = x^{n/2-1} \times \mathrm{e}^{-x/2}/[2^{n/2}(n/2-1)!] \quad \forall x > 0$ où $(n/2-1)! = (2n)!\sqrt{\pi}/2^{2n}n!$ (*cf.* p. 136).

La moyenne de la loi est égale à n le nombre de degrés de liberté et la variance est égale à $2 \times n$ soit $E(\chi^2_n) = n$, $V(\chi^2_n) = 2n$.

Propriété fondamentale. Soit n *v.a.* indépendantes $\mathcal{N}^{(1)}_{0;1}, \mathcal{N}^{(2)}_{0;1}, \ldots, \mathcal{N}^{(n)}_{0;1}$ qui suivent la loi normale standard. Alors la somme
$[\mathcal{N}^{(1)}_{0;1}]^2 + [\mathcal{N}^{(2)}_{0;1}]^2 + \ldots + [\mathcal{N}^{(n)}_{0;1}]^2 = \chi^2_n$. En particulier, $[\mathcal{N}_{0,1}]^2 = \chi^2_1$.

2.2 Propriétés de la fonction de répartition

– La table de la page 377 donne pour diverses valeurs de n et pour divers choix de α, la valeur de t correspondant à $\alpha = P(\chi^2_n \geqslant t)$.
Pour $n = 2$ on a $P(\chi^2_2 \leqslant t) = 1 - \mathrm{e}^{-t/2} \; \forall t > 0$.

– Pour $n > 30$ on peut utiliser :
soit l'approximation de Fisher : $\sqrt{2\chi^2_n} - \sqrt{2n-1} \cong \mathcal{N}_{0;1}$ c'est-à-dire

$$P(\chi^2_n \leqslant t) \cong \Phi(\sqrt{2t} - \sqrt{2n-1}) \; \forall t > 0 \; ;$$

soit l'approximation meilleure de Wilson-Hilferty :

$$P(\chi^2_n \leqslant t) \cong \Phi\left(\frac{\sqrt{t/n} - (1 - 2/9n)}{\sqrt{2/9n}}\right) \; \forall t > 0 .$$

Exemple

$X \sim\!> \chi^2(30)$. Pour déterminer la valeur a telle que $P(X \leqslant a) = 0{,}80$ c'est-à-dire $P(X \geqslant a) = 0{,}20$, on lit sur la table de la page 377, $P(\chi^2_{30} \geqslant 36{,}3) = 0{,}20$. Donc $a = 36{,}3$.

En utilisant l'approximation normale on a $P(\chi^2_{30} \leqslant a) \cong \Phi(\sqrt{2a} - \sqrt{2 \times 30 - 1}) = 0{,}80$ ce qui implique (*cf.* p. 374) $\sqrt{2a} - \sqrt{2 \times 30 - 1} = 0{,}84$ et par suite $a = 36{,}3$.

© Dunod. La photocopie non autorisée est un délit.

2.3 Somme de variables khi-deux indépendantes

Soient deux variables khi-deux indépendantes X_1 et X_2 : $X_1 \sim> \chi^2(n_1)$, $X_2 \sim> \chi^2(n_2)$.

Alors $Z = X_1 + X_2$ suit la loi khi-deux à $n = n_1 + n_2$ degrés de liberté :

$$Z \sim> \chi^2(n_1 + n_2).$$

Cette propriété peut être généralisée à une somme de ν variables khi-deux indépendantes. La justification résulte immédiatement de la propriété fondamentale.

2.4 Échantillon issu d'une distribution $\mathcal{N}(m, \sigma^2)$ et loi khi-deux

Théorème de la variance. Soit un échantillon de n *v.a* indépendantes $X_1, X_2, \ldots, X_n$ qui suivent une même loi normale $\mathcal{N}(m,\sigma^2)$, la moyenne et la variance standard de cet échantillon étant respectivement notées $\overline{X}$ et S^2 :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

On a les propriétés suivantes :

i) la variable $Z = \sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2$ suit la loi $\chi^2(n)$,

ii) la variable $Z^* = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$ suit la loi $\chi^2(n-1)$, autrement dit $\frac{(n-1)S^2}{\sigma^2} = \chi^2_{n-1}$

Justification de i). $[(X_i - m)/\sigma] = \mathcal{N}^{(i)}_{0;1}$ donc Z est une somme des carrés de n variables normales standard indépendantes.

3 Loi gamma $\gamma(\lambda, a)$

Une *v.a.* qui suit la loi gamma de paramètres (λ,a) sera notée $\Gamma_{\lambda,a}$ et de façon plus simple Γ_a lorsque $\lambda = 1$.

3.1 Caractérisation

La loi gamma de paramètres (λ,a), où λ et a désignent deux réels positifs donnés est caractérisée par :

– son support $\Delta = [0,\infty[$

– sa densité de probabilité $f(x) = kx^{a-1}e^{-x/\lambda}\ \forall\, x > 0$ où $k^{-1} = \lambda^a \times (a-1)!$.

Cette distribution a pour valeur moyenne $m = a\lambda$ et pour variance $\sigma^2 = a\lambda^2$; soit $E[\Gamma_{\lambda,a}] = a\lambda$, $V[\Gamma_{\lambda,a}] = a\lambda^2$. De façon équivalente : $\lambda = \sigma^2/m$, $a = (m/\sigma)^2$.

Les valeurs de la fonction de répartition $F(t) = P(\Gamma_{\lambda,a} \leqslant t)$ peuvent être obtenues en appliquant les procédures présentées dans la section 5 avec le logiciel Excel ou SPSS.

3.2 Propriétés

i) On a $\Gamma_{\lambda,a} = \lambda\Gamma_a$ et $\Gamma_a = (1/2)\chi^2_{2a}$; par suite $\Gamma_{\lambda,a} = \lambda \times (1/2) \times \chi^2_{2a}$ et donc $P(\Gamma_{\lambda,a} \leqslant t) = P(\chi^2_{2a} \leqslant 2t/\lambda)$.

ii) La somme de deux variables indépendantes $\Gamma_{\lambda,a}$ et $\Gamma_{\lambda,a'}$, de même paramètre λ, suit la loi $\gamma(\lambda, a + a')$: $\Gamma_{\lambda,a} + \Gamma_{\lambda,a'} = \Gamma_{\lambda,a+a'}$.

Exemple

La durée de vie (exprimée en dizaine de milliers de km) d'un pneu de type donné utilisé sur la roue avant droite peut être modélisée par une *v.a.* X qui suit la loi gamma de moyenne $m = 3$ et d'écart-type $\sigma = 1$. Pour évaluer la probabilité de ne pas avoir de problèmes techniques dus à l'usure des pneus avant 25 000 km, il faut calculer $P(X > 2{,}5)$. Compte tenu les valeurs de m et σ on constate que $a = 9$ et $\lambda = 1/3$: $X = \Gamma_{1/3,9} = (1/3)\Gamma_9 = (1/6)\chi^2_{18}$. Par suite

$$P(X \geqslant 2{,}5) = P(\chi^2_{18}/6 \geqslant 2{,}5) = P(\chi^2_{18} \geqslant 15) \cong 0{,}661\,.$$

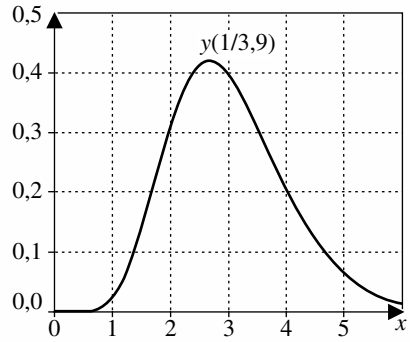


Figure 9.5 – *Fonction de densité de la loi gamma de paramètres (1/3, 9)*

© Dunod. La photocopie non autorisée est un délit.

4 Loi exponentielle Exp(a)

Une *v.a.* X qui suit la loi Exp(a) sera éventuellement notée $\mathcal{E}_a$.

4.1 Caractérisation

La loi exponentielle de paramètre a, où a désigne un nombre réel positif donné est caractérisée par

– son support $\Delta = [0, \infty[$

– sa densité de probabilité $f(x) = a\mathrm{e}^{-ax}\ \forall\, x > 0$.

Cette distribution a pour valeur moyenne $m = 1/a$ et pour écart-type $\sigma = 1/a$ autrement dit $E(\mathcal{E}_a) = 1/a$, $V(\mathcal{E}_a) = 1/a^2$.

Sa fonction de répartition a pour expression $F(t) = 1 - \mathrm{e}^{-at}\ \forall\, t > 0$.

Exemple

Le directeur d'un hypermarché a mis en place un service chargé d'analyser les files d'attente aux caisses. Sur l'une des caisses, ce service a constaté que la durée moyenne séparant une entrée d'une sortie était de 5 minutes. La durée X séparant une entrée d'une sortie étant supposée suivre une loi exponentielle, on a $E(X) = m = 1/a = 5$, donc $a = 1/5$ et par suite $F(t) = 1 - \mathrm{e}^{-t/5}$.

La connaissance de la valeur de m permet par exemple de déterminer la probabilité pour que la durée d'attente soit comprise entre 1 et 6 minutes :

$$P(1 \leqslant X \leqslant 6) = F(6) - F(1) = \mathrm{e}^{-1/5} - \mathrm{e}^{-6/5} = 0{,}5175\,.$$

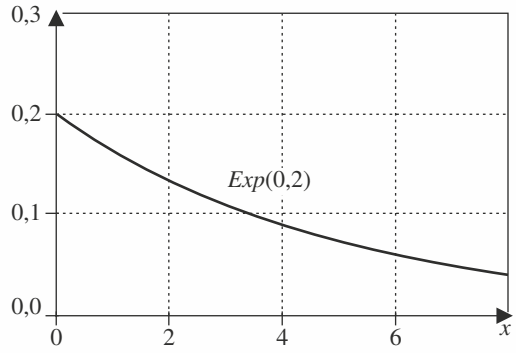


Figure 9.6 – *Fonction de densité de la loi exponentielle de paramètre* $a = 0{,}2$

Une propriété fondamentale. Soit une suite X_n de *v.a.* indépendantes qui suivent la loi Exp(a). Alors la *v.a.* $Y_t =$ (Nombre de $X_i \leqslant t$) suit la loi de Poisson $\mathcal{P}(at)$.

4.2 Théorème

Soit un échantillon de n *v.a* indépendantes $X_1, X_2, \ldots, X_n$ qui suivent la loi Exp(a) et soit $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne de cet échantillon. La *v.a.* $2a\overline{X} \times n$ suit la loi $\chi^2(2n)$: $2a\overline{X} \times n = \chi^2_{2n}$

5 Loi uniforme $\mathcal{U}(a, b)$

Une *v.a.* X qui suit la loi $\mathcal{U}(a,b)$ sera notée $\mathcal{U}_{a,b}$.

5.1 Caractérisation

La loi uniforme de paramètres (a,b), où a et b sont deux nombres réels donnés tels que $a < b$, est caractérisée par

– son support $\Delta = [a,b]$

– sa densité de probabilité qui est constante : $f(x) = 1/(b-a) \quad \forall\, x \in \Delta$.

Elle a pour valeur moyenne $m = (a+b)/2$ et pour écart-type $\sigma = (b-a)/\sqrt{12}$.

5.2 La fonction de répartition

Sa fonction de répartition a pour expression

$$F(t) = P(X \leqslant t) = (t-a)/(b-a)\ \forall\, t \in [a,b]$$

Le support $\Delta = [a,b]$ étant déterminé, on constate que la valeur de $P(\alpha \leqslant X \leqslant \beta)$ ne dépend que de la longueur de l'intervalle $[\alpha,\beta]$ inclus dans Δ.

5.3 Réalisation d'échantillons fictifs

Soit une *v.a.* X qui suit la loi $\mathcal{U}(a,b)$, la *v.a.* $Y = (X-a)/(b-a)$ suit la loi uniforme $\mathcal{U}(0,1)$. Cette propriété permet d'obtenir des échantillons fictifs de la loi $\mathcal{U}(a,b)$.

Exemple de procédé de simulation

La table de nombres au hasard fournit des échantillons fictifs de la loi $\mathcal{U}(0,1)$. Pour obtenir par exemple un échantillon fictif de taille 5 de cette loi considérons la colonne figurant page 377. Si l'on prend la colonne de 4 chiffres située en haut à gauche, on obtient l'échantillon fictif $y_1 = 0{,}1340$; $y_2 = 0{,}5027$; $y_3 = 0{,}8498$; $y_4 = 0{,}2211$; $y_5 = 0{,}6864$. Compte tenu de la propriété ci-dessus on constate que les nombres $x_i = (b-a)y_i + a$ fournissent la réalisation d'un échantillon fictif de taille 5 de la loi $\mathcal{U}(a,b)$. Notons que certains logiciels fournissent ce type d'échantillon fictif.

Cette propriété est en fait un cas particulier de celle présentée ci-après.

© Dunod. La photocopie non autorisée est un délit.

Théorème. Soit une *v.a.* X qui suit une loi continue $\mathcal{L}$ dont le support Δ est un intervalle et dont la fonction de répartition est notée $F : F(t) = P(X \leqslant t)$. Alors la *v.a.* $Y = F(X)$ suit la loi uniforme $\mathcal{U}(0,1)$.

Application. On souhaite obtenir un échantillon fictif de 5 valeurs prises par une *v.a.* X qui suit la loi exponentielle Exp(1). Cette loi a pour support $\Delta = [0,\infty[$ et pour fonction de répartition $F(t) = 1 - \mathrm{e}^{-t}\ \forall t \in [0,\infty[$. La variable $Y = 1 - \mathrm{e}^{-X}$ suit donc la loi uniforme $\mathcal{U}(0,1)$. Ayant fait choix d'un échantillon fictif de 5 valeurs prises par Y : $y_1 = 0{,}1455$; $y_2 = 0{,}2734$; $y_3 = 0{,}1392$; $y_4 = 0{,}2386$, $y_5 = 0{,}1252$ (*cf.* exemple ci-dessus) on obtient un échantillon fictif de 5 valeurs prises par $X = -\ln(1-Y)$: $x_1 = 0{,}1572$; $x_2 = 0{,}3194$; $x_3 = 0{,}1499$; $x_4 = 0{,}2726$; $x_5 = 0{,}1338$.

6 Loi de Fisher-Snédécor $\mathcal{F}(m, n)$

Une *v.a.* qui suit la loi $\mathcal{F}(m,n)$ sera notée $\mathcal{F}_n^m$.

6.1 Caractérisation

La loi de Fisher-Snédécor $\mathcal{F}(m,n)$, où m et n désignent deux entiers naturels strictement positifs, est caractérisée par

– son support $\Delta = [0,\infty[$ et

– sa densité de probabilité $f(x) = K(m,n) \times x^{m/2-1}/(mx+n)^{(m+n)/2}$

où
$$K(m,n) = \frac{m^{m/2} n^{n/2} \times \left(\dfrac{m+n}{2} - 1\right)!}{\left(\dfrac{m}{2} - 1\right)! \times \left(\dfrac{n}{2} - 1\right)!}.$$

La moyenne et la variance ont respectivement pour valeurs

$$E(\mathcal{F}_n^m) = n/(n-2)\ ;\ V(\mathcal{F}_n^m) = 2n^2(m+n-2)/m(n-4)(n-2)^2$$

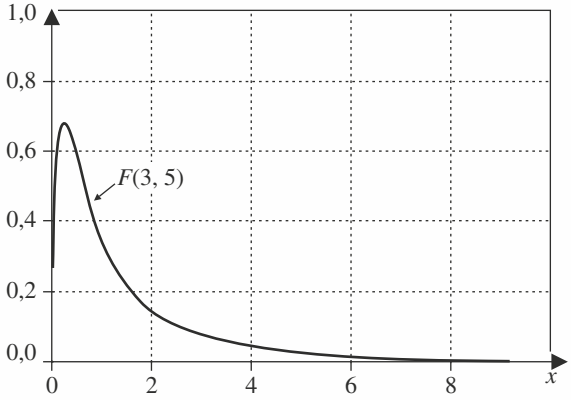


Figure 9.7 – *Fonction de densité de la loi $\mathcal{F}(3, 5)$*

Propriété fondamentale. χ^2_m et χ^2_n désignant deux variables khi-deux indépendantes, la variable quotient $Z = (\chi^2_m/m)/(\chi^2_n/n)$ suit la loi $\mathcal{F}(m,n)$. Autrement dit, $\dfrac{\hat{\chi}^2_m/m}{\hat{\chi}^2_n/n} = \mathcal{F}^m_n$.

6.2 Propriétés de la fonction de répartition

– Le lecteur trouvera page 379 la valeur de t pour laquelle $P(\mathcal{F}^m_n \geqslant t) = 5\ \%$. Pour trouver t tel que $P(\mathcal{F}^m_n \leqslant t) = 5\ \%$ ou $1\ \%$, on utilise la propriété :

$$P(\mathcal{F}^m_n \leqslant t) = P(\mathcal{F}^n_m \geqslant 1/t)\ \forall\, t > 0.$$

6.3 Échantillons issus de deux distributions normales et loi de Fisher-Snédécor

Les principales utilisations des lois de Fisher-Snédécor s'appuient sur le théorème suivant :

Théorème. Soit un échantillon $X_1, X_2, \ldots, X_m$ *iid* d'une loi normale $\mathcal{N}(m_1; \sigma^2_1)$ et soit un autre échantillon $Y_1, Y_2, \ldots, Y_n$ indépendant du précédent et *iid* d'une loi normale $\mathcal{N}(m_2, \sigma^2_2)$. S^2_X et S^2_Y désignant respectivement les variances standard de ces deux échantillons, on a la propriété : $\dfrac{S^2_X}{S^2_Y} \times \dfrac{\sigma^2_2}{\sigma^2_1} = \mathcal{F}^{m-1}_{n-1}$.

7 Loi de Cauchy $\mathcal{C}(m_e, \rho)$

Une *v.a.* notée $\mathcal{C}_{m,\rho}$ est une *v.a.* qui suit la loi $\mathcal{C}(m_e, \rho)$.

7.1 Caractérisation

La loi de Cauchy de paramètres (m_e, ρ), où m_e et ρ désignent deux réels donnés avec $\rho > 0$, est caractérisée par son support $\Delta = \mathbb{R}$ et sa fonction de densité :

$$f(x) = \frac{1}{\pi} \times \frac{1}{\rho} \times \frac{1}{1 + ((x - m_e)/\rho)^2}\ \forall\, x \in \Delta.$$

7.2 Fonction de répartition

Sa fonction de répartition a pour expression

$$F(t) = P(X \leqslant t) = 0{,}5 + \frac{1}{\pi} \times \text{Arctan}\left(\frac{t - m_e}{\rho}\right)\ \forall\, t \in \mathbb{R}$$

Pour $t = m_e$, on a $\mathcal{F}(m_e) = 0{,}5$ autrement dit m_e est la valeur médiane de la fonction de répartition. De plus $\mathcal{F}(m_e + \rho) = 0{,}75$ et $\mathcal{F}(m_e - \rho) = 0{,}25$. Donc 2ρ est la longueur de l'intervalle interquartile.

© Dunod. La photocopie non autorisée est un délit.

La valeur moyenne de cette loi n'étant pas définie, on l'assimile à sa valeur médiane en raison de la symétrie de la densité de probabilité $f(x)$ autour de la valeur m_e. On a évidemment $\mathcal{F}(m_e + t) + \mathcal{F}(m_e - t) = 1 \; \forall \, t \in \mathbb{R}$.

La variance a une valeur infinie.

Exemple

La distribution des commissions perçues par les représentants médicaux de la firme AZ peut être approximée par une distribution de Cauchy. La valeur médiane est égale à 4 K-euros et l'intervalle interquartile est égal à 2. Autrement dit, la commission perçue par les représentants suit sensiblement la loi $\mathcal{C}(4,1)$ puisque $m_e = 4$ et $2\rho = 2$. Alors la proportion de représentants qui gagnent une commission inférieure 1 K-euros est sensiblement égale à $\mathcal{F}(1) = 0{,}5 + \frac{1}{\pi} \times \text{Arctan}\left(\frac{1-4}{1}\right) = 0{,}10$. Est représentée ci-après la fonction de densité de la loi de Cauchy $\mathcal{C}(4,1)$.

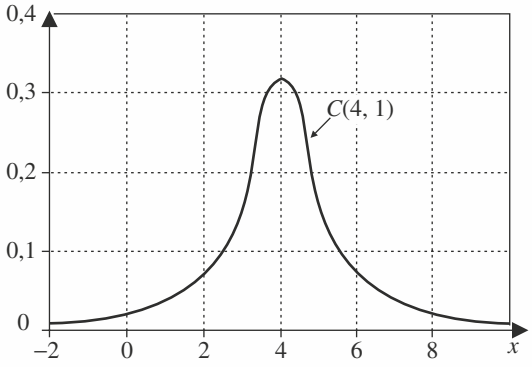


Figure 9.8 – *Fonction de densité de la loi $\mathcal{C}(4, 1)$*

7.3 Somme de variables de Cauchy indépendantes

Soient deux variables de Cauchy indépendantes X_1 et X_2 : $X_1 \sim> \mathcal{C}(m_{e1}, \rho_1)$, $X_2 \sim> \mathcal{C}(m_{e2}, \rho_2)$. Alors $Z = X_1 + X_2$ suit la loi de Cauchy $\mathcal{C}(m_{e1} + m_{e2}, \rho_1 + \rho_2)$.

Cette propriété peut être généralisée à une somme de ν variables de Cauchy indépendantes : $\mathcal{C}_{m_1,\rho_1} + \mathcal{C}_{m_2,\rho_2} + \ldots + \mathcal{C}_{m_n,\rho_n} = \mathcal{C}_{m^*,\rho^*}$ où $m^* = \sum_{i=1}^{n} m_i$ et $\rho^* = \sum_{i=1}^{n} \rho_i$.

8 Loi bêta β (*a*, *b*)

Une *v.a.* notée $\beta_{a,b}$ est une variable qui suit la loi $\beta(a,b)$.

8.1 Caractérisation

La loi bêta du premier genre de paramètres (a,b), où a et b désignent deux réels positifs donnés, est caractérisée par son support $\Delta = [0,1]$ et sa densité de probabilité $f(x) = k \times x^{a-1}(1-x)^{b-1}$ où la constante $k = (a+b-1)!/(a-1)! \times (b-1)!$

La valeur moyenne et la variance ont respectivement pour valeurs

$$E[\beta_{a,b}] = \frac{a}{a+b} \; ; \; V[\beta_{a,b}] = \frac{ab}{(a+b)^2(a+b+1)}$$

Propriété fondamentale liant variable bêta et variable Fisher-Snédécor :

$$\frac{\beta_{a,b}}{1-\beta_{a,b}} = \frac{a}{b}\mathcal{F}_{2a}^{2b}$$

8.2 La fonction de répartition

La fonction de répartition $F_{a,b}(t) = P(\beta_{a,b} \leqslant t)$ est tabulée pour certaines valeurs de $t \in [0,1]$. Les valeurs de la fonction de répartition $F_{a,b}(t) = P(\beta_{a,b} \leqslant t)$ peuvent être obtenues en utilisant les logiciels Excel ou SPSS.

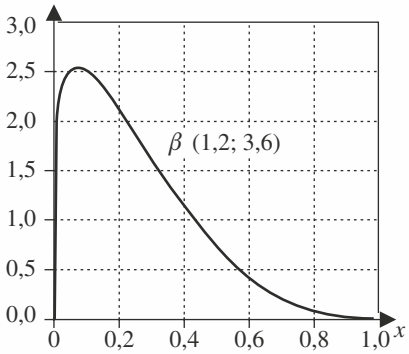


Figure 9.9 – *Fonction de densité de la loi* $\beta_{1,2,\ 3,6}$

9 Loi logistique $\mathcal{L}_g(m, \rho)$

9.1 Caractérisation

La loi logistique de paramètres (m,ρ), où m et ρ désignent deux réels avec $\rho > 0$, est caractérisée par son support $\Delta = \mathbb{R}$ et sa fonction de répartition.

9.2 Fonction de répartition

La fonction de répartition a pour expression $F(t) = 1/\left(1 + e^{-\frac{t-m}{\rho}}\right) \; \forall\, t \in \mathbb{R}$.

La symétrie de $f(x)$ autour de $x = m$ implique : $F(m+t) + F(m-t) = 1 \; \forall\, t$.

© Dunod. La photocopie non autorisée est un délit.

La valeur moyenne et la variance de cette distribution ont respectivement pour valeur m et $(\pi\rho)^2/3$.

Autrement dit, si $X \sim> \mathcal{L}(m,\rho)$ on a $E(X) = m$ et $V(X) = (\pi\rho)^2/3$.

Application. Le taux de croissance des ventes évalué en % est supposé suivre une loi logistique. La moyenne est égale 4 l'an et la variance est supposée égale à 7,4. On veut déterminer le risque d'une croissance négative des ventes au cours d'une année. Le taux de croissance aléatoire X suit la loi de paramètres $(m = 4, \rho = 1{,}5)$ puisque $V(X) = (\pi\rho)^2/3 = 7{,}4$.

Donc $P(X < 0) = F(0) = 1/\left(1 + \mathrm{e}^{\frac{0-m}{1,5}}\right) = 0{,}065$.

9.3 Cas de la forme réduite

Si $X_1 \sim> \mathcal{L}_g(m,\rho)$, la *v.a.* $Y = (X - m)/\rho \sim> \mathcal{L}(0,1)$ dont la fonction de répartition $G(y) = \dfrac{1}{1 + \mathrm{e}^{-y}}$. Cette loi est souvent utilisée pour modéliser des systèmes où la réponse est de type binaire (*cf.* chapitre 16). Remarquer que $y = \ln(G(y)/[1 - G(y)])$.

10 Loi log-normale ou loi de Galton LN(m, σ)

10.1 Caractérisation

La loi log-normale de paramètres (m,σ), où m et σ désignent deux réels donnés avec $\sigma > 0$, est caractérisée par son support $\Delta = [0; \infty[$ et sa densité de probabilité

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \times \mathrm{e}^{-\frac{(\ln(x)-a)^2}{2\sigma^2}} \quad \forall\, x > 0$$

Propriété fondamentale. Si Y suit la loi normale $\mathcal{N}(m,\sigma^2)$ alors $X = \mathrm{e}^Y$ suit loi log-normale de paramètres m,σ : soit $X \sim> \mathrm{LN}(m,\sigma)$.

10.2 La fonction de répartition

On a $F(t) = P(X \leqslant t) = \Phi\left(\dfrac{\ln(t) - m}{\sigma}\right) \ \forall\, t > 0$, où $\Phi(t) = P(\mathcal{N}_{0;1} \leqslant t)$.

10.3 Propriété

Soit n *v.a.* indépendantes $X_1, X_2, \ldots, X_n$. Si n est grand, $Z = (X_1 \times X_2 \times \ldots X_n)^{1/n}$ suit sensiblement une loi LN(m^*,σ^*) où $m^* = \left(\sum_i E[(\ln X_i]\right)/n$ et $\sigma^{*2} = \left(\sum_i V[(\ln X_i]\right)/n^2$.

En effet $\ln(Z)$ suit sensiblement la loi $\mathcal{N}(m^*; \sigma^*)$.

Exemple

Suite à une étude faite sur un nombre important de composants on estime que la durée de vie Y d'un composant en année suit la loi LN(2; 1,5).

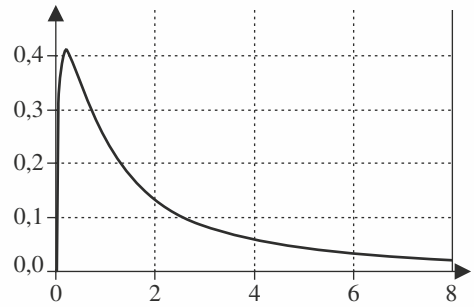


Figure 9.10 – *Fonction de densité de la loi* LN*(2,1.5)*

La probabilité pour que le composant tombe en panne avant un an est $P(Y < 1) = P(e^X < 1) = P(X < 0) = P(\mathcal{N}_{2;1,5} < 0) = \Phi[(0-2)/1,5] = 0,2571$

11 Loi de Weibull $\mathcal{W}$ (*a*, *b*)

Elle se caractérise par deux paramètres a et b qui sont des constantes positives.

Caractérisation

La loi $\mathcal{W}(a,b)$ est caractérisée par son support $\Delta = [0,\infty[$ et sa fonction de répartition : $F(t) = 1 - e^{-(x/a)^b}\ \forall\, t \geqslant 0$.

Cette loi a pour valeur moyenne $m = a \times \beta!$ où $\beta = 1/b$, pour moment d'ordre 2 $m^2 = a^2 \times (2\beta)!$ et donc pour variance $\sigma^2 = a^2 \times [(2\beta)! - (\beta!)^2]$.

Propriété fondamentale. Si X suit loi $\mathcal{W}(a,b)$ alors $Y = (X/a)^b$ suit loi Exp(1).
A contrario si Y suit la loi Exp(1), alors $X = aY^{1/b}$ suit la loi $\mathcal{W}(a,b)$.

Cette loi est utilisée pour étudier certains phénomènes d'usure.

Application. Le temps (exprimé en semaines) qui sépare deux pannes, suite à l'usure de pièces constituant la machine, est supposée suivre une loi de Weibull de moyenne $m = 20$ et d'écart-type $\sigma = 10,4$ et donc de paramètres $a \cong 22,5$ et $b \cong 2$. La probabilité qu'une panne survienne dans moins de sept semaines est de $F(7) = 1 - e^{-(7/22,5)^2} = 0,092$.

© Dunod. La photocopie non autorisée est un délit.

Section 4

LOIS MULTINOMIALES

1 Caractérisation et interprétation

L'ensemble des résultats possibles d'une expérience élémentaire est réparti en h classes : $C_1, C_2, \ldots, C_h$. La probabilité que le résultat X de l'expérience élémentaire envisagée appartienne à la classe C_i est notée p_i : $P(X \in C_i) = p_i$ avec $\sum_{i=1}^{h} p_i = 1$.

On se propose de renouveler n fois cette expérience élémentaire dans des conditions identiques. Le nombre aléatoire de résultats qui appartiendront à la classe C_i étant noté N_i on a évidemment $\sum_{i=1}^{h} N_i = n$.

Le h-uple aléatoire $(N_1, N_2, \ldots, N_h)$ satisfait à la loi de probabilité suivante dite loi multinomiale de dimension h (on dit aussi h-nomiale) :

« quel que soit le h-uple d'entiers naturels $(n_1, n_2, \ldots, n_h)$ vérifiant $\sum_{i=1}^{h} n_i = n$

$$P[(N_1, N_2, \ldots, N_h) = (n_1, n_2, \ldots, n_h)] = \frac{n!}{n_1! n_2! \ldots n_h!} p_1^{n_1} p_2^{n_2} \ldots p_h^{n_h} \text{ »}$$

On a $V(N_i) = np_i(1 - p_i)$ et $\mathrm{Cov}(N_i, N_j) = -np_i p_j$

2 Théorème fondamental

Lorsque $np_i \geqslant 5\ \forall i = 1, 2, \ldots, h$ on a l'approximation : $\sum_{i=1}^{h} \frac{(N_i - np_i)^2}{np_i} \cong \chi^2_{h-1}$.

Autrement dit, $P\left(\sum_{i=1}^{h} \frac{(N_i - np_i)^2}{np_i} \leqslant t\right) \cong P(\chi^2_{h-1} \leqslant t)\ \forall t > 0$.

La valeur numérique z que prend la *v.a.* $Z_n = \sum_{i=1}^{h} \frac{(N_i - np_i)^2}{np_i}$ peut s'exprimer à l'aide des « effectifs observés n_i » et des « effectifs dits théoriques np_i » des classes C_i : $z = \sum_{i=1}^{h} \frac{(n_i - np_i)^2}{np_i}$.

Généralisation. Considérons une partie J de $\{1, 2 \ldots, h\}$ ayant k éléments ($k \geqslant 2$) et les sommes qui lui sont associées : $p_J = \sum_{i \in J} p_i$ et $N_J = \sum_{i \in J} N_i$. Lorsque $np_i \geqslant 5$ $\forall i \in J$ on a $\sum_{i \in J} \frac{(N_i - N_J p_i / p_J)^2}{N_i} \cong \chi^2_{k-1}$.

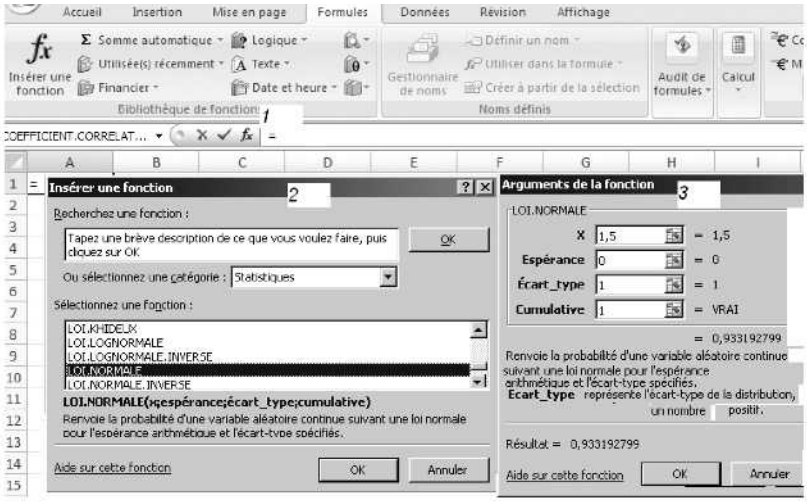
En choisissant par exemple $J = \{1,2\}$, cette propriété permet de trouver un intervalle de confiance du quotient $\omega = p_2/p_1$ ou de tester H_0 « $p_2 = \omega p_1$ » (ω constante donnée) puisque alors $p_1/p_J = (1+\omega)^{-1}$, $p_2/p_J = \omega/(1+\omega)$.

Section 5

PROCÉDURES AVEC EXCEL, SPSS

1 Traitements avec EXCEL

On souhaite connaître la valeur prise par la fonction de répartition $\Phi(t)$ de la loi normale centrée réduite soit $\mathcal{N}(0;1)$ pour $t = 1{,}5$ soit $\Phi(1{,}5)$.



Procédure.

1. On clique sur (fx)

2. On sélectionne la catégorie (Statistiques) et la fonction souhaitée soit ici (LOI.NORMALE), Puis on clique sur (OK).

3. Dans le menu arguments de la fonction on tape $(x = 1{,}5)$ puisque l'on calcule $P(X \leqslant t)$ avec $t = 1{,}5$, (espérance = m = 0), (écart_type = σ = 1) puisqu'il s'agit de la loi $\mathcal{N}(0;1)$, (Cumulative = 1) puisqu'il s'agit de la fonction de répartition de la loi normale (*a contrario* si l'on avait indiqué (Cumulative = 0) on aurait eu la valeur de la fonction de densité $f(x)$ pour $x = 1{,}5$). On obtient $P(X \leqslant 1{,}5) = 0{,}9331$.

L'on peut plus rapidement calculer cette probabilité en cliquant sur fx, et en tapant la formule LOI.NORMALE(x; espérance; écart-type; cumulative) avec ici x=1.5 ; espérance = 0 ; écart-type =1 ; cumulative=1

Excel permet de calculer pour les différentes lois étudiées $\boldsymbol{P(X = t)}$ et $\boldsymbol{P(X \leqslant t)}$ pour une loi discrète ou $\boldsymbol{f(t)}$ la valeur de la fonction de densité en $\boldsymbol{t}$ pour une loi continue. Notations sous excel.

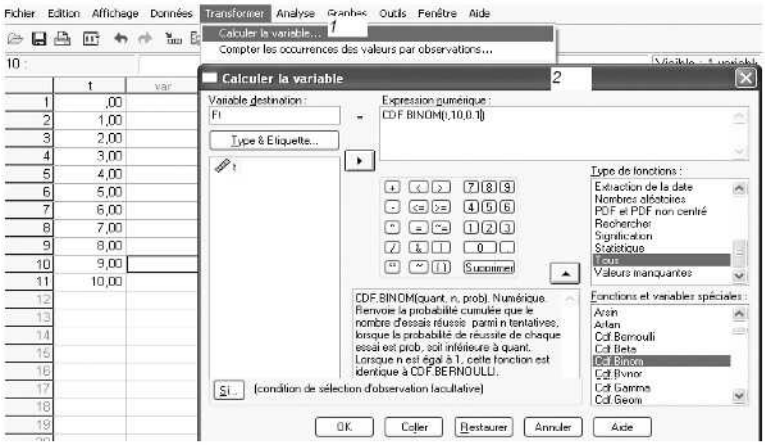
© Dunod. La photocopie non autorisée est un délit.

Se fixant *a priori* une probabilité α on peut déterminer la valeur t_α vérifiant $\alpha = \boldsymbol{P}(\boldsymbol{X} \leqslant \boldsymbol{t}_\alpha)$pour la loi LogNormale, Normale, de Student-Fisher de Fisher-Snédécor, de Weibull.

2 Traitements avec SPSS

Exemple

L'on souhaite connaître les valeurs prises par $F(t) = P(X \leqslant t)$ la fonction de répartition de la loi binomiale de paramètres 10, 0,1 en un point t du domaine.



Procédure.

1. On clique sur (Transformer)

2. On sélectionne (Calculer) et apparait (Calculer la variable)

3. Dans le menu **Calculer la variable** on sélectionne dans type de fonctions (Tous), on entre le nom donné à la **variable destination** (ici la fonction de répartition $F(t) = P(X \leqslant t)$ notée Ft). Dans fonctions et variables spéciales on saisit la fonction de répartition de la loi souhaitée soit ici la loi binomiale de paramètres 10, 0,1 : (Cdf.Binom(quant, *n*, prob)). La valeur *quant* correspond à t, $n = 10$ (le paramètre m_0) et *prob* = 0,1 (la probabilité p) .

Sont présentées ci-après les autres fonctions disponibles sur SPSS

– Les fonctions de SPSS commençant par CDF donnent les valeurs de la fonction de répartition d'une variable aléatoire qui suit une loi spécifiée. Ces fonctions permettent de calculer $\boldsymbol{P}(\boldsymbol{X} \leqslant \boldsymbol{quant})$ la probabilité qu'une variable aléatoire avec la distribution spécifiée soit inférieure à (quant) le premier argument de la fonction proposée par SPSS (autrement dit il s'agit du calcul de $F(t) = P(X \leqslant t)$ $\boldsymbol{t}$ étant ici désigné par la valeur ou la variable ***quant***). Les arguments ultérieurs des fonctions sont les paramètres de la distribution.

– Les fonctions de SPSS commençant par IDF correspondent aux fonctions de répartition inverse. Se fixant *a priori* une probabilité ***prob*** on détermine la valeur $\boldsymbol{t}$ vérifiant $\boldsymbol{prob} = \boldsymbol{P}(\boldsymbol{X} \leqslant \boldsymbol{t})$

Exercices (corrections fin d'ouvrage p. 349)

Exercice 1

Les erreurs algébriques des soldes des comptes clients suivent une loi normale de moyenne 0 et d'écart-type 5 euros. Calculez la probabilité pour que l'erreur soit de :

a) moins de 10 euros ; b) moins de 5 euros ; c) moins de – 5 euros ; d) entre – 5 et 10 euros.

Exercice 2

La demande d'un produit suit sensiblement la loi normale de moyenne 5 000 et d'écart-type 1 000.

1. Calculez la probabilité pour que la demande soit supérieure à 5 500.

2. Quel niveau de stock doit être maintenu pour que la demande soit satisfaite dans 90 % des cas ?

Exercice 3

5 % des pièces sorties d'un atelier d'usinage sont défectueuses.

1. On constitue un lot de 15 pièces. Quelle est la probabilité pour qu'il y ait dans ce lot

a) une seule pièce défectueuse? b) au moins deux pièces défectueuses ?

2. Quelle est la probabilité pour que dans un lot de 1 000 pièces, il y ait au plus 30 pièces défectueuses?

Exercice 4

Des étudiants employés dans un Call-Center pour faire du marketing téléphonique savent que la probabilité de vendre un article ménager suite à un coup de téléphone est égale à 0,01. Déterminer le nombre d'appels nécessaires pour que l'étudiant vendeur soit sûr à 95 % de vendre un article.

Exercice 5

Dans une unité de production, 220 postes de travail doivent fonctionner simultanément et en permanence. Chaque poste est occupé par un employé spécialisé. Pour pallier les absences éventuelles, la direction de ce service de production a créé une équipe E de n_0 employés, chacun d'eux pouvant occuper n'importe quel poste de travail. Le nombre X d'absences journalières, parmi les 220 employés affectés à des postes fixes, suit la loi de Poisson de moyenne égale à 17.

a) Déterminer le nombre minimum n_0 nécessaire pour que tous les postes de travail soient pourvus avec une probabilité de 0,96 si on néglige les absences possibles d'employés appartenant à l'équipe E.

b) Calculer la probabilité réelle d'interruption de la production sachant que $n = n_0$ et que la probabilité d'absence d'un individu appartenant à l'équipe E est de 0,04 [N désignant le nombre d'absents dans l'équipe E, remarquer que N suit sensiblement une loi de Poisson et qu'il y a interruption si $X + N > n_0$].

© Dunod. La photocopie non autorisée est un délit.

10 ESTIMATION PONCTUELLE ET INTERVALLE DE CONFIANCE

Partant d'un échantillon de taille n extrait d'une population statistique $\boldsymbol{P}$, on peut donner une estimation ponctuelle de la valeur de la moyenne, de la médiane, de l'écart-type ou de tout autre valeur d'un paramètre ℓ associé à la distribution sur cette population. On peut plus précisément, en fonction d'un risque d'erreur α fixé à l'avance (souvent 5 % ou 1 %) fournir une fourchette de valeurs possibles appartenant à un intervalle $]a,b[$ dit *intervalle de confiance* dont on est sûr, avec la probabilité $1-\alpha$ (et donc souvent 95 % ou 99 %), qu'il contienne la valeur du paramètre ℓ. Ce paramètre peut également être celui d'une distribution dont la nature est prédéfinie : la moyenne ou l'écart-type d'une distribution normale, le paramètre d'une loi de Poisson, le paramètre d'une loi exponentielle, la proportion p sur une population binomiale.

Section 1 ■ Variable aléatoire définie sur une population statistique $\boldsymbol{P}$

Section 2 ■ Constitution d'un échantillon

Section 3 ■ Estimation ponctuelle des paramètres

Section 4 ■ Estimation par intervalle de confiance

Section 5 ■ Estimation avec Excel et SPPS

Section 1

VARIABLE ALÉATOIRE DÉFINIE SUR UNE POPULATION STATISTIQUE *P*

1 Distribution de la mesure d'un caractère sur une population *P*

On s'intéresse à la mesure numérique $\boldsymbol{x}$ d'un caractère défini sur une population $\boldsymbol{P} = \{e_1, e_2, ..., e_N\}$ constituée de N éléments. À chaque élément e_i correspond sa mesure $\boldsymbol{x}(e_i)$, on a donc la série de N observations $x_1, x_2, \ldots, x_N$.

La série peut être présentée sous la forme d'un tableau où figurent les différentes valeurs x_i^* de la variable et le nombre de fois N_i où ces valeurs ont été observées, le nombre total d'observations étant égal à N (*cf.* chapitre 1 série sur une population).

Tableau 10.1 – *Distribution sur une population statistique P*

Valeurs prises par x	x_1^*	x_2^*	...	x_ν^*	**Total**
Effectifs N_i	N_1	N_2	...	N_ν	$N = \sum_{i=1}^{\nu} N_i$
Fréquences $p_i = N_i/N$	$p_1 = N_1/N$	$p_2 = N_2/N$	...	$p_\nu = N_\nu/N$	1
Fréquences relatives cumulées $F(x_i^*) = \pi_i$	N_1/N	$(N_1 + N_2)/N$	...	$(N_1 + N_2 + \ldots + N_\nu)/N$	

La population $\boldsymbol{P}$ est de fait partagée en ν classes $C_1, C_2, \ldots, C_\nu$ où C_h regroupe tous les éléments de $\boldsymbol{P}$ dont la mesure est égale à x_h^* : $C_h = \{e_i \in P / \boldsymbol{x}(e_i) = x_h^*\}$.

1.1 Les caractéristiques essentielles d'une distribution sur une population *P*

Sont considérées comme caractéristique de la distribution sur la population :

– la valeur moyenne $m_P = \dfrac{1}{N} \sum_{i=1}^{\nu} N_i x_i^*$ et les moments d'ordre h : $m_h = \dfrac{1}{N} \sum_{i=1}^{\nu} N_i (x_i^*)^h$, ($h$ entier naturel)

– la variance $\nu_P = \dfrac{1}{N} \sum_{i=1}^{\nu} N_i (x_i^* - m_P)^2$ (ou son écart-type $\sigma_P = \sqrt{\nu_P}$) et plus généralement les moments centrés d'ordre h : $\mu_{h,P} = \dfrac{1}{N} \sum_{i=1}^{\nu} N_i (x_i^* - m_P)^h$,

– la médiane et les quartiles ; la proportion dans le cas d'une population binomiale.

© Dunod. La photocopie non autorisée est un délit.

1.2 La fonction de répartition

La fonction de répartition F des N valeurs x_i, définie pour chaque $t \in \mathbb{R}$ par :
$F(t) =$(nombre de valeurs x_i telles que $x_i \leqslant t)/N =$ card.D_t/card.$\boldsymbol{P}$
où $D_t = \{$ensemble des éléments e_i de $\boldsymbol{P}$ dont la valeur $x_i \leqslant t\ \}$
caractérise complètement la répartition des N valeurs x_i dans $\boldsymbol{P}$ (*cf.* chapitre 1).

2 Tirage au hasard d'un élément dans la population *P*

Prélevant *au hasard* un élément e dans $\boldsymbol{P}$, la mesure aléatoire X du caractère considéré associée à cet élément e est une variable aléatoire qui satisfait à la loi de probabilité $\mathcal{L}_P$ caractérisée par :

$$\Delta_X = \{x_1^*, x_2^*, \ldots, x_\nu^*\},$$

$$P(X = x_h^*) = p(e \in C_h) = N_h/N = p_h \quad \forall x_h^* \in \Delta_X.$$

Autrement dit, les poids associés à chacune des valeurs x_1^*, x_2^*, …, x_ν^* correspondent aux fréquences relatives des différentes valeurs sur la population. On a :

$$E(X) = x_1^* \times (N_1/N) + x_2^* \times (N_2/N) + \ldots + x_\nu^* \times (N_\nu/N) = m_P$$

$$V(X) = (x_1^* - m_P)^2 \times (N_1/N) + \ldots + (x_\nu^* - m_P)^2 \times (N_\nu/N) = \nu_P.$$

La mesure aléatoire X associée à cet élément e suit la loi de probabilité caractérisée par $F(t)$, la fonction de répartition empirique sur $\boldsymbol{P}$. En effet,

$$(X(e) \leqslant t) \Longleftrightarrow (e \in D_t) \text{ où } D_t = \{e_i \in P / x(e_i) \leqslant t\}) \text{ et donc}$$

$$P(X \leqslant t) = P(e \in D_t) = \text{card.}D_t/\text{card.}\boldsymbol{P}. = F(t)$$

Section 2

CONSTITUTION D'UN ÉCHANTILLON

La constitution d'un échantillon se révèle nécessaire :

– lorsque l'on souhaite connaître certaines valeurs caractéristiques (moyenne, variance …) d'une population statistique $\boldsymbol{P}$ et que l'on ne peut réaliser une enquête systématique auprès de tous les éléments de la population $\boldsymbol{P}$;

– dans le cadre d'expérimentations réalisées dans des conditions identiques afin de déterminer les caractéristiques essentielles de la loi que suit le résultat X d'une expérience envisagée.

1 Échantillon prélevé sur une population statistique

1.1 Définition

Sonder une population consiste à prélever une fraction de la population $\boldsymbol{P}$, fraction appelée échantillon. Si la population $\boldsymbol{P}$ comporte N éléments, soit $\boldsymbol{P} = \{e_1, e_2, \ldots, e_N\}$ on constitue un échantillon de taille n en prélevant n éléments dans $\boldsymbol{P}$.

Deux procédés peuvent être envisagés pour constituer un échantillon :

– le *sondage simple* par prélèvement au hasard de n éléments dans $\boldsymbol{P}$,

– le *sondage stratifié* par partition de $\boldsymbol{P}$ en ν sous-populations (ou strates) $\boldsymbol{P_1}$, $\boldsymbol{P_2}$, …, $\boldsymbol{P_\nu}$ puis par prélèvement d'un échantillon dans chaque strate de façon à obtenir un échantillon global le plus représentatif possible de la population $\boldsymbol{P}$.

Repères

• *Dans un sondage aléatoire simple* chaque élément de la population a la même chance d'être extrait de la population $\boldsymbol{P}$ et donc de faire partie de l'échantillon. Si la population comprend N individus, chaque individu a une probabilité $1/N$ d'être tiré. Ce sondage aléatoire simple peut être réalisé à partir de tirages avec ou sans remise.

Un échantillon aléatoire avec remise (échantillon *non exhaustif*) est obtenu par prélèvements successifs d'éléments dans la population $\boldsymbol{P}$ où chaque élément prélevé et observé est remis dans la population après son observation, un même élément pouvant donc théoriquement être tiré et analysé plusieurs fois. Un échantillon aléatoire sans remise ou échantillon *exhaustif* est constitué d'éléments obligatoirement différents, un élément une fois tiré n'est pas remis dans la population.

Pour obtenir un échantillon aléatoire simple de taille n extrait d'une population $\boldsymbol{P}$ de taille N on peut, parmi divers procédés, attribuer de façon univoque à chaque élément de $\boldsymbol{P}$ un nombre entier compris entre 1 et N puis prélever au hasard (à l'aide d'une table de nombres au hasard ou d'un générateur de nombres aléatoires) n de ces N nombres entiers. Si dans le prélèvement on élimine tous les numéros déjà sortis, le sondage est exhaustif. Si on veut un échantillon exhaustif, on élimine de la liste les numéros dès la première fois où ils sont tirés.

• *Dans un sondage par strate* le prélèvement dans $\boldsymbol{P}$ consiste à réaliser une partition de la population en sous-populations en fonction d'une ou plusieurs caractéristiques, chaque élément de la population appartenant alors à une et une seule sous-population appelée *strate*. L'échantillon stratifié est constitué de l'ensemble des sous-échantillons aléatoires simples tirés au hasard de chaque strate. Pour déterminer la taille de chaque sous-échantillon le lecteur pourra se référer à l'exercice 5.

Exemple

Une machine a fabriqué 950 pièces au cours de l'heure et l'on veut vérifier la conformité des pièces à l'aide d'un échantillon de taille 10 prélevé au hasard. Pour cela on affecte fictivement à chaque pièce un chiffre compris entre 000 et 949 puis, à l'aide de

© Dunod. La photocopie non autorisée est un délit.

la table de nombres au hasard p. 377, on lit 10 valeurs en prenant les chiffres trois par trois. Si l'on commence au début on lit successivement 134 076 289 978 937 905 252 503 356 358. Ayant exclut la valeur 979 on prend un chiffre supplémentaire le 789.

1.2 Distribution d'un échantillon prélevé sur une population statistique

On prélève successivement *au hasard et avec remise* un échantillon $\{e'_1, e'_2, \ldots, e'_n\}$ de taille n.

– Au premier élément e'_1 tiré au hasard dans $\boldsymbol{P}$ va correspondre la mesure aléatoire X_1 du caractère considéré,
– au deuxième élément e'_2 va correspondre la mesure aléatoire X_2 du caractère considéré, …,
– et au n-ième élément e'_n, sa mesure X_n.

On a ainsi un échantillon de n *v.a.* indépendantes $X_1, X_2, \ldots, X_n$ qui suivent la loi de probabilité $\mathcal{L}_P$, c'est-à-dire *iid* sur $\mathcal{L}_P$.

Cette loi $\mathcal{L}_P$ est caractérisée par la distribution de probabilité : $\Delta_X = \{x_1^*, x_2^*, \ldots, x_\nu^*\}$, $P(X = x_h^*) = N_h/N = p_h \quad \forall h \in \Delta_X$.

On a évidemment $E(X_i) = E(X) = m_P$, $V(X_i) = V(X) = \nu_P$.

2 Échantillon associé à *n* expérimentations identiquement réalisées

On s'intéresse à la mesure numérique X d'un caractère associé à un certain type d'expérimentation. Pour ce faire on se propose de réaliser n expériences $e'_1, e'_2, \ldots, e'_n$ dans des conditions identiques. À la i-ème expérimentation e'_i est associée la *v.a.* X_i qui, après réalisation de l'expérience, prend la valeur numérique x_i. On obtient donc un échantillon $X_1, X_2, \ldots, X_n$ *iid* d'une loi $\mathcal{L}$ que suit X.

Exemple

On s'intéresse à la durée de vie T d'un certain type de composant électronique (l'unité de temps étant l'année). Afin d'appréhender la loi $\mathcal{L}(m, \sigma)$ que suit T ou de façon plus modeste pour obtenir une estimation de la valeur moyenne m et de l'écart-type σ de cette loi, on considère les durées de vie $T_1, T_2, \ldots, T_{1\,800}$ de 1 800 composants. Après réalisation des expérimentations, t_i désignant la valeur prise par T_i on constate que $t_1 = 1{,}06$, $t_2 = 0{,}98, \ldots, t_{1\,800} = 1{,}21$. L'échantillon étant de grande taille on peut penser qu'il y a une forte probabilité pour que la valeur exacte de m soit proche de $\bar{t} = (t_1 + t_2 + \ldots + t_{1\,800})/1\,800$.

3 Caractéristiques des échantillons

Soit un échantillon $X_1, X_2, \ldots, X_n$ *iid* d'une loi $\mathcal{L}$ de valeur moyenne m_P, d'écart-type σ_P et dont le moment centré d'ordre 4 est $\mu_{4,P}$. On peut définir

– la moyenne aléatoire de cet échantillon (dite aussi moyenne empirique) :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = (X_1 + X_2 + \ldots + X_n)/n$$

qui a pour espérance $E(\bar{X}_n) = m_P$ et variance $V(\bar{X}_n) = \sigma_P^2/n$ (*cf.* p. 121).

– la variance standard de l'échantillon $S_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ qui a pour espérance $E(S_n^2) = \sigma_P^2$ et variance $V(S_n^2) = \left(\mu_{4,P} - \dfrac{n-3}{n-1}\sigma_P^4\right)/n$

– l'écart-standard (ou *standard-déviation*) $S_n = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$

– le moment empirique centré d'ordre 4 $\hat{\mu}_{4,n} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^4$ qui est un estimateur convergent de $\mu_{4,P}$.

Lorsque la taille n de l'échantillon est fixée, $\bar{X}_n$, S_n^2 et $\hat{\mu}_{4,n}$ seront plus simplement notés $\bar{X}$, S^2 et $\hat{\mu}_4$. Selon la loi $\mathcal{L}$ que suit X, ces caractéristiques suivront tel ou tel type de loi. Dans le cas de grands échantillons, on pourra approcher les lois de ces caractéristiques par des lois normales.

4 Distributions associées à la moyenne aléatoire $\bar{X}$ d'un échantillon *iid*

Soit un échantillon de taille n qui est *iid* sur une loi $\mathcal{L}_P$. Lorsque n est grand ou lorsque l'on connaît la nature de la loi on dispose de distributions appropriées associées à la moyenne aléatoire de l'échantillon $\bar{X}$.

4.1 Cas d'un grand échantillon (distribution asymptotique)

Lorsque n est grand ($n \geqslant 30$), en vertu du théorème Lioupanoff, la distribution de $\bar{X}$ peut être approximée par la distribution normale $\mathcal{N}(m_P; \sigma_P^2/n)$ et l'on en déduit :

$$(i)\ Z = \frac{\bar{X} - m_P}{\sigma_P/\sqrt{n}} \cong \mathcal{N}_{0;1} \quad \text{et} \quad (ii)\ Z' = \frac{\bar{X} - m_P}{S/\sqrt{n}} \cong \mathcal{N}_{0;1}$$

© Dunod. La photocopie non autorisée est un délit.

La statistique *(i)* est employée pour obtenir des intervalles de confiance de la valeur moyenne m_P d'un caractère sur une population lorsque l'on connaît l'écart-type σ_P de la population. En général, ne disposant pas de la valeur de σ_P on l'estime par S et l'on utilise *(ii)*.

4.2 Cas d'un échantillon *iid* issu d'une distribution normale

Lorsque $X_1, X_2, \ldots, X_n$ est un échantillon *iid* d'une loi normale $\mathcal{N}(m, \sigma^2)$, la distribution associée à $\bar{X}$ est parfaitement déterminée et ce, quelque soit la taille de l'échantillon, même si n est petit :

$$(i)\ Z = \frac{\bar{X} - m}{\sigma/\sqrt{n}} = \mathcal{N}_{0;1} \quad (ii)\ Z' = \frac{\bar{X} - m}{S/\sqrt{n}} = \boldsymbol{t}_{n-1} \text{ (variable de Student)}$$

La statistique Z peut être utilisée pour obtenir des intervalles de confiance de la moyenne m de la distribution supposée normale si l'on connaît la valeur de l'écart-type σ. Dans le cas contraire (cas le plus fréquent) on utilise la statistique Z'.

4.3 Cas d'un échantillon *iid* issu d'une distribution exponentielle

Pour un échantillon $X_1, X_2, \ldots, X_n$ *iid* de la loi exponentielle de paramètre a on a : $2an\bar{X} = \chi^2_{2n}$, propriété permettant d'obtenir des intervalles de confiance de a et donc de $m = 1/a$.

5 Distributions associées à la variance standard S^2 d'un échantillon *iid*

Soit un échantillon de taille n qui est *iid* sur une loi $\mathcal{L}$. Lorsque n est grand ou lorsque la loi $\mathcal{L}$ est normale on dispose de distributions appropriées associées à S^2.

5.1 Cas d'un grand échantillon (distribution asymptotique)

Lorsque n est grand ($n \geqslant 30$), en vertu du théorème Lioupanoff, la distribution de S^2 peut être approximée par une distribution normale et l'on en déduit :

$$(j)\ \theta = \frac{S^2 - \sigma_P^2}{\sqrt{(\mu_{4,P} - \sigma_P^4)/n}} \cong \mathcal{N}_{0;1} \text{ et } (jj)\ \theta' = \frac{S^2 - \sigma_P^2}{\sqrt{(\hat{\mu}_4 - S^4)/n}} \cong \mathcal{N}_{0;1}.$$

Pour obtenir des intervalles de confiance de la variance σ_P^2 d'une distribution sur la population on utilise la statistique *(jj)*.

5.2 Cas d'un échantillon *iid* issu d'une distribution normale

Lorsque $X_1, X_2, \ldots, X_n$ est un échantillon *iid* d'une loi normale $\mathcal{N}(m,\sigma^2)$, on a quelque soit la taille n de l'échantillon :

$$(j)\ \ \theta = \sum_{i=1}^{n}[(X_i - m)/\sigma]^2 = \chi^2_n \qquad (jj)\ \ \theta' = \frac{(n-1)S^2}{\sigma^2} = \chi^2_{n-1}$$

Les statistiques θ et θ' sont respectivement utilisées pour obtenir les intervalles de confiance de la variance σ^2 de la distribution supposée normale, selon que l'on connaît ou non la valeur de la moyenne m de la distribution.

6 Distribution associée à la proportion aléatoire *F* d'un échantillon *iid*

Si $X_1, X_2, \ldots, X_n$ est un échantillon *iid* d'une loi de Bernoulli, (autrement dit X_i prend la valeur 1 si le i-ème élément tiré au hasard dans la population a le caractère étudié ou 0 dans le cas contraire) alors la moyenne aléatoire de l'échantillon $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$ correspond en fait à la proportion ou fréquence aléatoire sur l'échantillon et est le plus souvent notée F. Lorsque n est grand, par application du théorème central-limite, on a :

$$(k)\ \ \frac{F-p}{\sqrt{p(1-p)/n}} \cong \mathcal{N}_{0;1} \qquad (kk)\ \ \frac{F-p}{\sqrt{F(1-F)/n}} \cong \mathcal{N}_{0;1}$$

où p est la proportion d'éléments de la population qui ont le caractère étudié.

Ces statistiques sont utilisées pour obtenir des intervalles de confiance de p.

Section 3

ESTIMATION PONCTUELLE DES PARAMÈTRES

Les relations et lois décrites dans le paragraphe précédent sont à la base de l'estimation statistique. Celle-ci se propose en effet d'atteindre, à travers l'examen d'un échantillon, une information quantitative *à propos des paramètres* (essentiellement, moyenne m ou écart-type σ). Ces paramètres sont habituellement inconnus parce qu'il est en général impossible d'analyser la totalité de la population ***P***.

1 Définition

L'estimation est dite *ponctuelle* lorsque l'on se propose de substituer à la valeur ℓ d'un paramètre de ***P*** *un nombre unique*, construit à partir d'un échantillonnage. (exemple : $\bar{x}$ est un estimateur ponctuel de m_P).

© Dunod. La photocopie non autorisée est un délit.

REPÈRES : Choix des estimateurs ponctuels

Pour un paramètre ℓ il est possible d'imaginer plusieurs estimateurs construits à partir d'un même échantillon. Il convient donc de retenir l'estimateur Λ qui donnera la meilleure image de ℓ. Cela suppose que

① *l'espérance mathématique de* Λ *doit être égale à* ℓ : $E(\Lambda) = \ell$, on dit que Λ est un estimateur *sans biais de* ℓ. Par exemple, $\overline{X}$, S^2, F sont des estimateurs sans biais des paramètres respectifs m_p, σ_p^2, p.

② Λ *converge en probabilité vers* ℓ, lorsque la taille n de l'échantillon s'accroît, condition en particulier vérifiée lorsque $E(\Lambda) \longrightarrow \ell$ et $V(\Lambda) \longrightarrow 0$.

③ *la variance de* Λ *soit minimale*. – La taille n de l'échantillon étant fixée, entre plusieurs estimateurs de ℓ qui présentent les mêmes qualités quant au biais et à la convergence, il est logique de préférer celui dont la variance est la plus petite, puisque ses valeurs possibles seront moins dispersées.

Plus précisément, si Λ est un *estimateur sans biais de* ℓ, n étant fixé, on peut montrer qu'il existe pour la variance de Λ une *borne inférieure* (*cf.* annexe p. 372).

Lorsqu'un estimateur sans biais a une *variance égale à la borne inférieure*, il est le meilleur possible. On dit que c'est un ***estimateur efficace***.

2 Estimateurs ponctuels et maximum de vraisemblance

La méthode dite *du maximum de vraisemblance* permet de définir des estimateurs ponctuels dont les propriétés sont celles énoncées précédemment, ou bien s'en rapprochent lorsque la taille de l'échantillon croît. Le procédé consiste à choisir comme estimateur de ℓ la valeur particulière de ℓ (ℓ étant pris comme variable) qui rend maximale la probabilité, ou la densité de probabilité du n-uple $(X_1, X_2, \ldots, X_n)$. La méthode peut se généraliser à la construction simultanée de plusieurs estimateurs ponctuels. (*Cf.* annexe page 371) .

3 Propriétés des estimateurs usuels

Soit un échantillon $X_1, X_2, \ldots, X_n$ *iid* d'une loi $\mathcal{L}(m_P, \sigma_P, \mu_{4,P})$ de valeur moyenne m_P, d'écart-type σ_P et dont le moment centré d'ordre 4 est $\mu_{4,P}$.

– La moyenne aléatoire sur l'échantillon $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$ est un estimateur sans biais et convergent de m_P car $E(\bar{X}) = m_P$ et $V(\bar{X}) = \sigma_P^2/n \longrightarrow 0$ lorsque $n \longrightarrow \infty$.

– La variance standard de l'échantillon S^2 est un estimateur sans biais et convergent de la variance $\nu_P = \sigma_P^2$ car $E(S^2) = \sigma_P^2$ et $V(S^2) = (\mu_{4,P} - \dfrac{n-3}{n-1}\sigma_P^4)/n$ tend vers 0 lorsque $n \longrightarrow \infty$.

– L'écart-type standard (ou *standard- déviation*) S de l'échantillon est un estimateur convergent avec biais de σ_P car en général $E(S) \neq \sigma_P$.

4 Valeurs des estimations ponctuelles

À partir de valeurs numériques observées $x_1, x_2, \ldots, x_n$ d'un échantillon on obtient $\bar{x}$, s et μ_4^* qui sont les valeurs respectivement prises par $\bar{X}$, S et $\hat{\mu}_4$ après analyse de l'échantillon.

On sait que lorsque n est grand, il y a de fortes probabilités pour que $\bar{x} \cong m_P$, $s \cong \sigma_P$ et $\mu_4^* \cong \mu_{4,P}$. Aussi $\bar{x}$ fournit une estimation numérique de m_P et s une estimation numérique de σ_P.

Ces valeurs $\bar{x}$ et s résultent de l'analyse de l'échantillon considéré. Avec un autre échantillon de même taille, on aurait obtenu des valeurs numériques différentes $x'_1, x'_2, \ldots, x'_n$ et donc, sauf cas exceptionnel, $\bar{x}' \neq \bar{x}$ et $s' \neq s$.

Section 4

ESTIMATION PAR INTERVALLE DE CONFIANCE

L'estimation est dite par *intervalle de confiance* lorsque l'on construit à partir de l'échantillon un *intervalle* $]a,b[$ *qui peut contenir le paramètre* ℓ avec une probabilité que l'on se fixe à l'avance. À partir des valeurs $\bar{x}$ et s on peut obtenir, avec un niveau de confiance fixé *a priori*, un encadrement de la valeur de m_P ainsi qu'un encadrement de la valeur de σ_P, ces encadrements étant appelés intervalles de confiance.

1 Intervalle de confiance de la moyenne m_P et de l'écart-type σ_P d'une distribution lorsque l'échantillon est de grande taille

1.1 Intervalle de confiance de m_P

Variable statistique utilisée. L'échantillon $X_1, \ldots, X_n$ étant de grande taille, on utilise l'une ou l'autre des deux propriétés suivantes selon que la valeur de l'écart-type σ_P de la distribution est connue *(i)* $\dfrac{\bar{X} - m_P}{\sigma_P/\sqrt{n}} \cong \mathcal{N}_{0;1}$ ou inconnue *(ii)* $\dfrac{\bar{X} - m_P}{S/\sqrt{n}} \cong \mathcal{N}_{0;1}$ (*cf.* § 4.1 page 157).

© Dunod. La photocopie non autorisée est un délit.

Intervalle de confiance aléatoire. Pour trouver un intervalle de confiance de m_P de niveau $(1-\alpha)$ on cherche les nombres $a_{\alpha/2}$ et $b_{\alpha/2}$ tels que $P(\mathcal{N}_{0;1} < a_{\alpha/2}) = \alpha/2$ et $P(\mathcal{N}_{0;1} > b_{\alpha/2}) = \alpha/2$, ce qui implique que $P(a_{\alpha/2} \leqslant \mathcal{N}_{0;1} \leqslant b_{\alpha/2}) = (1-\alpha)$. En raison de la symétrie de la fonction de densité de $\mathcal{N}_{0;1}$ par rapport à l'axe des ordonnées on a $a_{\alpha/2} = -b_{\alpha/2}$ et donc $P(-b_{\alpha/2} \leqslant \mathcal{N}_{0;1} \leqslant b_{\alpha/2}) = 1-\alpha$.

Par exemple pour un niveau de confiance de 90 % et donc un risque d'erreur $\alpha = 10\% = 0{,}1$ on cherche $b_{\alpha/2}$ tel que $\Phi(b_{\alpha/2}) = 1-\alpha/2 = 0{,}95$ et on lit p. 374 $b_{\alpha/2} = 1{,}64$.

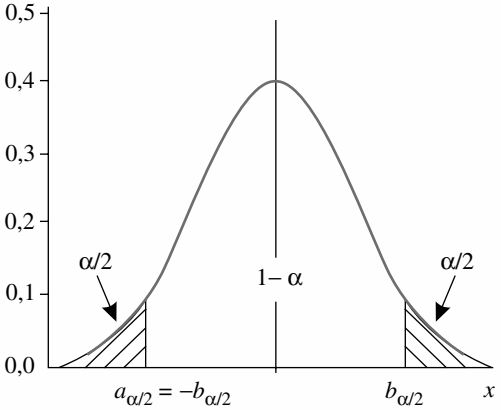


Figure 10.1

Lorsque σ_P a une valeur inconnue, on utilise l'approximation *(ii)*. On a donc

$$1-\alpha = P(-b_{\alpha/2} \leqslant \mathcal{N}_{0;1} \leqslant b_{\alpha/2}) \cong P(-b_{\alpha/2} \leqslant \frac{\bar{X}-m_P}{S/\sqrt{n}} \leqslant b_{\alpha/2})$$

$$= P(\bar{X} - b_{\alpha/2} \times \frac{S}{\sqrt{n}} \leqslant m_P \leqslant \bar{X} + b_{\alpha/2} \times \frac{S}{\sqrt{n}}).$$

L'événement « $\bar{X} - b_{\alpha/2} \times \frac{S}{\sqrt{n}} \leqslant m_P \leqslant \bar{X} + b_{\alpha/2} \times \frac{S}{\sqrt{n}}$ » a une probabilité d'environ $(1-\alpha)$ d'être réalisé.

Une réalisation de l'intervalle. Après expérimentation, constatant que $\bar{X}$ prend la valeur numérique $\bar{x}$ et S prend la valeur s, on peut affirmer avec une probabilité sensiblement égale à $(1-\alpha)$, que l'inégalité

[1] $\qquad \bar{x} - b_{\alpha/2}\, s/\sqrt{n} \leqslant m_P \leqslant \bar{x} + b_{\alpha/2}\, s/\sqrt{n}$ est vraie.

Autrement dit, l'inégalité [1] fournit un « intervalle de confiance de m_P de niveau $(1-\alpha)$ ».

Dans le cas purement théorique en gestion où σ_P a une valeur connue on utilise l'approximation *(i)* $\dfrac{\bar{X}-m_P}{\sigma_P/\sqrt{n}} \cong \mathcal{N}_{0;1}$ qui est plus précise que *(ii)* et l'on en déduit que

l'événement « $\bar{X} - b_{\alpha/2} \times \dfrac{\sigma_P}{\sqrt{n}} \leqslant m_P \leqslant \bar{X} + b_{\alpha/2} \times \dfrac{\sigma_P}{\sqrt{n}}$ » a une probabilité $\cong (1-\alpha)$ d'être réalisée. Puis l'on en déduit une réalisation :

$$\bar{x} - b_{\alpha/2}\, \sigma_P/\sqrt{n} \leqslant m_P \leqslant \bar{x} + b_{\alpha/2}\, \sigma_p/\sqrt{n}\,.$$

Application. On s'intéresse à la moyenne m_P des revenus déclarés pour l'année t par l'ensemble P des ménages fiscaux d'une région donnée. Pour cela on dispose d'un échantillon de 2 500 revenus obtenus par tirage au hasard et de façon non-exhaustive. Ces 2 500 revenus ont une valeur moyenne $\bar{x} = 15{,}1$ unités monétaires (*u.m.*) et un écart-standard $s = 4{,}3$ *u.m.* À partir de ces résultats, on souhaite déterminer un intervalle de confiance de m_P avec un niveau de confiance de 95 % soit $(1-\alpha) = 0{,}95$, le risque d'erreur étant donc de 5 %.

– *Statistique utilisée*. L'échantillon $X_1, \ldots, X_{2\,500}$ étant de grande taille et la valeur de σ_P étant inconnue, on utilise l'approximation $\dfrac{\bar{X} - m_P}{S/\sqrt{n}} \cong \mathcal{N}_{0;1}$.

– *Intervalle de confiance aléatoire*. On cherche le nombre $b_{\alpha/2}$ tel que $\Phi(b_{\alpha/2}) = 1 - \alpha/2 = 0{,}975$, soit par lecture de table $b_{\alpha/2} = 1{,}96$. On en déduit que

$$0{,}95 = P(-1.96 \leqslant \mathcal{N}_{0;1} \leqslant 1{,}96) \cong P(-1.96 \leqslant \frac{\bar{X} - m_P}{S/\sqrt{n}} \leqslant 1{,}96)$$
$$= P(\bar{X} - 1{,}96 \times \frac{S}{\sqrt{n}} \leqslant m_P \leqslant \bar{X} + 1.96 \times \frac{S}{\sqrt{n}}).$$

– *Réalisation, de l'intervalle*. $\bar{X}$ prend la valeur numérique $\bar{x} = 15{,}1$ et S la valeur $s = 4{,}3$. Avec un niveau de confiance de 95 % on a :

$$15{,}1 - 4{,}3 \times 1{,}96/\sqrt{2\,500} \leqslant m_P \leqslant 15{,}1 + 1{,}96 \times 4{,}3/\sqrt{2\,500}$$

soit $14{,}931 \leqslant m_P \leqslant 15{,}268$.

Remarque. Lorsque l'échantillon de taille n est exhaustif, c'est-à-dire extrait au hasard sans remise d'une population $P = \{e_1, e_2, \ldots, e_N\}$ comprenant N éléments, on utilise selon que l'on connaît ou non la valeur de l'écart-type σ_P de la population, l'une ou l'autre des approximations suivantes (exercice 4) :

$$\frac{\bar{X} - m_P}{\frac{\sigma_P}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}} \cong \mathcal{N}_{0;1}\ ; \quad \frac{\bar{X} - m_P}{\frac{S}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}} \cong \mathcal{N}_{0;1}.$$

1.2 Intervalle de confiance de σ_P

À partir des valeurs observées $x_1, x_2, \ldots, x_n$ on peut déterminer non seulement les valeurs estimées $\bar{x}$ et s de m_P et de σ_P mais également le nombre

© Dunod. La photocopie non autorisée est un délit.

$\mu_4^* = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$ qui est une estimation du moment centré d'ordre quatre $\mu_{4,P}$ de la distribution $\mathcal{L}_P$ dont est extrait l'échantillon.

– *Variable statistique utilisée*. L'échantillon $X_1,\ldots,X_n$ étant de grande taille et la valeur de $\mu_{4,P}$ étant généralement inconnue, on utilise l'approximation

$$\frac{S^2 - \sigma_P^2}{\sqrt{(\hat{\mu}_4 - S^4)/n}} = \mathcal{N}_{0;1} \text{ où } \hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4 \text{ est un estimateur de } \mu_{4,P}.$$

– *Intervalle de confiance aléatoire*. Pour trouver un intervalle de confiance de σ_P^2 de niveau $1-\alpha$, on cherche le nombre $b_{\alpha/2}$ tel que $\Phi(b_{\alpha/2}) = 1 - \alpha/2$.

Compte tenu de l'approximation au sens des distributions on en déduit

$$1 - \alpha = P(-b_{\alpha/2} \leqslant \mathcal{N}_{0;1} \leqslant b_{\alpha/2}) \cong P(-b_{\alpha/2} \leqslant \frac{S^2 - \sigma_P^2}{\sqrt{(\hat{\mu}_4 - S^4)/n}} \leqslant b_{\alpha/2})$$

et donc $P(S^2 - b_{\alpha/2} \times \sqrt{(\hat{\mu}_4 - S^4)/n} \leqslant \sigma_P^2 \leqslant S^2 + b_{\alpha/2} \times \sqrt{(\hat{\mu}_4 - S^4)/n)}) \cong 1 - \alpha$.

– *Une réalisation de l'intervalle de confiance*. Après expérimentation on constate que S prend la valeur s et $\hat{\mu}_4$ la valeur μ_4^*.

D'où l'encadrement de σ_P^2 :

$$s^2 - b_{\alpha/2}\sqrt{(\mu_4^* - s^4)/n} \leqslant \sigma_P^2 \leqslant s^2 + b_{\alpha/2}\sqrt{(\mu_4^* - s^4)/n}$$

Application. Afin d'estimer la dispersion des revenus autour de leur valeur moyenne, on se propose de trouver un IC de σ_P de niveau $(1-\alpha) = 0{,}90$. Partant d'un échantillon de 2 500 revenus déclarés x_i, on détermine au préalable la valeur numérique $\mu_4^* = \dfrac{1}{2\,500}\sum_{i=1}^{2\,500}(x_i - \bar{x})^4 = 720$ et la valeur de l'écart-standard $s = 4{,}3$.

– *Variable statistique utilisée*. L'échantillon $X_1, X_2, \ldots, X_n$ étant de grande taille on utilise l'approximation $\theta' = \dfrac{S^2 - \sigma_P^2}{\sqrt{(\hat{\mu}_4 - S^4)/n}} \cong \mathcal{N}_{0;1}$.

– *Intervalle de confiance aléatoire*. On cherche le nombre $b_{\alpha/2}$ tel que $\Phi(b_{\alpha/2}) = 1 - \alpha/2 = 0{,}95$ car alors $P(-b_{\alpha/2} \leqslant \mathcal{N}_{0;1} \leqslant b_{\alpha/2}) = 1 - \alpha = 0{,}90$. Par lecture de la table de $\mathcal{N}(0;1)$ on obtient $b_{\alpha/2} = 1.64$. De $\theta' \cong \mathcal{N}_{0;1}$ on déduit :

$0{,}90 = P(-1.64 \leqslant \mathcal{N}_{0;1} \leqslant 1{,}64) \cong P(-1{,}64 \leqslant \dfrac{S^2 - \sigma_P^2}{\sqrt{(\hat{\mu}_4 - S^4)/n}} \leqslant 1{,}64)$, d'où

$$0{,}90 \cong P(S^2 - 1.64 \times \sqrt{(\hat{\mu}_4 - S^4)/n} \leqslant \sigma_P^2 \leqslant S^2 + 1{,}64 \times \sqrt{(\hat{\mu}_4 - S^4)/n}).$$

– *Une réalisation de l'intervalle*. Avec un niveau de confiance égal à 0,90 on a :

$$4{,}3^2 - 1{,}64\sqrt{(720-4{,}3^4)/2\,500} \leqslant \sigma_P^2 \leqslant 4{,}3^2 + 1{,}64\sqrt{(720-4{,}3^4)/2\,500}$$

et donc $\sqrt{17{,}85} \leqslant \sigma_P \leqslant \sqrt{19{,}12}$.

2 Intervalles de confiance de m et de σ lorsque $X_1, X_2, \ldots, X_n$ est un échantillon, éventuellement de petite taille, *iid* d'une loi normale $\mathcal{N}(m, \sigma^2)$

2.1 Intervalle de confiance de m

$x_1, x_2, \ldots, x_n$ sont les valeurs observées d'un échantillon *iid* d'une loi normale $\mathcal{N}(m, \sigma^2)$ où les valeurs de m et de σ sont inconnues.

– *Variable statistique utilisée*. $X_1, X_2, \ldots, X_n$ étant un échantillon *iid* d'une loi normale $\mathcal{N}(m, \sigma^2)$ on utilise l'une ou l'autre des propriétés suivantes (*cf.* p. 158)

(i) $\dfrac{\bar{X} - m}{\sigma/\sqrt{n}} = \mathcal{N}_{0;1}$ si l'on connaît la valeur de la variance σ^2 (cas théorique),

(ii) $\dfrac{\bar{X} - m}{S/\sqrt{n}} = \boldsymbol{t}_{n-1}$ si la valeur de la variance σ^2 de la distribution est inconnue. ($\boldsymbol{t}_{n-1}$ est une variable de Student à $n-1$ degrés de liberté)

– *Intervalle de confiance aléatoire quand la valeur de σ est inconnue*. Pour trouver un intervalle de confiance de m de niveau $1-\alpha$, on utilise la propriété *(ii)* et on cherche le nombre $b_{\alpha/2}$ tels que $P(-b_{\alpha/2} < \boldsymbol{t}_{n-1} < b_{\alpha/2}) = (1-\alpha)$. Par lecture de table page 378 on obtient $b_{\alpha/2}$ vérifiant $P(\boldsymbol{t}_{n-1} \leqslant b_{\alpha/2}) = 1 - \alpha/2$ (par exemple pour $(n-1) = 8$ et $\alpha = 0{,}1$ on trouve $b_{\alpha/2} = 1{,}86$).

Compte tenu de *(ii)* on a

$$1 - \alpha = P(-b_{\alpha/2} \leqslant \boldsymbol{t}_{n-1} \leqslant b_{\alpha/2}) = P\left(-b_{\alpha/2} \leqslant \frac{\bar{X} - m}{S/\sqrt{n}} \leqslant b_{\alpha/2}\right)$$

et donc $P\left(\bar{X} - b_{\alpha/2} \times \dfrac{S}{\sqrt{n}} \leqslant m \leqslant \bar{X} + b_{\alpha/2} \times \dfrac{S}{\sqrt{n}}\right) = 1 - \alpha$.

– *Une réalisation de l'intervalle*. Après expérimentation, $\bar{X}$ prend la valeur numérique $\bar{x}$ et S prend la valeur s. Avec une probabilité égale à $(1-\alpha)$, on peut affirmer que : $\bar{x} - b_{\alpha/2}\, s/\sqrt{n} \leqslant m \leqslant \bar{x} + b_{\alpha/2}\, s/\sqrt{n}$.

Remarque. Lorsque la valeur de l'écart-type σ est connue, on obtient en général un intervalle plus étroit en utilisant la propriété *(i)*.

© Dunod. La photocopie non autorisée est un délit.

Application. On veut estimer la durée de vie X d'un certain type de composant électronique. Pour cela on dispose d'un échantillon des durées de vie de 10 composants de ce type. Classés par ordre de valeurs croissantes les résultats sont les suivants (l'unité de temps étant l'année) : $1{,}01 - 1{,}04 - 1{,}11 - 1{,}23 - 1{,}24 - 1{,}25 - 1{,}32 - 1{,}34 - 1{,}46 - 2{,}00$ d'où $\bar{x} = 1{,}3$ et $s \cong 0{,}282$. Considérant que X suit sensiblement une loi normale $\mathcal{N}(m,\sigma^2)$ on se propose de trouver un intervalle de confiance (de niveau 0,99) de la valeur moyenne m de la durée de vie de ce type de composant.

À cette fin, on utilise la propriété $\dfrac{\bar{X} - m}{S/\sqrt{10}} = \boldsymbol{t}_9$ puisque $n = 10$. Sur la table de la page 378 on lit « $P(\boldsymbol{t}_9 \leqslant 3{,}25) = 0{,}995$ », donc $P(-3{,}25 \leqslant \boldsymbol{t}_9 \leqslant 3{,}25) = 0{,}99$.

Par suite $P(-3{,}25 \leqslant \dfrac{\bar{X} - m}{S/\sqrt{n}} \leqslant 3{,}25) = 0{,}99$

soit $P(\bar{X} - 3{,}25 \times S/\sqrt{10} \leqslant m \leqslant \bar{X} + 3{,}25 \times S/\sqrt{10}) = 0.99$.

Compte tenu des valeurs $\bar{x} = 1{,}3$ et $s \cong 0{,}282$ respectivement prises par $\bar{X}$ et S, on peut affirmer avec un niveau de confiance de 99 % que « $1{,}3 - (3{,}25) \times (0{,}282)/\sqrt{10} \leqslant m \leqslant 1{,}3 + (3{,}25) \times (0{,}282)/\sqrt{10}$ » soit $1{,}01 \leqslant m \leqslant 1{,}59$.

2.2 Intervalle de confiance de σ

À partir des valeurs observées $x_1, x_2, \ldots, x_n$ on détermine la valeur estimée $s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ de σ^2.

– *Variable statistique utilisée*. $X_1, X_2, \ldots, X_n$ étant un échantillon *iid* d'une loi normale $\mathcal{N}(m,\sigma^2)$, on utilise la propriété : $(n-1)S^2/\sigma^2 = \chi^2_{n-1}$.

– *Intervalle de confiance aléatoire*. Pour trouver un intervalle de confiance de σ^2 de niveau $1 - \alpha$, il faut chercher les nombres $a_{\alpha/2}$ et $b_{\alpha/2}$ tel que $P(\chi^2_{n-1} < a_{\alpha/2}) = \alpha/2$ et $P(\chi^2_{n-1} > b_{\alpha/2}) = \alpha/2$ ce qui implique $P(a_{\alpha/2} \leqslant \chi^2_{n-1} \leqslant b_{\alpha/2}) = 1 - \alpha$. Par lecture de la table de χ^2_{n-1} page 377, on obtient les valeurs $a_{\alpha/2}$ et $b_{\alpha/2}$ en remarquant $P(\chi^2_{n-1} > a_{\alpha/2}) = 1 - \alpha/2$.
Compte tenu de la propriété $(n-1)S^2/\sigma^2 = \chi^2_{n-1}$ on déduit
$1 - \alpha = P(a_{\alpha/2} \leqslant \chi^2_{n-1} \leqslant b_{\alpha/2}) = P(a_{\alpha/2} \leqslant (n-1)S^2/\sigma^2 \leqslant b_{\alpha/2})$ soit

$$1 - \alpha = P\left(\frac{(n-1)S^2}{b_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{a_{\alpha/2}}\right).$$

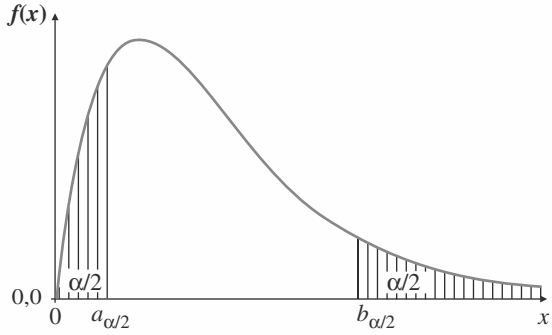


Figure 10.2

– *Une réalisation de l'intervalle*. Après expérimentation on constate que S prend la valeur s. Avec une probabilité égale à $(1-\alpha)$, on peut affirmer que : $\dfrac{(n-1)s^2}{b_{\alpha/2}} \leqslant \sigma^2 \leqslant \dfrac{(n-1)s^2}{a_{\alpha/2}}$. Cet encadrement est l'intervalle de confiance de σ^2.

Exemple

Reprenant l'exemple précédent où la taille de l'échantillon est $n = 10$ et l'écart-type standard $s = 0{,}282$, on peut déterminer un intervalle de confiance à 90 % de σ^2. Sachant que $(n-1)S^2/\sigma^2 = \chi^2_{n-1}$ et que $(1-\alpha) = 0{,}9$ (soit $\alpha/2 = 0{,}05$), il faut chercher les nombres $a_{\alpha/2}$ et $b_{\alpha/2}$ tel que $P(\chi^2_9 < a_{\alpha/2}) = 0{,}05$ et $P(\chi^2_9 > b_{\alpha/2}) = 0{,}05$.
Obtenant par lecture de table $a_{\alpha/2} = 3{,}325$ et $b_{\alpha/2} = 16{,}919$ on en déduit : $P\left(\dfrac{9S^2}{16{,}919} \leqslant \sigma^2 \leqslant \dfrac{9S^2}{3{,}325}\right) = 0.90$ et la réalisation

$$\frac{9 \times 0{,}282^2}{16{,}919} \leqslant \sigma^2 \leqslant \frac{9 \times 0{,}282^2}{3{,}325}, \text{ soit } 0{,}0423 \leqslant \sigma^2 \leqslant 0{,}215\,.$$

3 Intervalle de confiance du paramètre d'une loi de Poisson

$x_1, x_2, \ldots, x_n$ sont les valeurs observées d'un échantillon de grande taille *iid* d'une loi de Poisson $P(\lambda)$ de valeur moyenne λ.

La somme $t = x_1 + x_2 + \ldots + x_n$ est la réalisation d'une *v.a.* T qui suit la loi de Poisson de paramètre $\Lambda = n\lambda$ (*cf.* § 5.3 p. 134).

On détermine un intervalle de confiance $]\Lambda_2; \Lambda_1[$ de Λ de niveau $(1-\alpha)$ en cherchant :

– Λ_1 tel que *poids de* $[0; t] = P(\mathcal{P}_{\Lambda_1} \leqslant t) = \displaystyle\sum_{h=0}^{t} \mathrm{e}^{-\Lambda_1} \frac{\Lambda_1^h}{h!} = \alpha/2$

© Dunod. La photocopie non autorisée est un délit.

– puis Λ_2 tel que *poids de* $[t;\infty[= P(\mathcal{P}_{\Lambda_2} \geqslant t) = \alpha/2$

et donc $P(\mathcal{P}_{\Lambda_2} \leqslant t-1) = \sum_{h=0}^{t-1} \mathrm{e}^{-\Lambda_2} \frac{\Lambda_2^h}{h!} = 1-\alpha/2$.

Par suite, $]\Lambda_2/n;\ \Lambda_1/n[$ est un intervalle de confiance de λ de niveau $(1-\alpha)$.

Méthodologie équivalente. On peut plus simplement déterminer Λ_1 et Λ_2 en utilisant la propriété « $P(\mathcal{P}_\Lambda \leqslant h) = 1 - P(\chi^2_{2h+2} \leqslant 2\Lambda)$ ».

Soit θ_1 et θ_2 définis par $P(\chi^2_{2t} \leqslant \theta_2) = \alpha/2$ et $P(\chi^2_{2t+2} \geqslant \theta_1) = 1-\alpha/2$ où $t = x_1 + \ldots + x_n$.

Avec un niveau de confiance $1-\alpha$ on a l'encadrement : $\theta_2/2n \leqslant \lambda \leqslant \theta_1/2n$.

Exemple

Le nombre aléatoire d'arrivées par 1/2 heure à un guichet de banque suit une loi de Poisson. Prélevant au hasard cinq tranches horaires de 1/2 heure, on observe le nombre d'arrivées correspondantes. Sur cet échantillon de taille $n=5$ on constate que : $x_1=3$, $x_2=1$, $x_3=1$, $x_4=3$ et $x_5=2$. À partir de ces résultats on se propose de trouver un intervalle de confiance du nombre moyen d'arrivées λ avec un niveau de confiance de 0,95.

Ayant constaté que $t = x_1 + \ldots + x_5 = 10$, on cherche θ_1 et θ_2 définis par

$P(\chi^2_{2t} < \theta_2) = P(\chi^2_{20} < \theta_2) = 0{,}025$; $P(\chi^2_{2t+2} \geqslant \theta_1) = P(\chi^2_{22} \geqslant \theta_1) = 0{,}025$.

Page 377, on lit $\theta_2 = 9{,}59$ et $\theta_1 = 36{,}79$. Avec un niveau de confiance de 95 % on a l'encadrement : $9{,}59/(2\times 5) \leqslant \lambda \leqslant 36{,}79/(2\times 5)$ soit $0{,}959 \leqslant \lambda \leqslant 3{,}679$.

Remarque. Lorsque la taille de l'échantillon n est suffisamment grande, on utilise l'approximation la plus simple $2(\sqrt{T} - \sqrt{n\lambda}) \cong \mathcal{N}_{0;1}$. Se fixant un niveau de confiance $1-\alpha$ on cherche $b_{\alpha/2}$ tel que $\Phi(b_{\alpha/2}) = 1-\alpha/2$. On obtient alors l'intervalle de confiance : $(t^{1/2} - b_{\alpha/2}/2)^2 \leqslant n\lambda \leqslant (t^{1/2} + b_{\alpha/2}/2)^2$ où t désigne la valeur prise par T. La condition $(t^{1/2} - b_{\alpha/2}/2)^2 > 10$ doit être satisfaite pour utiliser cette approximation.

4 Intervalle de confiance de la valeur moyenne d'une loi exponentielle

Soit un échantillon $X_1, X_2, \ldots, X_n$ *iid* d'une loi exponentielle non décalée $\mathrm{Exp}(a)$ dont la valeur du paramètre a est inconnue.

La somme $T_n = X_1 + X_2 + \ldots + X_n$. satisfait à la propriété $2aT_n = \chi^2_{2n}$ (*cf.*§ 4.2 p. 141) ou de façon équivalente puisque la moyenne de la loi $m = 1/a$ on a

$$2T_n/m = \chi^2_{2n} \text{ où } n \text{ est la taille de l'échantillon.}$$

Pour trouver un intervalle de confiance de niveau $1-\alpha$ de la moyenne m de la loi on cherche les nombres $a_{\alpha/2}$ et $b_{\alpha/2}$ tels que :

$$P(\chi^2_{2n} \leqslant a_{\alpha/2}) = \alpha/2 \qquad P(\chi^2_{2n} \geqslant b_{\alpha/2}) = \alpha/2\,.$$

Donc $1-\alpha = P(a_{\alpha/2} \leqslant \chi^2_{2n} \leqslant b_{\alpha/2}) = P(a_{\alpha/2} \leqslant 2T_n/m \leqslant b_{\alpha/2})$
$= P(2T_n/b_{\alpha/2} \leqslant m \leqslant 2T_n/a_{\alpha/2})\,.$

Notons t la réalisation de T_n : $t = x_1 + x_2 + \ldots + x_n$.

Avec un niveau de confiance $1-\alpha$ on a l'encadrement : $2t/b_{\alpha/2} \leqslant m \leqslant 2t/a_{\alpha/2}$.

5 Intervalle de confiance d'une proportion p

5.1 La distribution sur une population binomiale

Soit une population $\boldsymbol{P} = \{e_1, e_2, \ldots, e_N\}$ dans laquelle N_1 éléments parmi les N possèdent le caractère considéré $\mathcal{A}$. Le quotient $p = N_1/N$ représente la proportion d'éléments de $\boldsymbol{P}$ qui possèdent le caractère $\mathcal{A}$. Autrement dit, on peut associer à chaque élément e_i sa mesure x_i qui vaut 1 si e_i possède le caractère $\mathcal{A}$ et 0 sinon.

Valeurs prises par x	1	0	**Total**
Effectifs N_i	N_1	N_2	N
Fréquences	$p = N_1/N$	$1-p$	1

La moyenne m_p sur la population est égale à la proportion p et la variance sur la population ν_P est égale à $p(1-p)$.

5.2 Constitution d'un échantillon

On extrait de $\boldsymbol{P}$ un échantillon non exhaustif de taille n. Autrement dit on prélève au hasard et avec remise n éléments $e'_1, e'_2, \ldots, e'_n$ dans une population binomiale $\boldsymbol{P}$. À chaque élément e'_i tiré de $\boldsymbol{P}$ on peut associer la *v.a.* X_i de Bernoulli qui prend la valeur 1 si e'_i possède le caractère étudié et 0 dans le cas contraire. X_i suit la loi de Bernoulli de paramètre $p = N_1/N$:

$$\Delta_X = \{0,1\},\ P(X=1) = P(e \text{ possède } \mathcal{A}) = N_1/N = p,\ P(X=0) = 1-p\,.$$

Le nombre aléatoire $K = (X_1 + X_2 + \ldots + X_n)$ d'éléments de l'échantillon de taille n qui possèdent le caractère étudié suit donc la loi binomiale de paramètres $(n; p)$. La proportion aléatoire $F = K/n$ d'éléments qui, sur l'échantillon, possèdent le caractère $\mathcal{A}$ est un estimateur de p. (En effet dans le cas particulier de population binomiale, la moyenne sur la population $m_P = p$ et la moyenne aléatoire de l'échantillon $\bar{X} = F$).

Sur l'échantillon observant qu'une proportion f possède le caractère considéré (autrement dit f étant la valeur prise par F) on souhaite fournir un intervalle de confiance de la proportion p sur la population.

© Dunod. La photocopie non autorisée est un délit.

5.3 Intervalle de confiance de p à partir d'un grand échantillon

– *Variable statistique utilisée*. L'échantillon étant de grande taille, on peut utiliser l'approximation $\dfrac{F-p}{\sqrt{F(1-F)/n}} \cong \mathcal{N}_{0;1}$.

– *Intervalle de confiance aléatoire*. Pour trouver un intervalle de confiance de p avec un niveau de confiance de $(1-\alpha)$ on cherche le nombre $b_{\alpha/2}$ défini par $\Phi(b_{\alpha/2}) = 1-\alpha/2$. On a alors : $1-\alpha = P(-b_{\alpha/2} \leqslant \mathcal{N}_{0;1} \leqslant b_{\alpha/2})$ $\cong P(-b_{\alpha/2} \leqslant \dfrac{F-p}{\sqrt{F(1-F)/n}} \leqslant b_{\alpha/2})$, ou de façon équivalente :

$$P(F - b_{\alpha/2} \times \sqrt{F(1-F)/n} \leqslant p \leqslant F + b_{\alpha/2} \times \sqrt{F(1-F)/n}) \cong 1-\alpha\,.$$

– *Une réalisation de l'intervalle*. Après expérimentation, on constate que F prend la valeur numérique f qui est la proportion dans l'échantillon. Avec un niveau de confiance $1-\alpha$, on obtient l'encadrement :

$$f - b_{\alpha/2} \times \sqrt{f(1-f)/n} \leqslant p \leqslant f + b_{\alpha/2} \times \sqrt{f(1-f)/n}\,.$$

Exemple

On envisage de changer de fournisseurs de pièces en raison d'un pourcentage élevé de pièces défectueuses. Sur un lot de 100 pièces tirées au hasard on observe qu'il y a huit pièces défectueuses. À partir de ces données on souhaite déterminer un intervalle de confiance de la proportion p de pièces défectueuses parmi celles actuellement utilisées, et ce avec un niveau de confiance de 95 % ($\alpha = 0{,}05$). Pour cela on cherche $b_{\alpha/2}$ défini par $\Phi(b_{\alpha/2}) = 1-\alpha/2 = 0{,}975$, puis ayant lu $b_{\alpha/2} = 1{,}96$, on utilise l'encadrement ci-dessus où $f = 8/100 = 0{,}08$ est la proportion de pièces défectueuses observées dans l'échantillon :

$0{,}08 - 1{,}96\sqrt{0{,}08 \times (1-0{,}08)/100} \leqslant p \leqslant 0{,}08 + 1{,}96\sqrt{0{,}08 \times (1-0{,}08)/100}$,

soit $0{,}0268 \leqslant p \leqslant 0{,}1331$.

Remarques.

1/ L'approximation $\dfrac{F-p}{\sqrt{p(1-p)/n}} \cong \mathcal{N}_{0;1}$ est de meilleure qualité que celle utilisée mais nécessite des calculs plus complexes, soit les racines p_1 et p_2 de l'équation du second degré $p^2(1 + b^2_{\alpha/2}/n) - p(2f + b^2_{\alpha/2}/n) + f^2 = 0$. Avec un niveau de confiance sensiblement égal à $1-\alpha$ on a alors $p_1 \leqslant p \leqslant p_2$ (voir exercice 3).

2/ Lorsque l'échantillon de taille n est extrait de $P = \{e_1, e_2, \ldots, e_N\}$ au hasard de *façon exhaustive* on utilise l'approximation $\dfrac{F-p}{\sqrt{\frac{N-n}{N-1}p(1-p)/n}} \cong \mathcal{N}_{0,1}$.

5.4 Intervalle de confiance de p dans le cas de petits échantillons

Soit un échantillon de taille n quelconque, on constate après analyse, que ν des n éléments de l'échantillon possèdent le caractère $\mathcal{A}$ (autrement dit le nombre aléatoire K d'éléments de l'échantillon de taille n qui possèdent le caractère étudié prend pour valeur ν). Au vu de ce résultat, on détermine un intervalle de confiance $]p_2;\ p_1[$ de p de niveau $(1-\alpha)$ en cherchant :

– p_1 tel que *poids de* $[0;\nu] = P(0 \leqslant B_{n;p_1} \leqslant \nu) = \displaystyle\sum_{i=0}^{\nu} C_n^i p_1^i (1-p_1)^{n-i} = \alpha/2$

– p_2 tel que *poids de* $[\nu;n] = P(B_{n,p_2} \geqslant \nu) = \displaystyle\sum_{i=\nu}^{n} C_n^i p_2^i (1-p_2)^{n-i} = \alpha/2$

et donc $P(0 \leqslant B_{n;p_2} \leqslant \nu-1) = \displaystyle\sum_{i=0}^{\nu-1} C_n^i p_2^i (1-p_2)^{n-i} = 1-\alpha/2\,.$

Méthodologie équivalente. On peut déterminer p_1 et p_2 en utilisant

$$\text{« } P(B_{n;p} \leqslant h) = P(\boldsymbol{F}_{2n-2h}^{2h+2} > \frac{p}{1-p} \times \frac{n-h}{h+1}) \text{ » :}$$

– $p_1 = \dfrac{(\nu+1)\xi_1}{(n-\nu)+(\nu+1)\xi_1}$ où ξ_1 est défini par $P(\boldsymbol{F}_{2n-2\nu}^{2\nu+2} \geqslant \xi_1) = \alpha/2,$

– $p_2 = \dfrac{\nu\xi_2}{n-\nu+1+\nu\xi_2}$ où ξ_2 est défini par $P(\boldsymbol{F}_{2n-2\nu+2}^{2\nu} < \xi_2) = \alpha/2$

La procédure proposée est évidemment valable quelque soit la taille de l'échantillon.

Application. On dispose d'un échantillon de 20 contribuables ($n = 20$) auxquels on demande s'ils connaissent ou non le montant maximal de réduction d'impôt auxquels ils ont droit pour l'emploi d'une personne à domicile. Le nombre aléatoire X d'individus qui connaissent ce plafond suit donc une loi $B(20;\ p)$. Ayant constaté que seules deux personnes de l'échantillon connaissent la réponse ($\nu = 2$) on souhaite déterminer un IC de p de niveau 0,90.

On cherche :

– ξ_1 tel que $P(\boldsymbol{F}_{2n-2\nu}^{2\nu+2} > \xi_1) = \alpha/2$ avec $\alpha = 0{,}1$, $n = 20$ et $\nu = 2$. Par lecture de table p. 379 on sait que $P(F_{36}^{6} > 2{,}36) = 0{,}05$ donc $\xi_1 = 2{,}36$ et $p_1 = 0{,}2823$.

– ξ_2 tel que $P(\boldsymbol{F}_{2n-2\nu+2}^{2\nu} < \xi_2) = \alpha/2$. Par lecture de table, on sait que $P(F_4^{38} > 5{,}72) = 0{,}05$ donc $\xi_2 = (5{,}72)^{-1} = 0{,}1748$ et $p_2 = 0{,}0180$.

Avec un niveau de confiance de 0,90 on a $0{,}0180 \leqslant p \leqslant 0{,}2823$.

© Dunod. La photocopie non autorisée est un délit.

Remarque. Si on dispose d'un échantillon de h valeurs $x_1, x_2, \ldots, x_h$ prises par une *v.a.* X qui suit la loi $\mathcal{B}(n; p)$, pour déterminer un IC de p on doit considérer la somme $t = \Sigma_i x_i$ qui est la valeur prise par une *v.a.* T qui suit la loi $\mathcal{B}(nh; p)$.

6 Intervalle de confiance d'un p-quantile ξ_p d'une loi

6.1 Les variables d'ordre

Soit un échantillon $X_1, X_2, \ldots, X_n$ *iid* d'une loi continue. À toute réalisation $x_1, x_2, \ldots, x_n$ de cet échantillon faisons correspondre leur rangement par ordre croissant : $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ avec $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. Autrement dit, $x_{(1)}$ est le plus petit des n nombres $x_1, x_2, \ldots, x_n$, $x_{(2)}$ est le plus petit des $(n-1)$ nombres restants etc., $x(n)$ étant évidemment le plus grand.

La *v.a.* $X_{(i)}$ qui prend la valeur $x_{(i)}$ est appelée i-ème statistique d'ordre. En particulier $X_{(1)}$ et $X_{(n)}$ sont respectivement la plus petite et la plus grande variable d'ordre : $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$.

Théorème. Désignant par ξ_p le p-quantile de la loi suivie par $X_1, X_2, \ldots, X_n$, on a pour tout i et j appartenant à $\{1, 2, \ldots, n\}$:

$$P(X_{(i)} \leqslant \xi_p) = P(i \leqslant \mathcal{B}_{n;p}), P(\xi_p \leqslant X_{(j)}) = P(\mathcal{B}_{n;p} \leqslant j-1) \text{ et conséquemment}$$

$$P(X_{(i)} \leqslant \xi_p \leqslant X_{(j)}) = P(i \leqslant \mathcal{B}_{n;p} \leqslant j-1)$$

où $\mathcal{B}_{n;p}$ est une variable binomiale de paramètres n, p.

6.2 Détermination d'un intervalle de confiance d'un p-quantile ξ_p

Pour obtenir un intervalle de confiance centré de niveau $1-\alpha$ on cherche :

– le plus grand entier i_0 vérifiant $P(\mathcal{B}_{n;p} \leqslant i_0 - 1) \leqslant \alpha/2$

– le plus petit entier j_0 tel que $P(\mathcal{B}_{n;p} \geqslant j_0) \geqslant \alpha/2$

et l'on en déduit : $P(X_{(i_0)} \leqslant \xi_p \leqslant X_{(j_0)}) = P(i_0 \leqslant \mathcal{B}_{n;p} \leqslant j_0 - 1) \leqslant 1 - \alpha$.

Avec un niveau de confiance $1-\alpha$ on a l'encadrement $x_{(i_0)} \leqslant \xi_p \leqslant x_{(j_0)}$.

Cas particulier de la médiane $m_e = \xi_{0,5}$. Dans ce cas particulier $j_0 = n - i_0 + 1$ puisque la loi de $\mathcal{B}_{n;0,5}$ est symétrique.

Exemple

À partir d'un échantillon de 15 entreprises d'un secteur S on se propose de déterminer un intervalle de confiance de niveau 0,95 pour la valeur médiane $\xi_{0,5}$ du taux de rentabilité économique des firmes de ce secteur en %. Les 15 valeurs numériques sont 0,69, 1,84,

7,83, 1,51, 0,39, – 4,1, 0,33, 1,15, – 1,2, 6,67, 3,61, 3,58, – 1,9, 0,48, 1,92. La classification par rang croissant donne $x_{(1)} = -4{,}1, x_{(2)} = -1{,}91, \ldots x_{(14)} = 6{,}67, x_{(15)} = 7{,}83$

Rang i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$x_{(i)}$	– 4,1	– 1,9	–1,2	0,33	0,39	0,48	0,69	1,15	1,51	1,84	1,92	3,58	3,61	6,67	7,83

Le quantile considéré étant la médiane $\xi_{0,5}$, on cherche le plus grand entier i_0 tel que $P(\mathcal{B}_{15;0,5} \leqslant i_0 - 1) \leqslant 0{,}025$. Sur la table p. 375 on lit $P(\mathcal{B}_{15;0,5} \leqslant 3) = 0.0176$ et $P(\mathcal{B}_{15;0,5} \leqslant 4) = 0{,}0592$. Donc $i_0 - 1 = 3$ et $j_0 = n - i_0 + 1 = 12$. Par suite, $P(X_{(4)} \leqslant \xi_p \leqslant X_{(12)}) = P(4 \leqslant \mathcal{B}_{15;0,5} < 12 - 1) = 1 - 2 \times 0{,}0176 \cong 0{,}965$.

Avec un niveau de confiance de 96.6 % on a $x_{(4)} \leqslant \xi_{0,5} \leqslant x_{(12)}$, c'est-à-dire $0{,}33 \leqslant \xi_{0,5} \leqslant 3{,}58$.

Remarque. Cette procédure s'applique également aux p-quantiles des lois discrètes usuelles après classement des n valeurs $x_1, x_2, \ldots, x_n$ par ordre de valeurs croissantes, les ex-aequo étant également classés.

Section **5**

ESTIMATION AVEC SPSS ET EXCEL

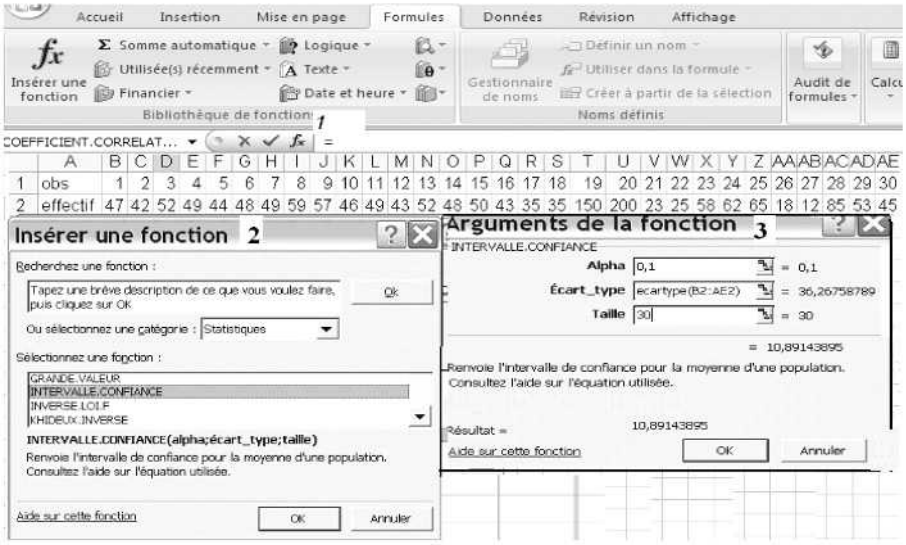
1 Traitements avec EXCEL

Sous EXCEL, on utilise la propriété $\dfrac{\bar{X} - m_P}{\sigma_P / \sqrt{n}} \cong \mathcal{N}_{0;1}$ qui est à appliquer lorsque σ_P est connu ou $\dfrac{\bar{X} - m_P}{S / \sqrt{n}}$ si σ_P inconnu.

Exemple

La direction des études d'une banque souhaite estimer, par un intervalle de confiance à 90%, la taille moyenne des entreprises d'un secteur donné. Prélevant au hasard et avec remise 30 entreprises, elle relève leur effectif afin d'en inférer sur la population.

© Dunod. La photocopie non autorisée est un délit.



Procédure.

1. On clique sur (Insertion) puis (Fonction) puis (OK)

2. On sélectionne le menu (Statistiques) on retient (INTERVALLECONFIANCE).

3. On sélectionne le risque d'erreur α (Alpha = 0,1) puis l'on pose (écart-type = ECARTYPE(B2 :AE2)) afin de calculer s l'écart-standard de l'échantillon dont les données sont comprises entre B2 et AE2 et enfin on précise la taille de l'échantillon (taille = 30)

Le résultat figure immédiatement dans Valeur =10,8914 , autrement dit $-10,8914 + \bar{x} \leqslant m \leqslant 10,8914 + \bar{x}$. En se mettant dans une autre cellule et en tapant

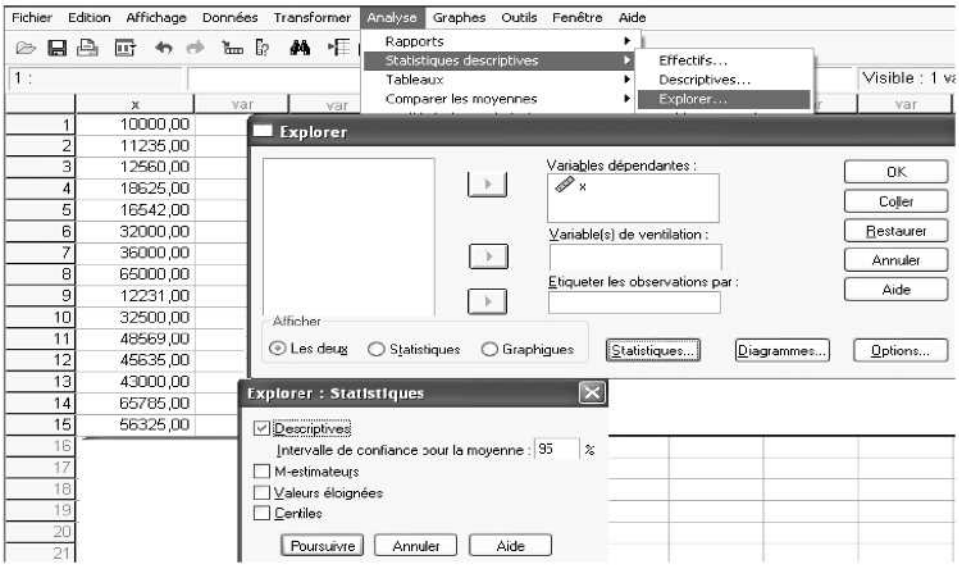
= MOYENNE(B2 :AE2) on obtient $\bar{x} = 54,8$ et on déduit $43,9086 \leqslant m \leqslant 65,6914$.

2 Traitements avec SPSS

SPSS calcule les intervalles de confiance dans le cas où X suit une loi normale $\mathcal{N}(m_P, \sigma_P^2)$ avec σ_P^2 inconnu et utilise pour cela la propriété $\dfrac{\bar{X} - m}{S/\sqrt{n}} = \boldsymbol{t}_{n-1}$.

Exemple

Dans le cadre d'un audit comptable, on souhaite estimer la moyenne des soldes X des balances fournisseurs. Prélevant un échantillon de 15 soldes $x_1, x_2, \ldots, x_{15}$ et supposant que X suit une loi normale $N(m_P, \sigma_P^2)$, on souhaite obtenir un intervalle de confiance de la valeur moyenne des soldes avec un niveau de confiance de 95 %.



Procédure.

1. Cliquer sur (Analyse), (Statistiques descriptives) et (Explorer).

2. Dans le menu **Explorer** on saisit la variable (*x*) que l'on envoie dans (Variables dépendantes)

3. Cliquant sur (Statistiques), on obtient le menu **Explorer statistiques** dans lequel on fixe le niveau de confiance (95)% puis on clique sur (Poursuivre)

Résultats obtenus

			Statistique	Erreur standard
	Moyenne		33733,1	5106.3
	Intervalle de confiance à 95 % pour la moyenne	Borne inférieure	22781,1	
		Borne supérieure	44685,1	

Interprétation : $\bar{x} = 33733{,}1$, $s/\sqrt{n} = 5106{,}3$. Il y a 95 chances sur 100 pour que la moyenne m sur la population soit comprise dans cette fourchette : $22781{,}1 \leqslant m \leqslant 44685{,}1$.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 351)

Exercice 1

Une assurance automobile avait sélectionné en fin d'année 900 véhicules d'une cylindrée supérieure à 1200 cm^3 afin de connaître la distance moyenne m_P parcourue au cours de l'année par ce type de véhicules. L'unité de distance retenue est 10^3 km. L'analyse des résultats a donné une valeur moyenne $\bar{d} = 15$, un écart- standard $s = 5$ et un moment centré d'ordre 4 $\mu_4^* = 800$.

1. Définir avec précision la population statistique $\boldsymbol{P}$.

2. Trouver un intervalle de confiance de niveau 95 % de m_p.

3. Trouver un intervalle de confiance de niveau 90 % de l'écart-type σ_p de la distribution sur P.

Exercice 2

Une machine fabrique en série des rondelles. Leur diamètre D suit sensiblement une loi normale. On prélève au hasard un échantillon de 10 rondelles. Les mesures des diamètres, en millimètres sont les suivantes : 20,1 ; 19,9 ; 20,0 ; 19,8 ; 19,7 ; 20,2 ; 20,1 ; 20,3 ; 20,0 ; 19,9.

a) Construire un intervalle qui contienne, avec une probabilité de 95 %, la valeur moyenne m_P de la loi que suit D.

b) Même question pour l'écart-type $\sigma_P = \sigma(D)$.

Exercice 3

La société Landchaus spécialisée dans la vente de chaussures a dix magasins dans la région Ile-de-France. Ce grand distributeur souhaitant savoir si son enseigne est suffisamment connue du public dans la région Ile de France décide de faire procéder à une enquête auprès des consommateurs.

Connaissant de façon aléatoire l'âge de 10 clientes : 16, 32, 28, 45, 18, 60, 52, 48, 45, 35 construire un intervalle de confiance de l'âge médian du client avec un niveau de confiance de 85%.

Exercice 4

Une grande entreprise mène un intense politique de formation du personnel. Au cours de l'année, 200 employés ont suivi une même formation FT. Le Directeur des ressources humaines a demandé à 50 de ces employés choisis au hasard sans remise (échantillon exhaustif) s'ils étaient satisfaits de leur formation et leur ancienneté dans l'entreprise.

1. Sur les cinquante interrogés, 40 se sont déclarés satisfaits. Construire un intervalle de confiance à 90 % du pourcentage d'employés satisfaits de la formation FT.

2. L'ancienneté dans l'entreprise des 50 employés a une moyenne de 68 mois et un écart-standard de répartition s = 24 mois. Trouver un intervalle de confiance à 95 % de l'ancienneté moyenne des employés ayant suivi la formation.

Exercice 5

On s'intéresse à la mesure numérique X d'un caractère sur une population statistique P de taille n. Cette population P étant partagée en deux strates P_1 et P_2 d'effectifs respectifs ν_1 et ν_2 ($\nu_1 + \nu_2 = \nu$), $X^{(1)}$ et $X^{(2)}$ désignent respectivement les *v.a.* associées à la mesure du caractère observé dans les sous-populations P_1 et P_2. La moyenne et la variance de la distribution de X sur P sont respectivement notées m et σ^2 : $m = E(X)$, $\sigma^2 = V(X)$. De même $m_1 = E(X^{(1)})$, $\sigma_1^2 = V(X^{(1)})$ et $m_2 = E(X^{(2)})$, $\sigma_2^2 = V(X^{(2)})$. Effectuant un sondage aléatoire simple avec remise, de taille n_1 dans P_1 et de taille n_2 dans P_2 on obtient les échantillons $(X_1^{(1)},\ldots,X_{n_1}^{(1)})$ et $(X_1^{(2)},\ldots,X_{n_2}^{(2)})$. Soit

$$\overline{X}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)} \quad \text{et} \quad \overline{X}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i^{(2)}$$

1. Montrer que $Z = \dfrac{\nu_1}{\nu}\overline{X}^{(1)} + \dfrac{\nu_2}{\nu}\overline{X}^{(2)}$ est un estimateur convergent sans biais de m.

2. Pour des raisons de coût, la taille n de l'échantillonnage pris dans sa globalité étant fixée ($n_1 + n_2 = n$ entier fixé), montrer que l'estimateur Z de m a une variance minimale lorsque

$$n_1 = \frac{\nu_1\sigma_1}{\nu_1\sigma_1 + \nu_2\sigma_2}n, \quad n_2 = \frac{\nu_2\sigma_2}{\nu_1\sigma_1 + \nu_2\sigma_2}n = n - n_1 \ .$$

Exercice 6

La veille d'une consultation électorale un sondage a été réalisé auprès d'électeurs afin d'avoir une estimation du pourcentage de votants pour la liste EXT considérée comme extrémiste. Pour éviter les réponses non sincères le questionnaire a été présenté de la façon suivante : « répondez oui si vous êtes nés en Janvier ou si vous souhaitez voter pour la liste EXT » ; « si aucune des conditions précédentes n'est réalisée, répondez non ». Sur un échantillon de 2500 électeurs ayant décidés de voter, 375 ont répondus « oui ». L'échantillon étant considéré comme tiré au hasard, trouver un intervalle de confiance de niveau 0.95 de la proportion p^* d'électeurs qui dans la population des votants satisfont à la condition C « être né en Janvier ou souhaiter voter pour la liste EXT » et en déduire un intervalle de confiance de même niveau pour la proportion p de votants pour la liste EXT.

© Dunod. La photocopie non autorisée est un délit.

11 TESTS D'HYPOTHÈSES PARAMÉTRIQUES

Réaliser un test d'hypothèse paramétrique consiste à émettre une hypothèse concernant la distribution de la mesure d'un caractère sur une population et à décider d'accepter ou de refuser cette hypothèse à partir d'un ou de plusieurs échantillons. On teste systématiquement une hypothèse donnée appelée hypothèse nulle H_0 contre une hypothèse alternative. Deux choix sont possibles : prendre la décision D_0 d'accepter H_0 ou bien la décision D_1 de rejeter H_0.

Lorsque les hypothèses portent sur la valeur d'un ou de plusieurs paramètres (moyenne, écart-type ...) d'une même population, on parle de test d'hypothèse paramétrique. Si l'hypothèse concerne une étude comparative des valeurs d'un même paramètre dans des populations statistiques distinctes, ce test paramétrique est appelé test de comparaison. Dans le cas où le test utilisé ne fait pas intervenir la nature de la distribution dont est ou sont issus le ou les échantillons, on parle de méthodes non paramétriques (tests d'ajustement, d'indépendance...).

Dans ce chapitre, sont présentés les tests paramétriques concernant les valeurs de paramètres de distribution.

Section 1 ■ Méthodologie des tests
Section 2 ■ Tests relatifs à la valeur de la moyenne d'une distribution
Section 3 ■ Tests relatifs à la valeur de l'écart-type d'une distribution
Section 4 ■ Tests relatifs à la valeur d'une proportion
Section 5 ■ Tests de la moyenne sous SPSS et EXCEL
Section 6 ■ Test de symétrie d'une distribution

Section 1

MÉTHODOLOGIE DES TESTS

1 Les différents types d'hypothèses

Dans le cas où le test d'hypothèse porte sur la valeur d'une caractéristique θ d'une population statistique P (souvent la moyenne m ou l'écart-type σ) ou sur la valeur du paramètre θ d'une loi de probabilité $\mathcal{L}(\theta)$, on formule des hypothèses concernant la valeur de ce paramètre que l'on essaie de valider au vu des n valeurs numériques $x_1, x_2, \ldots, x_n$ associées à la réalisation d'un échantillon. Pour ce type de test, il est d'usage de distinguer les tests d'hypothèse simple des tests d'hypothèses multiples de type bilatéral ou unilatéral.

Repères

Le test est dit d'hypothèse simple lorsqu'il s'agit de choisir entre 2 valeurs numériques θ_0 et θ_1 pour le paramètre θ étudié. Ainsi on peut tester l'hypothèse H_0 « $\theta = \theta_0$ » contre l'hypothèse alternative « $\theta = \theta_1$ ».

Le test est dit d'hypothèse multiple lorsque l'hypothèse alternative correspond à un ensemble de valeurs possibles pour le paramètre étudié. Ainsi on peut tester H_0 « $\theta = \theta_0$ » contre $\bar{H}_0$ « $\theta \neq \theta_0$ », H_0 « $\theta = \theta_0$ » contre H_1 « $\theta > \theta_0$ »; H_0 contre H_1' « $\theta < \theta_0$ ». Si l'hypothèse alternative est de type $\theta < \theta_0$ ou bien $\theta > \theta_0$, il s'agit d'*un test unilatéra*l. Si elle est de type $\theta \neq \theta_0$, alors il s'agit d'un *test bilatéral*.

Exemple

Un réviseur doit contrôler 50 000 opérations comptables. La norme professionnelle est qu'il y ait au plus 1% de fautes d'enregistrement sur ces opérations comptables pour que le contrôle interne des comptes soit considéré comme bon. Ne pouvant vérifier toutes les écritures, il constitue un échantillon pour savoir s'il doit prendre la décision D_0 d'admettre l'hypothèse H_0 « que la proportion p de fautes d'enregistrements a une valeur inférieure ou égale à la limite « $p \leqslant 1\ \%$ » ou s'il doit prendre la décision D_1 de retenir l'hypothèse alternative H_1 « $p > 1\ \%$ ».

Prenant sa décision à partir de résultats sur un échantillon, il peut commettre des erreurs dues au caractère partiel de l'enquête menée. Aussi travaillant sur un échantillon de 100 opérations d'enregistrement tirées au hasard, il se fixe pour simplifier, comme règle que s'il y a moins de 3 erreurs il prend la décision D_0 d'admettre l'hypothèse H_0 « $p \leqslant 1\ \%$ et conclut que le contrôle interne est bon ». Dans le cas contraire (3 erreurs ou +) il prend la décision D_1 de retenir l'hypothèse alternative H_1 .

© Dunod. La photocopie non autorisée est un délit.

2 Méthode de Neyman

Prenant sa décision au vu de la réalisation $x_1, x_2, \ldots, x_n$ d'un échantillon *iid* issu d'une population statistique $\boldsymbol{P}$, on peut commettre deux types d'erreurs du fait du caractère partiel de l'enquête.

Validité de l'hypothèse sur la population / **Décision à partir des données de l'échantillon**	H_0 **vraie**	H_1 **vraie**
D_0 **accepter** H_0	décision correcte	erreur de 2ème espèce β
D_1 **accepter** H_1	erreur 1ère espèce α	décision correcte

– La probabilité de prendre la décision D_1 alors que H_0 est vraie est appelée *risque d'erreur de première espèce*, elle est notée α : $\alpha = P(D_1/H_0 \text{ vraie})$[1]

– La probabilité de prendre la décision D_0 alors que H_1 est vraie est appelée *risque d'erreur de seconde espèce*. Elle est notée β : $\beta = P(D_0/H_1 \text{ vraie})$

– La puissance d'un test est mesurée par la quantité $\gamma = 1 - \beta = P(D_1/H_1 \text{ vraie})$

Tous les tests présentés seront basés sur le principe de Neyman-Pearson. On fixe *a priori* une valeur à α ou du moins une limite supérieure à ce risque d'erreur α (par exemple 5 % ou 10 %) et on en déduit, au vu des résultats sur l'échantillon, la règle de décision « on décide ou non de rejeter H_0 ». Dans ce type de test, l'hypothèse H_0 et son hypothèse alternative ne jouent pas un rôle symétrique. H_0 est l'hypothèse de base pour laquelle on limite *a priori* le risque de rejet à tort. Dans le cadre de notre exemple introductif, on prend la décision D_1 si $F \geqslant c$ avec $c = 0{,}03$. Aussi le risque de première espèce est-il $\alpha = P(F \geqslant c/H_0 \text{ vraie})$ noté $P_0(F \geqslant c)$ [lire probabilité pour que la proportion F dans l'échantillon soit $\geqslant c$ si l'hypothèse H_0 est vraie]. On prend la décision D_0 si la valeur numérique f prise par F est telle que $f < c$.

Le risque de seconde espèce $\beta = P(D_0/H_1 \text{ vraie}) = P(F < c/H_1 \text{ vraie})$ noté $P_1(F < c)$ [lire probabilité pour que la proportion F dans l'échantillon soit $< c$ si l'hypothèse H_1 est vraie].

1. Lire : α est la probabilité de prendre la décision D_1 sachant qur H_0 est vraie.

REPÈRES : Règle générale de décision.

À partir de l'hypothèse de base H_0 que l'on souhaite tester, on se fixe un niveau de signification α (c'est-à-dire une faible probabilité, par exemple 5 % ou 1 %). Compte tenu de l'hypothèse alternative retenue on détermine alors un domaine de réels δ_α tel que, si l'*hypothèse H_0 est vraie*, la probabilité pour que la variable statistique utilisée T_0 prenne une valeur appartenant à δ_α soit égale à α une valeur très faible. Si après expérimentation la valeur prise par T_0 appartient à δ_α , étant en présence d'un événement rare, on décide de rejeter H_0. En général δ_α est un intervalle ou une réunion de deux intervalles.

Parmi tous les domaines δ_α dont le poids, sous H_0, est égal à α : « $P(T_0 \in \delta_\alpha$ si H_0 est vraie) = α », on doit choisir celui pour lequel, si H_0 est faux (et donc H_1 vraie), la probabilité d'appartenance de T_0 à δ_α soit la plus grande possible : « $P(T_0 \in \delta_\alpha$ si H_1 est vraie) maximum ».

Notations. $P(T_0 \in \delta_\alpha$ si H_0 est vraie) sera noté $P_0(T_0 \in \delta_\alpha)$; $P(T_0 \in \delta_\alpha$ si H_1 est vraie) sera noté $P_1(T_0 \in \delta_\alpha)$.

3 Procédure pour la réalisation d'un test paramétrique

S'intéressant à la valeur d'un paramètre θ (*moyenne m, écart-type σ ou proportion p*...) d'une loi $\mathcal{L}$ dont la nature de la distribution est connue ou inconnue, on peut, à partir la réalisation $x_1, x_2, \ldots, x_n$ d'un échantillon *iid* $X_1, X_2, \ldots, X_n$ de la loi $\mathcal{L}$, émettre une hypothèse H_0 sur la valeur ce paramètre θ contre une hypothèse alternative qui sera notée selon le cas H_1, H_1' ou $\bar{H}_0$.

REPÈRES

Suivant la forme prise par l'hypothèse alternative, on regroupe les différents tests proposés en trois catégories :

[1] Test bilatéral	H_0 « $\theta = \theta_0$ » contre $\overline{H}_0$ « $\theta \neq \theta_0$ »		
[2] Test unilatéral avec rejet à droite de H_0	**[2a]** H_0 « $\theta = \theta_0$ » contre H_1 « $\theta > \theta_0$ »	**[2b]** H_0 « $\theta \leqslant \theta_0$ » contre H_1 « $\theta > \theta_0$ »	**[2c]** H_0 « $\theta = \theta_0$ » contre H_1 « $\theta = \theta_1$ avec $\theta_1 > \theta_0$ »
[3] Test unilatéral avec rejet à gauche de H_0	**[3a]** H_0 « $\theta = \theta_0$ » contre H_1' « $\theta < \theta_0$ »	**[3b]** H_0 « $\theta \geqslant \theta_0$ » contre H_1' « $\theta < \theta_0$ »	**[3c]** H_0 « $\theta = \theta_0$ » contre H_1' « $\theta = \theta_1$ avec $\theta_1 < \theta_0$ »

© Dunod. La photocopie non autorisée est un délit.

– **Dans une première étape** il faut déterminer la statistique utilisée.
Les informations apportées par l'échantillon $X_1, \cdots, X_n$ sont résumées dans une statistique $T(X_1, \cdots, X_n)$ (statistique correspondant à celles présentées pour les intervalles de confiance p 157-159) **qui, sous l'hypothèse nulle H_0 a une distribution connue et est notée** $T_0(X_1, \cdots, X_n)$ ou plus simplement T_0.

– **Dans une deuxième étape** il faut déterminer la règle de décision obtenue à partir du risque d'erreur de 1ère espèce α que l'on s'est fixé (souvent α = 1 % ou 5 %) puis conclure au vu des résultats de l'échantillon. Les règles de décision sont :

[1] *Pour un test bilatéral où l'hypothèse alternative* « $\theta \neq \theta_0$ » est notée $\overline{H}_0$, on cherche $c_{1,\alpha/2}$ et $c_{2,\frac{\alpha}{2}}$ tels que $P_0(c_{1,\alpha/2} \leqslant T_0 \leqslant c_{2,\alpha/2}) = 1 - \alpha$ puis on rejette l'hypothèse H_0 au profit de $\overline{H}_0$ lorsque la valeur t_0^* prise par T_0 n'appartient pas à l'intervalle $[c_{1,\alpha/2} ; c_{2,\alpha/2}]$.

[2] *Pour un test avec rejet à droite de l'hypothèse nulle* H_0, l'hypothèse alternative H_1 étant « $\theta > \theta_0$ » ou bien « $\theta > \theta_1$ *avec* $\theta_1 > \theta_0$ », on détermine le nombre c_α tel que $P_0(T_0 \geqslant c_\alpha) = \alpha$ et l'on rejette H_0 au profit de l'hypothèse alternative H_1 lorsque la valeur numérique t_0^* prise par T_0 est supérieure à c_α. Le domaine de rejet de H_0 est du type $t_0^* \geqslant c_\alpha$.

En effet, fixant par exemple α = 1 %, on sait que si l'hypothèse H_0 est vraie, il y a seulement une chance sur 100 pour que T_0 prenne une valeur supérieure à $c_{0,01}$. Aussi, si après expérience, la valeur t_0^* prise par T_0 est effectivement supérieure à $c_{0,01}$ on peut penser que l'hypothèse H_0 qui génère sa distribution est fausse.

[3] *Pour un test avec rejet à gauche de l'hypothèse nulle* H_0 on détermine le nombre c'_α tel que $P_0(T_0 \leqslant c'_\alpha) = \alpha$ et l'on rejette H_0 au profit de l'hypothèse alternative H'_1 lorsque la valeur numérique t_0^* prise par T_0 est inférieure à c'_α. Le domaine de rejet de H_0 est de type $t_0^* \leqslant c'_\alpha$.

Section **2**

TESTS RELATIFS À LA VALEUR DE LA MOYENNE D'UNE DISTRIBUTION

À partir des valeurs numériques observées $x_1, x_2, \ldots, x_n$ d'un échantillon de taille n prélevé au hasard avec remise dans une population $\boldsymbol{P}$, on obtient une estimation numérique $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ de m_P et une estimation $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ de l'écart-type σ_P de la distribution.

1 Tests relatifs à la valeur de m_P avec échantillon de grande taille

Ces estimations permettent de réaliser un des tests d'hypothèses présentés au § 1.3 concernant la valeur de la moyenne m_P sur la population :

[1] Test bilatéral	[2] Test unilatéral avec rejet à droite de H_0	[3] Test unilatéral avec rejet à gauche de H_0
H_0 « $m_P = m_0$ » contre $\overline{H}_0$ « $m_P \neq m_0$ »	**[2a]** H_0 « $m_P = m_0$ » contre H_1 « $m_P > m_0$ »	**[3a]** H_0 « $m_P = m_0$ » contre H_1' « $m_P < m_0$ »
	[2b] H_0 « $m_P \leqslant m_0$ » contre H_1 « $m_P > m_0$ »	**[3b]** H_0 « $m_P \geqslant m_0$ » contre H_1' « $m_P < m_0$ »
	[2c] H_0 « $m_P = m_0$ » contre H_1 « $m_P = m_1$ avec $m_1 > m_0$ »	**[3c]** H_0 « $m_P = m_0$ » contre H_1' « $m_P = m_1$ avec $m_1 < m_0$ »

Variable statistique utilisée. Si on dispose d'un échantillon de grande taille $X_1, \ldots, X_n$ *iid* d'une loi quelconque $\mathcal{L}(m_P, \sigma_P)$ où m_P et σ_P sont inconnus, on utilise la propriété «$T = \dfrac{\overline{X} - m_p}{S/\sqrt{n}} \cong \mathcal{N}_{0,1}$ » (*cf.* 4.1 p. 157).

Sous l'hypothèse H_0 « $m_P = m_0$ *valeur donnée* », l'expression de T devient $T_0 = \dfrac{(\overline{X} - m_0)}{S/\sqrt{n}} \cong \mathcal{N}_{0;1}$, autrement dit $P(T_0 \leqslant t) \cong P(\mathcal{N}_{0;1} \leqslant t) = \Phi(t)$.

Après expérimentation, on connaît les valeurs numériques $\bar{x}$ et s prises par $\overline{X}$ et S et donc la valeur t_0^* prise par T_0 .

Règle de décision selon le cas envisagé

[1] Test bilatéral de H_0 « $m_P = m_0$ » **contre** $\overline{H}_0$ « $m_P \neq m_0$ »

Ayant choisi un risque d'erreur de 1ère espèce α (par exemple $\alpha = 1$ %) le domaine d'acceptation de H_0 est du type $]-c_{\alpha/2}, c_{\alpha/2}[$ où $\Phi(c_{\alpha/2}), = 1 - \alpha/2$.

On prendra la décision D_1 de rejeter H_0 au profit de $\overline{H}_0$ lorsque la valeur numérique t_0^* prise par T_0 n'appartient pas à l'intervalle $-]c_{\alpha/2}, c_{\alpha/2}[$.

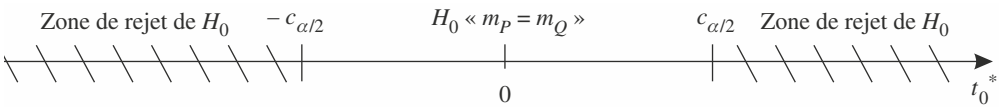


Figure 11.1

En effet sous l'hypothèse alternative $\overline{H}_0$ « $m_P \neq m_0$ », $\overline{X}$ converge en probabilité vers une valeur m_P différente de m_0 (soit $\overline{X} \xrightarrow[n\to\infty]{\text{pr}} m_P \neq m_0$) et donc $T_0 \cong \dfrac{\sqrt{n}(\overline{X} - m_0)}{S}$ tend à prendre de grandes valeurs positives ou négatives selon que $m_P > m_0$ ou *a contrario* $m_P < m_0$. On prend la décision D_1 de rejeter H_0 et donc d'accepter $\overline{H}_0$ si la valeur t_0^* prise par T_0 s'écarte de façon importante de 0 par valeur inférieure ou supérieure. Ces bornes de la zone de rejet sont calculées en fonction du *risque d'er-*

© Dunod. La photocopie non autorisée est un délit.

reur de première espèce $\alpha = P([T_0 \leqslant c_{1,\,\alpha/2}] \cup [T_0 \geqslant c_{2,\,\alpha/2}]/H_0$ vraie) à risque symétrique puisque $c_{1,\alpha/2}$ est défini par $\alpha/2 = P_0(T_0 \leqslant c_{1,\,\alpha/2}) \cong P(\mathcal{N}_{0;1} \leqslant c_{1,\,\alpha/2}) = \Phi(c_{1,\,\alpha/2})$ et $c_{2,\,\alpha/2}$ l'est par $\alpha/2 = P_0(T_0 \geqslant c_{2,\,\alpha/2}) \cong P(\mathcal{N}_{0;1} \geqslant c_{2,\,\alpha/2}) = 1 - \Phi(c_{2,\,\alpha/2})$ soit $-c_{1,\alpha/2} = c_{2,\alpha/2}$ noté $c_{\alpha/2}$.

[2] Test avec rejet à droite de l'hypothèse H_0

Le domaine de rejet de H_0 est du type $[c_\alpha, \infty[$ où c_α est défini par $\Phi(c_\alpha) \cong 1 - \alpha$ car $\alpha = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0;1} \geqslant c_\alpha)$.

Figure 11.2

On prend la décision D_1 de rejeter H_0 au profit de H_1 lorsque la valeur numérique t_0^* prise par T_0 est supérieure à $c_\alpha (i.e.\ t_0^* \geqslant c_\alpha)$.

En effet, sous l'hypothèse alternative H_1, $\overline{X} \xrightarrow[n\to\infty]{\text{pr}} m_P > m_0$ donc T_0 tend à prendre de grandes valeurs positives et par suite le domaine de rejet de H_0 est du type $[c_\alpha, \infty[$. On prend la décision D_1 de rejet de H_0 si $t_0^* \geqslant c_\alpha$. Le risque d'erreur de première espèce choisi étant α, on en déduit la valeur de c_α qui est défini par $\alpha = P(D_1/H_0$ vraie$) = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0;1} \geqslant c_\alpha)$ soit $\Phi(c_\alpha) = 1 - \alpha$.

[3] Test avec rejet à gauche de l'hypothèse H_0

Le domaine de rejet de H_0 est du type $]-\infty, c'_\alpha]$ où c'_α est défini par $\alpha = P_0(T_0 \leqslant c'_\alpha) \cong P(\mathcal{N}_{0;1} \leqslant c'_\alpha) = \Phi(c'_\alpha)$. On prend la décision D_1 de rejeter H_0 au profit de H'_1 lorsque $t_0^* \leqslant c'_\alpha$.

Figure 11.3

En effet, sous l'hypothèse alternative H'_1, $\overline{X}$ converge vers m_P qui est inférieur à m_0 donc T_0 tend à prendre de grandes valeurs négatives.

Exemple

Une machine produit en série des plaques de chocolat. Bien réglée, le poids X (évalué en grammes) est en moyenne m = 102 gr. Afin de tester si la machine s'est déréglée en cours de production, on prélève 50 plaques et on constate que le poids moyen $\bar{x}$ de ces 50 plaques est égal à 100,9 et que $s = \sqrt{\sum_{i=1}^{50}(x_i - \bar{x})^2/49} = 4$ gr. À partir de cette observa-

tion, on désire faire un « choix optimal » entre la décision D_0 « *laisser la production se poursuivre si l'on pense que* $m = 102$ » et la décision D_1 « *arrêter la machine pour la régler si l'on pense que* $m < 102$ », le risque d'erreur de première espèce α étant fixé à 5%.

La décision est subordonnée à la réalisation du test de H_0 « $m = 102$ » contre H_1' « $m < 102$ » avec $\alpha = 0{,}05$. Sous l'hypothèse de base H_0, la variable statistique utilisée

$$T_0 = \frac{(\overline{X} - m_0)}{S/\sqrt{n}} = \frac{(\overline{X} - 102)}{S/\sqrt{50}} \quad \text{suit sensiblement la loi } \mathcal{N}(0;\,1) :$$

$$T_0 = \frac{(\overline{X} - 102)}{S/\sqrt{50}} \cong \mathcal{N}_{0;1}.$$

Sous l'hypothèse alternative H_1' « $m < 102$ », T_0 tend à prendre des valeurs négatives car alors $\overline{X} \xrightarrow[n\to\infty]{\text{pr}} m < 102$. Nous sommes dans le cas [**3**], le domaine de rejet de l'hypothèse H_0 est donc du type $]-\infty,\, c'_\alpha]$ où $c'_\alpha = -1{,}64$ car $0{,}05 = P_0(T_0 \leqslant c'_\alpha) \cong \Phi(c'_\alpha)$. On prend la décision D_1 de rejeter H_0 si la valeur t_0^* prise par T_0 est inférieure à –1,64 soit si $t_0^* \leqslant -1{,}64$. Ici $t_0^* = \sqrt{50}(100{,}9 - 102)/4 = -1{,}94$ donc on rejette H_0 au profit de H_1'. Pour rendre moins rigide le choix de la valeur de α on calcule le ***niveau de signification observé*** $\alpha_c = P_0(T_0 \leqslant t_0^*) \cong P(\mathcal{N}_{0;1} \leqslant -1{,}94) = \Phi(-1{,}94) = 0{,}026$. Autrement dit avec un risque d'erreur de 2,6 % on décide de rejeter H_0.

REPÈRES : Puissance d'un test et erreur de seconde espèce.

Considérant le cas [**2**] d'un test avec rejet à droite, le domaine δ_α de rejet de l'hypothèse de base H_0 est de type $\delta_\alpha = [c_\alpha, \infty[$. On appelle puissance du test la valeur $\gamma = P(T_0 \in \delta_\alpha / H_1 \text{ vraie}) = P(T_0 \geqslant c_\alpha / H_1 \text{ vraie})$

Pour obtenir par exemple la puissance du test H_0 « $m_P = m_0$ » contre H_1 « $m_P > m_0$ » dans le cas de grands échantillons et lorsque la valeur de σ_P est connue, on utilise la variable statistique $T_0 = \sqrt{n}(\overline{X} - m_0)/\sigma_P$ dont la fonction de répartition, sous H_0, est approximée par la fonction de répartition de la loi normale centrée réduite $\Phi(t)$. On a

$\gamma = P(T_0 \geqslant c_\alpha / H_1 \text{ vraie}) = P(\sqrt{n}(\overline{X} - m_0)/\sigma_P \geqslant c_\alpha / H_1 \text{ vraie}) \cong \Phi(-c_\alpha + \sqrt{n}(m_P - m_0)/\sigma_P)$ avec $m_P > m_0$.

En effet, $\gamma = P([\sqrt{n}(\overline{X} - m_P)/\sigma_P + \sqrt{n}(m_P - m_0)/\sigma_P \geqslant c_\alpha] / H_1 \text{ vraie})$

$= P(\sqrt{n}(\overline{X} - m_P)/\sigma_P \geqslant c_\alpha + \sqrt{n}(m_0 - m_P)/\sigma_P \text{ sachant que } H_1 \text{ vraie})$

$\cong P(\mathcal{N}_{0\,;\,1} \geqslant c_\alpha + \sqrt{n}(m_0 - m_P)/\sigma_P)$

Sous H_1, $[\sqrt{n}(m_P - m_0)/\sigma_P - c_\alpha] \to \infty$ lorsque $n \to \infty$, donc $\gamma \to 1$ et par suite $\beta = 1 - \gamma \to 0$. On dit que le test est *convergent*.

Dans le cas [3] d'un rejet à gauche, γ serait estimée par $\Phi(c'_\alpha + \sqrt{n}(m_0 - m_P)/\sigma_P)$.

Remarque. Lorsque la valeur de σ_P n'est pas connue on lui substitue la valeur s prise par son estimateur : « $\gamma \cong \Phi(-c_\alpha + \sqrt{n}(m_P - m_0)/s)$ ». En effet l'échantillon étant de grande taille, il y a une forte probabilité pour que $s \cong \sigma_P$.

© Dunod. La photocopie non autorisée est un délit.

Courbe de puissance d'un test. Dans le cadre de l'exemple précédent, pour diverses valeurs supposées prises par m_P qui soient inférieures à 102 on peut calculer la puissance du test, ainsi pour $m = 101$ on trouve une puissance du test égale à $\gamma = \Phi(c'_\alpha + \sqrt{n}(m_0 - m_P)/s) = \Phi(-1{,}64 + \sqrt{50}(102 - 101)/4)$ $= \Phi(0{,}1277) = 0{,}44$ soit un risque d'erreur de seconde espèce de 56%.

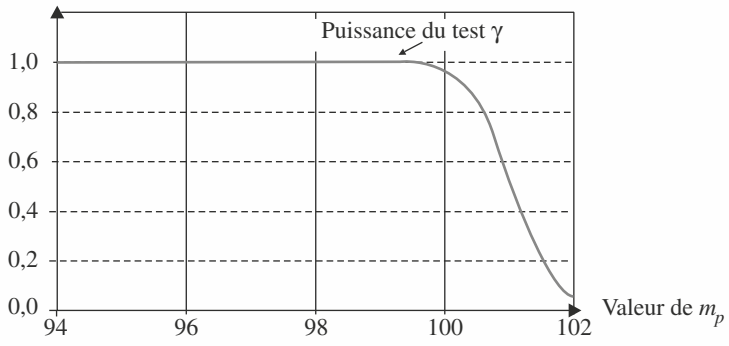


Figure 11.4

2 Tests relatifs à la valeur de m_P avec échantillon de petite taille

Si on ne dispose que d'un échantillon de petite taille, il convient de connaître la nature de la distribution afin de pouvoir utiliser les tests paramétriques.

2.1 Cas d'une distribution normale.

Si on est assuré que la distribution est normale, autrement dit si $X_1, \ldots, X_n$ est un échantillon *iid* d'une loi $\mathcal{N}(m,\sigma^2)$ dont la moyenne et l'écart-type ont des valeurs inconnues, on utilise la variable $T = \dfrac{\overline{X} - m}{S/\sqrt{n}} = \boldsymbol{t}_{n-1}$ (*cf.* p. 158). Sous l'hypothèse H_0 « $m_P = m_0$ », la variable $T_0 = \dfrac{\overline{X} - m_0}{S/\sqrt{n}}$ suit la loi de Student $\mathcal{S}t(n-1) : T_0 = \dfrac{\overline{X} - m_0}{S/\sqrt{n}} = \boldsymbol{t}_{n-1}$.

La méthodologie est analogue en tous points à celle présentée dans le § 2.1, la variable $\mathcal{N}_{0;1}$ étant ici remplacée par $\boldsymbol{t}_{n-1}$. Autrement dit, on utilise pour déterminer $-c_{\alpha/2}$ et $c_{\alpha/2}$ ou c_α ou c'_α la loi de Student à $n-1$ degrés de liberté, loi suivie par T_0 si l'hypothèse H_0 est vraie.

Remarque. On rejette H_0 dès lors que le niveau de signification observé α_c est inférieur ou égal au risque de première espèce que l'on s'est fixé. Cette remarque vaut pour tous les types de test et tous les paramètres testés.

Exemple

Un Institut d'Administration des Entreprises réalisant une enquête emploi auprès de ses diplômés se demande si le salaire d'entrée moyen de ses étudiants est supérieur ou non à 3 000 euros bruts par mois comme il est indiqué dans leur plaquette. Interrogeant dix étudiants on trouve un salaire moyen de $\bar{x} = 2\,500$ et un écart-type standard $s = 1\,000$. Après test d'ajustement on accepte l'hypothèse selon laquelle $X_1, \ldots, X_{10}$ est un échantillon *iid* d'une loi $\mathcal{N}(m,\sigma^2)$ dont la moyenne et l'écart-type ont des valeurs inconnues. À partir de ces observations on se propose de tester, avec risque d'erreur de première espèce de 5 %, l'hypothèse H_0^* « $m_P \geqslant 3\,000$ » contre H_1' « $m_P < 3\,000$ », **cas [3]** . À ce test on doit techniquement substituer le test de H_0 « $m_P = 3\,000$ » contre H_1' « $m_P < 3\,000$ ».

Sous l'hypothèse de base H_0, la variable statistique utilisée

$$T_0 = \frac{(\overline{X} - m_0)}{(S/\sqrt{n}} = \frac{\overline{X} - 3\,000)}{S/\sqrt{10}} = \boldsymbol{t}_9 .$$

Dans le **cas [3]** de test avec rejet à gauche de l'hypothèse H_0, le domaine de rejet de H_0 est $]-\infty, c'_\alpha]$ où c'_α est défini par $\alpha = P_0(T_0 \leqslant c'_\alpha) = P(\boldsymbol{t}_9 \leqslant c'_\alpha)$ soit $1-\alpha = 0{,}95 = P(\boldsymbol{t}_9 \leqslant -c'_\alpha)$ et on lit p. 378 $-c'_\alpha = 1{,}83$. La valeur $t_0^* = \dfrac{(2\,500 - 3\,000)}{1\,000/\sqrt{10}} = -1{,}58$ prise par T_0 étant supérieure à – 1,83 on ne peut rejeter H_0 et donc H_0^* au profit de H_1'.

Le niveau de signification observé $\alpha_c = P_0(T_0 \leqslant t_0^*) = P(\boldsymbol{t}_9 \leqslant -1{,}796) = 0{,}074$. Autrement dit, avec un risque d'erreur de 7,5 % on peut rejeter H_0 et donc *a fortiori* H_0^* au profit de H_1'.

2.2 Cas d'une distribution exponentielle

Si $X_1, \ldots, X_n$ est un échantillon *iid* d'une loi exponentielle non décalée de valeur moyenne m, on sait que la *v.a.*

$$T = \frac{2}{m}\sum_{i=1}^{n} X_i = 2 \times (X_1 + X_2 + \ldots + X_n)/m \text{ suit la loi } \chi^2(2n) \text{ [}cf.\text{ § 4.2 p. 141].}$$

Pour tester l'hypothèse H_0 « $m = m_0$ » contre l'une des hypothèses alternatives précédemment envisagées, on utilise la propriété :

$$\text{« sous } H_0,\ T_0 = \frac{2}{m_0} \times \sum_{i=1}^{n} X_i = \chi^2_{2n} \text{ ».}$$

La procédure est la même que pour un grand échantillon, la seule différence étant que la loi utilisée pour déterminer c_α ou c'_α ou $c_{\alpha/2}$ définis p. 184, est la loi suivie par T_0 sous H_0 et non la loi normale centrée réduite.

© Dunod. La photocopie non autorisée est un délit.

Section 3

TESTS RELATIFS À LA VALEUR DE L'ÉCART-TYPE D'UNE DISTRIBUTION

À partir des valeurs numériques observées $x_1, x_2, \ldots, x_n$ d'un échantillon de taille n prélevé au hasard avec remise dans une population statistique $\boldsymbol{P}$, on peut réaliser l'un des tests d'hypothèses présentés au § 1.3 concernant la valeur de l'écart-type σ_P de la distribution sur la population. Plus généralement, si on dispose d'un échantillon de taille n, $X_1, \ldots, X_n$ *iid* d'une loi $\mathcal{L}(m,\sigma)$ dont les valeurs de m et de σ sont inconnues, on peut être conduit à réaliser les mêmes types de tests concernant la valeur de σ.

Les différents types de test où H_0 désigne l'hypothèse « $\sigma_P^2 = \sigma_0^2$ *valeur donnée* » diffèrent suivant la forme prise par l'hypothèse alternative (*cf.* § 1.3 le paramètre θ étant ici la variance σ^2).

Trois cas peuvent être envisagés :

[1] l'hypothèse alternative est $\overline{H}_0$ « $\sigma_P^2 \neq \sigma_0^2$ »

[2] l'hypothèse alternative est H_1 « $\sigma_P^2 > \sigma_0^2$ » ou « $\sigma_P^2 = \sigma_1^2$ avec $\sigma_1^2 > \sigma_0^2$ », (rentre dans cette catégorie le cas où H_0 est formulée de la façon suivante « $\sigma_P^2 \leqslant \sigma_0^2$ » au lieu de $\sigma_P^2 = \sigma_0^2$).

[3] l'hypothèse alternative est H_1' « $\sigma_P^2 < \sigma_0^2$ » ou « $\sigma_P^2 = \sigma_1^2$ avec $\sigma_1^2 < \sigma_0^2$ » (est incluse dans cette catégorie le cas où H_0 s'exprime de la façon suivante « $\sigma_P^2 \geqslant \sigma_0^2$ » au lieu de $\sigma_P^2 = \sigma_0^2$.

1 Tests relatifs à la valeur de σ avec échantillon de grande taille

– **Variable statistique utilisée.** Si on dispose d'un grand échantillon $X_1, \ldots, X_n$, pour réaliser le test, on utilise la variable $T_0 = \dfrac{S^2 - \sigma_0^2}{\sqrt{(\hat{\mu}_4 - S^4)/n}}$ qui, sous l'hypothèse H_0 « $\sigma_p^2 = \sigma_0^2$ *valeur donnée* », suit sensiblement la loi normale standard $\mathcal{N}_{(0,1)}$:

$$\text{sous l'hypothèse } H_0,\ T_0 = \frac{(S^2 - \sigma_0^2)}{\sqrt{(\hat{\mu}_4 - S^4)/n}} \cong \mathcal{N}_{0,1}$$

Après expérimentation on connaît les valeurs numériques $\bar{x}, s, \mu_4^*$

et donc la valeur $t_0^* = \dfrac{(s^2 - \sigma_0^2)}{\sqrt{(\mu_4^* - s^4)/n}}$ prise par T_0.

– Règle de décision selon le cas envisagé.

Dans le cas [1] de test bilatéral le domaine d'acceptation de H_0 est du type $[-c_{\alpha/2}, c_{\alpha/2}]$ où $1-\alpha/2 = \Phi(c_{\alpha/2})$.

Dans le cas [2] de test avec rejet à droite de l'hypothèse H_0 le domaine de rejet de H_0 est du type $[c_\alpha, \infty[$ où c_α est tel que $\alpha = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0;1} \geqslant c_\alpha)$ et donc $\Phi(c_\alpha) \cong 1-\alpha$. On prend la décision D_1 de rejeter H_0 au profit de H_1 lorsque $t_0^* \geqslant c_\alpha$.
A contrario si $t_0^* \leqslant c_\alpha$ on prend la décision D_0 d'accepter l'hypothèse H_0 et ce, avec un risque d'erreur de seconde espèce $\beta = 1-\gamma$ où la puissance γ du test est estimée par $\gamma \cong \Phi\left(-c_\alpha + \dfrac{\sqrt{n}(\sigma_P^2 - \sigma_0^2)}{\sqrt{(\mu_4^* - s^4)}}\right)$.

Dans le cas [3] de test avec rejet à gauche de l'hypothèse H_0, le domaine de rejet de H_0 est du type $]-\infty, c'_\alpha]$ où c'_α est défini par $\alpha = P_0(T_0 \leqslant c'_\alpha) \cong P(\mathcal{N}_{0;1} \leqslant c'_\alpha) = \Phi(c'_\alpha)$, la puissance du test étant estimée par $\Phi\left(c'_\alpha + \dfrac{\sqrt{n}(\sigma_0^2 - \sigma_P^2)}{\sqrt{(\mu_4^* - s^4)}}\right)$.

Exemple

Une machine embouteille de l'eau minérale. Bien réglée, la variance de la quantité d'eau embouteillée (évalué en millilitres) est égale à 25 ml. Afin de tester si la machine s'est déréglée en cours de production on a prélevé 100 bouteilles d'eau et constaté sur cet échantillon un écart-standard $s = 6$ et $\mu_4^* = \dfrac{1}{100}\sum_{i=1}^{n}(x_i - \bar{x})^4 = 1400$.

À partir de ces observations on se propose de tester, avec un risque d'erreur de première espèce de 10 %, l'hypothèse H_0^* « $\sigma_P \leqslant 5$ » contre H_1 « $\sigma_P > 5$ ». À ce test on doit techniquement substituer le test de H_0 « $\sigma_P = 5$ » contre H_1 « $\sigma_P > 5$ ».

L'échantillon étant grand on sait que sous l'hypothèse H_0, $T_0 = \dfrac{\sqrt{n}(S^2 - 5^2)}{\sqrt{\hat{\mu}_4 - S^4}} \cong \mathcal{N}_{0;1}$.

Règle de décision. Le domaine de rejet de H_0 est du type $[c_\alpha, \infty[$ où. En effet, sous l'hypothèse alternative H_1 on a : $S^2 \xrightarrow[n\to\infty]{\text{pr}} \sigma_P^2$ avec $\sigma_P^2 > 5^2$, donc $T_0/\sqrt{n} \xrightarrow[n\to\infty]{\text{pr}} \dfrac{\sigma_P^2 - 5^2}{\sqrt{\mu_{4,P} - \sigma_P^4}}$ et par suite T_0 tend donc à prendre des valeurs positives grandes. Le nombre c_α est défini par $\alpha = 0{,}1 = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0;\,1} \geqslant c_\alpha)$.
Donc $\Phi(c_\alpha) \cong 0{,}90$ et par suite $c_\alpha = 1{,}28$.

La valeur $t_0^* = (6^2 - 5^2)/\sqrt{(1\,400 - 6^4)/100} = 10{,}79$ prise par T_0 étant supérieure à 1,28 on prend la décision D_1 de rejeter H_0 et donc *a fortiori* H_0^* au profit de H_1 et ce avec un risque d'erreur inférieur à 10 %.

© Dunod. La photocopie non autorisée est un délit.

2 Tests relatifs à la valeur de σ avec échantillon de petite taille

Si on ne dispose que d'un échantillon de petite taille, les tests paramétriques ne permettent pas d'apporter de réponse au problème étudié. Toutefois si on est assuré que la ***distribution est normale***, autrement dit si $X_1, \ldots, X_n$ est un échantillon *iid* d'une loi $\mathcal{N}(m, \sigma^2)$ dont la moyenne et l'écart-type ont des valeurs inconnues on a la propriété : « $T = (n-1)S^2/\sigma^2 = \chi^2_{n-1}$ ».

– **Variable statistique utilisée.** Pour réaliser le test on utilise la variable $T_0 = (n-1)S^2/\sigma_0^2$ qui, sous l'hypothèse H_0 « $\sigma = \sigma_0$ », suit la loi $\chi^2(n-1)$: $T_0 = (n-1)S^2/\sigma_0^2 = \chi^2_{n-1}$.

– **Règle de décision selon le cas envisagé.**

Dans le cas [1] de test bilatéral, le domaine d'acceptation de H_0 est du type $]c_{1,\alpha/2}, c_{2,\alpha/2}[$ où $\alpha/2 = P_0(T_0 \leqslant c_{1,\,\alpha/2}) = P(\chi^2_{n-1} \leqslant c_{1,\alpha/2})$ et où $\alpha/2 = P(T_0 \geqslant c_{2,\,\alpha/2}) = P(\chi^2_{n-1} \geqslant c_{2,\alpha/2})$. On rejette H_0 si la valeur $t_0^* = (n-1)s^2/\sigma_0^2$ prise par T_0 n'appartient pas à $]c_{1,\alpha/2}, c_{2,\alpha/2}[$. Le niveau de signification observé α_c est tel que $\alpha_c/2 = \text{Min}\,[P(\chi^2_{n-1} \leqslant t_0^*), P(\chi^2_{n-1} \geqslant t_0^*)]$. Avec un risque d'erreur égal à α_c on peut rejeter H_0 au profit de $\overline{H}_0$.

Dans le cas [2] de test avec rejet à droite de l'hypothèse H_0, on détermine c_α tel que $P_0(T_0 \geqslant c_\alpha) = P(\chi^2_{n-1} \geqslant c_\alpha) = \alpha$ puis l'on rejette H_0 au profit de H_1 si la valeur t_0^* est supérieure à c_α. Le niveau de signification observé $\alpha_c = P(\chi^2_{n-1} \geqslant t_0^*)$.

Dans le cas [3] de test avec rejet à gauche de l'hypothèse H_0 où par exemple H_1' est « $\sigma < \sigma_0$ », on détermine c'_α tel que $P_0(T_0 \leqslant c'_\alpha) = P(\chi^2_{n-1} \leqslant c'_\alpha) = \alpha$ et l'on rejette H_0 si t_0^* est inférieure à c'_α. Le niveau de signification observé $\alpha_c = P(\chi^2_{n-1} \leqslant t_0^*)$.

Exemple

Reprenant l'exemple de la page 187 où $s = 1\,000$, on se propose de tester, avec risque d'erreur de première espèce de 2 %, l'hypothèse H_0 « $\sigma_P = 1\,500$ » contre H_1' « $\sigma_P \neq 1\,500$ ».

– ***Variable statistique utilisée.*** Pour réaliser le test on utilise la variable, $T_0 = (n-1)S^2/\sigma_0^2$ qui, sous l'hypothèse H_0 « $\sigma_P = \sigma_0$ », suit la loi $\chi^2(n-1)$: $T_0 = (9)S^2/1\,500^2 = \chi^2_9$.

– ***Règle de décision.*** Dans ce **cas [1]** de test bilatéral, le domaine d'acceptation de H_0 est du type $[c_{1,\alpha/2}, c_{2,\alpha/2}]$ où $0{,}01 = \alpha/2 = P_0(T_0 \leqslant c_{1,\,\alpha/2}) = P(\chi^2_9 \leqslant c_{1,\,\alpha/2})$ et où $0{,}01 = \alpha/2 = P(T_0(c_{2,\,\alpha/2}) = P(\chi^2_9 \geqslant c_{2,\,\alpha/2})$ d'où $c_{1,\,\alpha/2} = 2{,}09$ et $c_{2,\,\alpha/2} = 21{,}66$. On ne peut rejeter H_0 car la valeur $t_0^* = 9 \times 1\,000^2/1\,500^2 = 4$ prise par T_0 appartient au domaine d'acceptation.

Remarque. Lorsque $X_1, \ldots, X_n$ est un échantillon *iid* d'une loi $\mathcal{N}(m,\sigma^2)$ dont la moyenne m a une valeur connue et l'écart-type est inconnu on utilise, au lieu de la statistique présentée ci-dessus, la variable statistique $T_0 = \dfrac{1}{n\sigma_0^2}\sum_{i=1}^{n}(X_i - m)^2 = \chi_n^2$.

REPÈRES : Puissance du test

Si l'on note δ_α le domaine de rejet de H_0 lorsque l'on teste H_0 contre H_1' « $\sigma^2 < \sigma_0^2$ », on obtient la puissance du test en calculant

$$\gamma = P(T_0 \in \delta_\alpha / H_1' \text{ vraie}) = P((n-1)S^2/\sigma_0^2 \leqslant c_\alpha')$$
$$= P((n-1)S^2 / \sigma^2 \leqslant \sigma_0^2 c_\alpha' / \sigma^2) = P(\chi_{n-1}^2 \leqslant \sigma_0^2 c_\alpha' / \sigma^2).$$

Pour s'assurer que ce test est convergent il suffit d'approximer la loi du khi-deux par la loi normale et remarquer que $(c_\alpha' - n + 1)/\sqrt{2n-2} = a_\alpha$ où $\Phi(a_\alpha) = 1 - \alpha$

Section **4**

TESTS DE VALEUR D'UNE PROPORTION

Soit une population binomiale ***P*** dont une proportion p possède le caractère considéré $\mathcal{A}$. Si on s'intéresse à la valeur de p, l'on peut être conduit à réaliser un des tests de même type que ceux proposés au § 2.1 où p remplace m_P.(*Cf.* § 5.1 p. 169).

Ayant extrait de ***P*** un échantillon non exhaustif de taille n, notons K le nombre aléatoire d'éléments de l'échantillon qui possèdent le caractère $\mathcal{A}$ et $F = K/n$ la proportion aléatoire d'éléments de l'échantillon ayant ce caractère considéré.

Les différents types de test où H_0 désigne l'hypothèse « $p = p_0$ *valeur donnée* » diffèrent suivant la forme prise par l'hypothèse alternative. Dans le cas

[1] l'hypothèse alternative est $\overline{H}_0$ « $p \neq p_0$ »

[2] l'hypothèse alternative est H_1 « $p > p_0$ » ou « $p = p_1$ avec $p_1 > p_0$ »

[3] l'hypothèse alternative est H_1' « $p < p_0$ » ou H « $p = p_1$ avec $p_1 < p_0$ »

Variable statistique utilisée. Si on dispose d'un grand échantillon non exhaustif on a la propriété : « $T = \dfrac{F-p}{\sqrt{p(1-p)/n}} \cong \mathcal{N}_{0\,;\,1}$ ». Sous l'hypothèse H_0 « $p = p_0$ v*aleur donnée* », la *v.a.* $T_0 = \dfrac{F-p_0}{\sqrt{p_0(1-p_0)/n}} \cong \mathcal{N}_{0\,;\,1}$.

Après expérimentation on constate que F prend la valeur numérique $f = k/n$ et donc T_0 prend la valeur $t_0^* = \dfrac{f-p_0}{\sqrt{p_0(1-p_0)/n}}$.

© Dunod. La photocopie non autorisée est un délit.

Règle de décision selon le cas envisagé

[1] Test de H_0 « $p = p_0$ » contre $\overline{H}_0$ « $p \neq p_0$ ». Le domaine d'acceptation de H_0 est du type $[-c_{\alpha/2}, c_{\alpha/2}]$ où $c_{\alpha/2}$ est défini par $1 - \alpha = P_0(-c_{\alpha/2} \leqslant T_0 \leqslant c_{\alpha/2}) \cong P_0(-c_{\alpha/2} \leqslant \mathcal{N}_{0\,;\,1} \leqslant c_{\alpha/2})$ et donc $\Phi(c_{\alpha/2}) = 1 - \alpha/2$. On prend la décision D_1 de rejeter H_0 au profit de $\overline{H}_0$ lorsque $t_0^* \notin [-c_{\alpha/2}, c_{\alpha/2}]$. Le niveau de signification observé α_c est tel que $\alpha_c/2 = P(\mathcal{N}_{0\,;\,1} \geqslant |t_0^*|)$.

[2] Test avec rejet à droite de l'hypothèse H_0. Le domaine de rejet de H_0 est du type $[c_\alpha, \infty[$ où c_α est défini par $\alpha = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0\,;\,1} \geqslant c_\alpha)$ et donc par $\Phi(c_\alpha) = 1 - \alpha$. On prend la décision D_1 de rejeter H_0 au profit de H_1 lorsque $t_0^* \geqslant c_\alpha$. Le niveau de signification observé $\alpha_c = P_0(T_0 \geqslant t_0^*) \cong P(\mathcal{N}_{0\,;\,1} \geqslant t_0^*)$.

[3] Test avec rejet à gauche de l'hypothèse H_0. Le domaine de rejet de H_0 est du type $]-\infty, c'_\alpha\,]$ où c'_α est telle que $\Phi(c'_\alpha) = \alpha$. On prend la décision D_1 de rejeter H_0 au profit de H'_1 lorsque $t_0^* \leqslant c'_\alpha$. En effet, sous l'hypothèse alternative : $F \xrightarrow[n\to\infty]{\text{pr}} p$ avec $p < p_0$, donc T_0 tend à prendre des valeurs négatives. Le niveau de signification observé $\alpha_c = P(\mathcal{N}_{0\,;\,1} \leqslant t_0^*)$.

REPÈRES : Puissance du test

Pour évaluer la puissance du test on utilise la propriété « $T = \dfrac{F - p}{\sqrt{p(1-p)/n}} \cong \mathcal{N}_{0\,;\,1}$ » et l'on considère la variable $T_0 = \dfrac{F - p_0}{\sqrt{p_0(1-p_0)/n}}$ qui sert à réaliser le test.

– Pour déterminer par exemple la puissance γ du test H_0 « $p = p_0$ » contre H_1 « $p > p_0$ » il suffit de remarquer que $\gamma = P(T_0 > c_\alpha \;/\; H_1 \text{ vraie}) = P\left(\dfrac{F - p}{\sqrt{p(1-p)/n}} \geqslant k_n\right)$ où $k_n = \dfrac{c_\alpha\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}} + \dfrac{\sqrt{n}(p_0 - p)}{\sqrt{p(1-p)}}$ et donc $\gamma \cong P(\mathcal{N}_{0\,;\,1} \geqslant k_n) = 1 - \Phi(k_n)$ expression qui intègre ***la vraie valeur*** p qui n'est pas connue et à laquelle on substitue sa valeur estimée f. On en déduit une estimation de l'erreur de seconde espèce $\beta = 1 - \gamma$ que l'on peut commettre en acceptant H_0 alors que c'est H_1 qui est vraie. Le test est bien convergent puisque $k_n \to -\infty$ et par suite $\gamma \to 1$ lorsque $n \to \infty$.

– De même on établit que la puissance γ' du test H_0 « $p = p_0$ » contre H'_1 « $p < p_0$ » est sensiblement égale à $\Phi(k'_n)$ où $k'_n = \dfrac{c'_\alpha\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}} + \dfrac{\sqrt{n}(p_0 - p)}{\sqrt{p(1-p)}}$.

Remarque. Lorsque l'échantillon de taille n extrait de la population $P = \{e_1, e_é, \ldots, e_N\}$ est exhaustif on utilise la *v.a.* $\boldsymbol{T}'_0 = \dfrac{F - p_0}{\sqrt{\dfrac{N-n}{N-1}\,p_0(1-p_0)/n}}$ qui, sous H_0, suit sensiblement la loi normale $\mathcal{N}(0; 1)$.

Exemple

Une machine produit en série des pièces dont une proportion p de défectueuses. Si la machine est bien réglée $p = 3\ \%$. Si elle est déréglée $p = 5\ \%$. Pour savoir si elle s'est déréglée en cours de production, on se propose de prélever un échantillon de 1 000 unités. F désigne la proportion aléatoire de pièces défectueuses dans l'échantillon. À partir de la valeur prise par F soit $f = 0{,}035$ on se propose de tester, avec un niveau de signification α fixé à 5 %, l'hypothèse de base H_0 « $p = 3\ \%$ » contre l'hypothèse alternative H_1 « $p = 5\ \%$ », aucune autre éventualité n'étant possible.

Statistique utilisée. Sous l'hypothèse H_0 : $T_0 = \dfrac{(F - 0{,}03)}{\sqrt{0{,}03(1 - 0{,}03)/1\,000}} \cong \mathcal{N}_{0\,;\,1}$

Règle de décision. **Cas [2]**. Le domaine de rejet de H_0 est du type $[c_\alpha, \infty[$ où c_α est défini par $\alpha = 0{,}05 = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0\,;\,1} \geqslant c_\alpha)$ et donc $\Phi(c_\alpha) = 1 - \alpha = 0{,}95$ soit $c_\alpha = 1{,}64$. La variable T_0 prend la valeur numérique $t_0^* = (0{,}035 - 0{,}03)/\sqrt{0{,}03 \times 0{,}97/1\,000} = 0{,}93$ qui n'appartient pas au domaine de rejet de H_0 donc on prend (avec un risque d'erreur de seconde espèce β qu'il faut évaluer) la décision D_0.

Risque de seconde espèce. Sous H_1, on a $\dfrac{(F - 0{,}05)}{\sqrt{0{,}05(1 - 0{,}05)/1\,000}} \cong \mathcal{N}_{0\,;\,1}$ donc

$$\beta = P(T_0 \leqslant c_\alpha / H_1 \text{ vrai} = P\left(\frac{(F - 0{,}03)}{\sqrt{0{,}03(1 - 0{,}03/1\,000}} \leqslant 1{,}64/p = 0{,}05\right)$$

$$= P\left(F \leqslant 0{,}039/p = 0{,}05\right) = P\left(\frac{(F - 0{,}05)}{\sqrt{0{,}05(1 - 0{,}05)/1\,000}} \leqslant -1{,}6/p = 0{,}05\right)$$

$$= \Phi(-1{,}60) = 5{,}5\ \%$$

Si l'on s'intéresse au test de H_0 « $p = 3\ \%$ » contre H_1 « $p = \pi$ » prenons successivement $\pi = 3{,}5\ \%$, 4 %, 5 % etc. On peut dans chaque cas calculer l'erreur de seconde espèce $\beta(\pi) = P(F > c/p = \pi)$ puis tracer la **courbe des caractéristiques opérationnelles** (CCO) $\pi \mapsto \beta(\pi)$ et la **courbe de puissance du test** $\pi \mapsto \gamma(\pi)$ où $\gamma(\pi) = 1 - \beta(\pi)$.

Dans le cas des tests unilatéraux où l'hypothèse nulle H_0^* peut s'écrire « $p \leqslant p_0$ » dans le cas **[2]**, « $p \geqslant p_0$ » dans le cas **[3]**, utiliser $T_0' = \dfrac{(F - p_0)}{\sqrt{F(1 - F)/(n)}} \cong \mathcal{N}_{0\,;\,1}$, la statistique T_0' étant utilisée en lieu et place de T_0.

© Dunod. La photocopie non autorisée est un délit.

Section 5

TEST DE SYMÉTRIE D'UNE DISTRIBUTION

Soit un échantillon $X_1, X_2 \ldots, X_n$ *iid* d'une loi continue F (où F désigne la fonction de répartition) dont on ignore la nature et pour laquelle on souhaite tester l'hypothèse H_0 de symétrie : $F(m + x) + F(m - x) = 1 \ \forall x$.

Si la valeur de la moyenne m est connue il suffit d'appliquer le test des rangs signés de Wilcoxon à l'échantillon $Y_1, Y_2 \ldots, Y_n$ (où $Y_i = X_i - m$) dont la distribution est symétrique autour de 0 (*cf.* § 6 p. 134). Généralement, la valeur de la moyenne sur la population m étant inconnue, on lui substitue la valeur moyenne $\overline{x}$ de l'échantillon lorsque n est grand.

Variable statistique utilisée

Les n valeurs numériques y_i prises par les n *v.a.* Y_i étant rangées par ordre de croissance des valeurs absolues : $|y_1| < |y_2| < \ldots < |y_n|$, la somme t^+ des rangs des valeurs y_i positives est la valeur prise par une variable T_n^+ qui suit la loi $W^+(n)$.

Règle de décision

Soit h_0 le plus grand entier tel $P(W_n^+ \leqslant h_0) \leqslant \alpha/2$. Si « $t^+ \leqslant h_0$ ou $t^+ \geqslant n(n+1)/2 - h_0$ » on décide de rejeter l'hypothèse de symétrie avec un risque d'erreur α.

Exemple

Souhaitant réaliser une étude concernant l'évolution de l'activité dans un secteur, le responsable des études de la banque B a relevé la variation du chiffre d'affaires réel x réalisé par 20 entreprises de ce secteur entre 2005 et 2006 (unité = 10^5 euros) :

Entreprise	1	2	3	4	5	6	7	8	9	10
x_i	–8,5	–16,2	–55	–37,3	–5,6	–14,6	1,0	7	16	–5,1
Entreprise	11	12	13	14	15	16	17	18	19	20
x_i	25	2,1	–11,6	–12,7	53	88	–4,1	9	4,3	46,6

On souhaite tester l'hypothèse de symétrie de la distribution autour de la valeur 0. Pour cela on soumet les 20 réalisations $x_1, x_2, \ldots, x_{20}$ de l'échantillon au classement de leurs

valeurs absolues : $|x_{i1}| < |x_{i2}| < \ldots < |x_{i20}|$. Notons r_i^a le rang de la i-ième observation x_i dans le classement des valeurs absolues et soulignons les rangs associés à des valeurs x_i positives.

$\lvert x_i \rvert$	8,5	16,2	55	37,3	5,6	14,6	1,0	7	16	5,1
r_i^a	8	14	19	16	6	12	1	7	13	5
$\lvert x_i \rvert$	25	2,1	11,6	12,7	53	88	4,1	9	4,3	46,6
r_i^a	15	2	10	11	18	20	3	9	4	17

La *v.a.* W_{20}^+ peut prendre n'importe quelle valeur entière comprise entre 0 et $n(n+1)/2$ où $n = 20$. La valeur w prise par W_{20}^+ est la somme des rangs r_i^a des observations qui ont une valeur positive : $W_{20}^+ = 1 + 7 + 13 + 15 + 2 + 18 + 20 + 9 + 4 = 89$.

Sous l'hypothèse H_0 et puisque $n \geqslant 15$, on peut utiliser l'approximation normale avec correction de continuité : pour tout $h \in \Delta_{W^+}$, $P_0(W_{20}^+ \leqslant h) \cong \Phi[\theta(h)]$ où $\theta(h) = [h + 0.5 - n(n+1)/4]/[n(n+1)(2n+1)/24]^{1/2}$ et ici $n = 20$.

On constate que $P_0(W_{20}^+ \leqslant 89) \cong \Phi[\theta(89)] = \Phi(-0{,}578) \cong -28{,}17$ %. On ne peut donc pas rejeter H_0 (le niveau de signification observé est de 56, 34%).

Section 6

TEST DE LA MOYENNE AVEC SPSS ET EXCEL

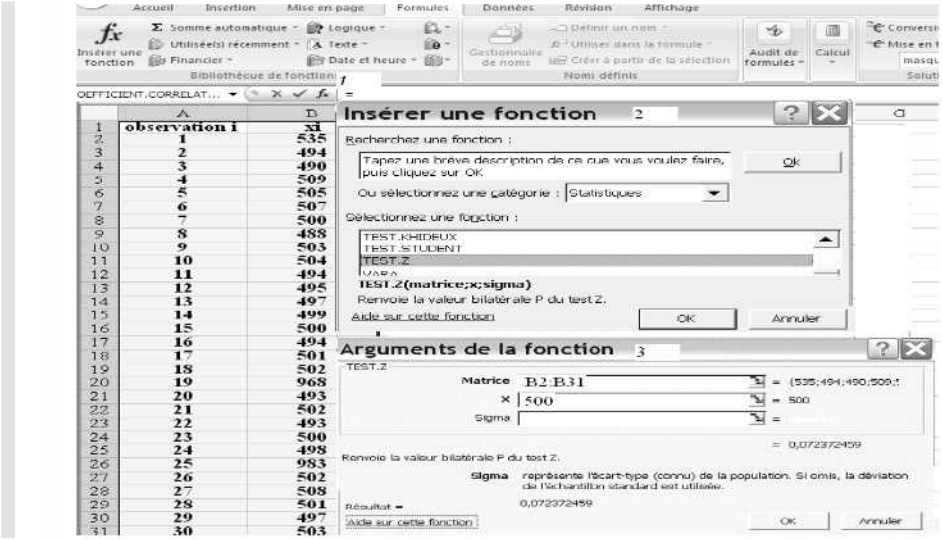
1 Traitement avec EXCEL

Pour les tests sur la valeur de la moyenne m_P lorsque σ_P est inconnu, on utilise $\dfrac{\overline{X} - m_p}{S/\sqrt{n}} \cong \mathcal{N}_{0;1}$. Si au contraire l'on connaît la valeur de σ_P. on peut opter pour $\dfrac{\overline{X} - m_p}{\sigma_P/\sqrt{n}} \cong \mathcal{N}_{0;1}$.

Exemple

Le directeur des ventes d'une usine de conditionnement de lait pour enfant souhaite contrôler le poids net des boîtes de 500 gr de lait pour premier âge. Il fait procéder à un tirage au hasard non exhaustif de 30 boîtes et obtient les résultats ci-dessous où x_i désigne la contenance de la i-ème boîte de l'échantillon (en ml) $i = 1, 2, \ldots, 30$ (*cf.* colonne A et B ci-après).

© Dunod. La photocopie non autorisée est un délit.



Procédure

1. On se place sur une cellule non utilisée on clique **Fx**

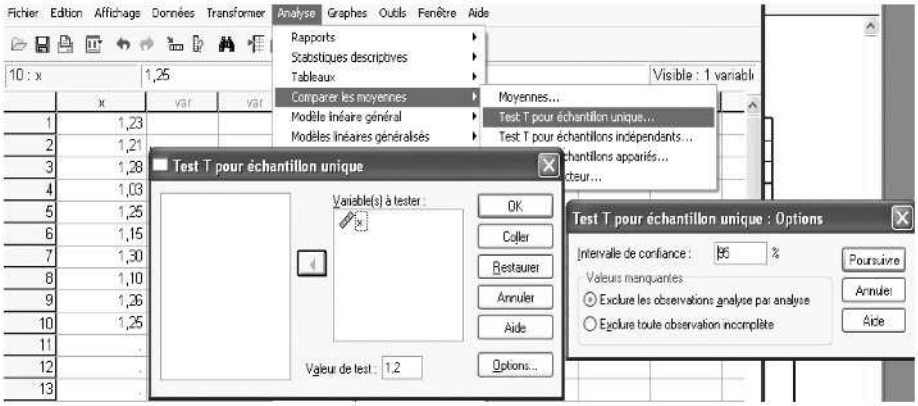
2. On sélectionne le menu **Statistiques** et le sous-menu **TEST.Z** puis cliquer sur **OK**

3. On obtient le menu **Arguments de la fonction**, dans lequel on sélectionne les références des cellules des données soit **B2 : B31** en tant que **Matrice**, puis on entre dans **x** la valeur testée du paramètre soit **500**. L'écart-type étant inconnu on ne met aucune valeur dans **Sigma** puis **OK**

Dans le coin droit de la fenêtre on obtient le niveau de signification observé α égal à 0,07237 $[\alpha/2 = P_0(T_0 \geqslant |(\bar{x} - m_0)/(s/\sqrt{n})|)]$, soit $\alpha = 0{,}07\,237$. Si on décide de rejeter H_0 le risque minimum d'erreur est de 7,24 %.

2 Traitement avec SPSS (uniquement ici lorsque $X_1, \ldots, X_n$ est un échantillon *iid* d'une loi $\mathcal{N}(m, \sigma^2)$ avec m et σ inconnus)

On a recensé la durée de marche ininterrompue de 10 appareils de même type et obtenu les résultats suivants : 1,23 – 1,21 – 1,28 – 1,03 – 1,25 – 1,15 – 1,30 – 1,10 – 1,26 – 1,25 . Considérant que « la durée de marche ininterrompue suit sensiblement une loi normale », on teste l'hypothèse H_0 selon laquelle la durée moyenne « $m = 1{,}2$ » contre $\overline{H_0}$ « $m \neq 1{,}2$ ».



Procédure.

1. Cliquer sur **Analyse**, **Comparer les moyennes** et **Test T pour échantillon unique**.

2. On saisit ***x*** la variable que l'on souhaite tester que l'on envoie dans **Variable à tester**. On met la valeur testée du paramètre soit 1,2 dans **Valeur du test**.On peut cliquer sur **OK** et l'on obtient le résultat du test pour un risque d'erreur de 5 %

3. Si l'on veut changer ce risque d'erreur on clique sur **Options**et on obtient le menu ci-dessous qui permet de modifier le niveau de confiance souhaité $(1-\alpha)$.

Résultats. Intitulés « Statistiques sur échantillon unique »

	N	Moyenne	Écart-type	Erreur standard moyenne
X	10	1,206	0,08630695	0,02729265

Test sur échantillon unique						
Valeur du test = 1,2						
	t	*ddl*	Sig. (bilatérale)	Différence moyenne	Intervalle de confiance 95% de la différence	
					inférieure	supérieure
X	0,21983939	9	0,83090161	0,006	– 0,0557	0,0677

Du premier tableau on déduit que $\bar{x} = 1{,}206$ et $s = 0{,}0863$ et $s/\sqrt{n} = 0{,}02729$.

Du deuxième tableau on déduit que la statistique T_0, qui sert à faire un test sur la moyenne, suit la loi de Student à 9 *ddl* et prend la valeur $t_0^* = (\bar{x} - 1{,}2)/(s/\sqrt{n}) = 0{,}2189$. Le Sig. correspond au niveau de signification observé 0,8309 d'un test bilatéral : $\alpha_c/2 = P_0(T_0 \geqslant |(\bar{x} - m_0)/(s/\sqrt{n})|)$, d'où $\alpha_c = 0{,}8309$. Si on décide de rejeter H_0 le risque minimum d'erreur est de 83,09 %.

On constate que la différence $\bar{x} - m_0 = 0{,}006$ et on décide d'accepter H_0 avec un niveau de confiance de 95 % lorsque $-0{,}0557 < \overline{x} - m_0 < 0{,}0677$.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 355)

Exercice 1

La limite légale d'un certain polluant contenu dans les déchets d'une usine est de 6 mg par kg de déchet. On effectue un dosage sur 12 prélèvements de 1 Kg pour lesquels on obtient comme valeurs de la moyenne $\bar{x} = 6{,}8$ et de la variance standard $s^2 = 3{,}6$. On suppose que la variable aléatoire X dont les valeurs représentent la quantité de polluants en mg/kg a une répartition normale de moyenne m et d'écart-type σ inconnus.

1. Sachant qu'au delà de cette limite de 6 mg/kg l'usine doit verser une amende, quel test doit-on envisager ? Compte tenu des résultats des prélèvements, quelle est la décision de ce test pour un risque d'erreur de 5 % ? Déterminer le niveau de signification observé c'est-à-dire le risque minimum d'erreur si on décide de rejeter H_0 ?

2. Tester avec un risque d'erreur de première espèce de 5 % l'hypothèse H_0 « $\sigma = 2$ » contre $\overline{H}_0$ « $\sigma \neq 2$ ».

Exercice 2

Le service commercial de la société se demande s'il doit accélérer, au prix d'une dépense supplémentaire, le lancement d'une campagne publicitaire. Il estime que si la proportion p d'individus qui connaissent la marque est inférieure ou égale à 30 %, il est nécessaire d'accélérer le lancement de la campagne publicitaire car l'insuffisance de notoriété des produits risque d'entraîner des pertes de marché considérables. Par contre si la proportion p d'individus qui connaissent la marque est supérieure à 30 % il n'est pas nécessaire de lancer cette campagne publicitaire. Sur un échantillon de taille 500 le service responsable des études de marché observe que 100 personnes connaissent la marque.

Le service commercial doit-il oui ou non accélérer le lancement de la campagne publicitaire ? (formuler un test au seuil de 10 %), en explicitant le choix de base.

Exercice 3

La chambre de commerce et d'industrie souhaite savoir si la croissance du chiffre d'affaires des restaurateurs est supérieure au taux d'inflation constaté sur la période soit 4 %. Disposant d'un échantillon de 60 restaurants, le responsable des études observe sur cet échantillon que le chiffre d'affaires a augmenté en moyenne de 3,8 %, que l'écart-standard $s = 2{,}1$ et que le moment centré d'ordre 4 $\mu_4^* = 19{,}55$.

1. m_P désignant la moyenne des taux de croissance sur l'ensemble du secteur, tester l'hypothèse H_0 « $m_P = 4$ » contre $\overline{H}_0$ « $m_P \neq 4$ » avec un risque d'erreur de 5 % puis déterminer le niveau de signification du test. (En fait si les données statistiques le permettaient, une analyse par strate s'imposerait).

2. Tester l'hypothèse H_0 « $\sigma_P = 2{,}2$ » contre H_1' « $\sigma_P < 2{,}2$ » avec un risque d'erreur de 10 % puis déterminer le niveau de signification du test.

12 TESTS DE COMPARAISON

Il s'agit de comparer les moyennes, écarts-type, voire les fonctions de répartition d'un même caractère sur deux populations distinctes ***P*** et ***Q***.

En général, la fonction de répartition F sur ***P*** et la fonction de répartition G sur ***Q*** sont inconnues, conséquemment les valeurs moyennes notées respectivement m_P et m_Q, les écarts-type σ_P et σ_Q ont des valeurs inconnues.

Disposant d'un échantillon $X_1, \ldots, X_{n_P}$ de taille n_P extrait de ***P*** et d'un échantillon $Y_1, \ldots, Y_{n_Q}$ de taille n_Q extrait de ***Q***, on peut tester à l'aide d'un test unilatéral ou bilatéral, l'hypothèse « $m_P = m_Q$ » d'égalité des moyennes dans les deux populations, l'hypothèse « $\sigma_P = \sigma_Q$ » d'égalité des écart-types, ou enfin l'hypothèse d'égalité des fonctions de répartition « $F = G$ *c'est-à-dire X et Y ont même distribution* ». On peut également tester « $m_P \leqslant m_Q$ », « $\sigma_P \leqslant \sigma_Q$ » ou « $F < G$ ».

Section 1 ■ Tests paramétriques de comparaison de moyennes, d'écart-types, de proportions

Section 2 ■ Tests de comparaison de deux distributions (fonctions de répartition)

Section 3 ■ Tests de comparaison avec Excel et SPSS

© Dunod. La photocopie non autorisée est un délit.

Section 1

TESTS PARAMÉTRIQUES DE COMPARAISON

Les variables statistiques utilisées pour réaliser les tests dépendent évidemment du paramètre qui fait l'objet du test (moyenne ou écart-type).

Lorsque les deux échantillons $X_1,\ldots,X_{n_P}$ et $Y_1,\ldots,Y_{n_Q}$ respectivement extraits de $\boldsymbol{P}$ et $\boldsymbol{Q}$ sont de grande taille, ces statistiques ont généralement une distribution très proche de la loi normale standard. Si l'on dispose d'échantillons de petite taille, le choix de la statistique utilisée présuppose connues les natures des distributions sur $\boldsymbol{P}$ et sur $\boldsymbol{Q}$ (soit par exemple « les distributions sur $\boldsymbol{P}$ et sur $\boldsymbol{Q}$ sont normales »).

1 Tests paramétriques de comparaison de moyennes

Les différents types de test où H_0 désigne l'hypothèse « $m_P = m_Q$ » (ou de façon équivalente « $m_P - m_Q = 0$ ») diffèrent suivant la forme de l'hypothèse alternative :

– **[1]** l'hypothèse alternative est $\bar{H}_0$ « $m_P \neq m_Q$ » équivalente à « $m_P - m_Q \neq 0$ »

– **[2]** l'hypothèse alternative est H_1 « $m_P > m_Q$ » équivalente à « $m_P - m_Q > 0$ » (rentre dans cette catégorie le cas où H_0 est formulée de la façon suivante « $m_P \leqslant m_Q$ »)

– **[3]** l'hypothèse alternative est H_1' « $m_P < m_Q$ » équivalente à « $m_P - m_Q < 0$ » (rentre dans cette catégorie le cas où H_0 est « $m_P \geqslant m_Q$ »).

Disposant de deux échantillons $X_1, X_2,\ldots,X_{n_P}$ et $Y_1, Y_2,\ldots,Y_{n_Q}$ de taille n_P et n_Q respectivement extraits de $\boldsymbol{P}$ et $\boldsymbol{Q}$, on sait que $\bar{X} = \frac{1}{n_P}\sum_{i=1}^{n_P} X_i$ est un estimateur de m_P, $\bar{Y} = \frac{1}{n_Q}\sum_{i=1}^{n_Q} Y_i$ est un estimateur de m_Q et donc $(\bar{X} - \bar{Y})$ est un estimateur de $m_P - m_Q$.

La variance de $(\bar{X} - \bar{Y})$ est égale à $\left(\frac{\sigma_P^2}{n_P} + \frac{\sigma_Q^2}{n_Q}\right)$ et peut être estimée par $\frac{S_X^2}{n_P} + \frac{S_Y^2}{n_Q}$ car $S_X^2 = \frac{1}{n_P - 1}\sum_{i=1}^{n_P}(X_i - \bar{X})^2$ et $S_Y^2 = \frac{1}{n_Q - 1}\sum_{i=1}^{n_Q}(Y_i - \bar{Y})^2$ sont des estimateurs de σ_P^2 et σ_Q^2.

1.1 Test à partir de deux échantillons de grande taille

Variable statistique utilisée. Les deux échantillons $X_1, X_2, \ldots, X_{n_P}$ et $Y_1, Y_2, \ldots, Y_{n_Q}$ respectivement extraits de $\boldsymbol{P}$ et $\boldsymbol{Q}$ étant de grande taille et les valeurs des écarts-types σ_P et σ_Q étant inconnues, on utilise la propriété :

$$\text{« } T = \frac{\bar{X} - \bar{Y} - (m_P - m_Q)}{\sqrt{\frac{S_X^2}{n_P} + \frac{S_Y^2}{n_Q}}} \cong \mathcal{N}_{0;1} \text{ »}$$

Aussi, si l'hypothèse H_0 « $m_P - m_Q = 0$ » est vraie, la *v.a.*

$T_0 = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_P} + \frac{S_Y^2}{n_Q}}}$ suit sensiblement la loi normale standard : $T_0 \cong \mathcal{N}_{0;1}$.

Après expérimentation, on connaît les valeurs numériques $\bar{x}$ et s_x^2 prises par $\bar{X}$ et S_X^2 ainsi que les valeurs $\bar{y}$ et s_y^2 prises par $\bar{Y}$ et S_Y^2 et donc la valeur t_0^* prise après expérimentation par T_0.

Règle de décision selon la forme de l'hypothèse alternative

[1] Test bilatéral avec l'hypothèse alternative $\bar{H}_0$ « $m_P \neq m_Q$ »

En raison de la pseudo symétrie de la loi que suit T_0, le domaine de rejet δ_α de H_0 est de type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$ où $c_{\alpha/2}$ est défini par $\Phi(c_{\alpha/2}) \cong 1 - \alpha/2$.

Si la valeur t_0^* prise par T_0 n'est pas comprise entre $-c_{\alpha/2}$ et $c_{\alpha/2}$, on décide de rejeter l'hypothèse H_0 « $m_P = m_Q$ » au profit de l'hypothèse $\bar{H}_0$ (décision notée D_1). Dans le cas contraire, avec le niveau de signification retenu α, on conclut que l'on ne peut pas rejeter H_0 (décision notée D_0).

Le niveau de signification observé α_c est tel que $\alpha_c = 2 \times P_0(T_0 \geqslant |t_0^*|) = 2 \times P(\mathcal{N}_{0;1} \geqslant |t_0^*|)$.

Justification du choix de δ_α domaine de rejet de H_0. Sous l'hypothèse H_0, $\bar{X} - \bar{Y} \xrightarrow[n_P \text{ et } n_Q \to \infty]{\text{pr}} (m_P - m_Q) = 0$, aussi la variable statistique T_0 tend à prendre des valeurs positives ou négatives mais proches de 0. *A contrario*, si la valeur t_0^* prise par T_0 est éloignée de 0, on décide de rejeter H_0 au profit de $\bar{H}_0$. Le domaine de rejet δ_α de H_0 est de type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, +\infty[$.

[2] Test unilatéral avec l'hypothèse alternative H_1 « $m_P > m_Q$ »

Le domaine de rejet δ_α de H_0 est du type $[c_\alpha, \infty[$. Ayant choisi un niveau de signification α, on cherche le nombre c_α défini par $\alpha = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0;1} \geqslant c_\alpha)$ soit $\Phi(c_\alpha) \cong 1 - \alpha$.

© Dunod. La photocopie non autorisée est un délit.

On prend la décision D_1 de rejeter H_0 au profit de H_1 lorsque la valeur numérique t_0^* prise par T_0 est supérieure à c_α.

Justification du choix de δ_α domaine de rejet de H_0. Si l'hypothèse alternative H_1 « $m_P > m_Q$ » est vraie : $\bar{X} - \bar{Y} \xrightarrow[n_P \text{ et } n_Q \to \infty]{\text{pr}} (m_P - m_Q) > 0$. Donc, T_0 tend à prendre de grandes valeurs positives.

Le niveau de signification observé est $\alpha_c = P(T_0 \geqslant t_0^*) \cong P(\mathcal{N}_{0;1} \geqslant t_0^*)$.

[3] Test unilatéral avec l'hypothèse alternative H_1' « $m_P < m_Q$ »

Le domaine de rejet δ_α de H_0 est de type $]-\infty, c'_\alpha]$. Ayant choisi un niveau de signification α, on cherche le nombre c'_α défini par $\alpha = P_0(T_0 \leqslant c'_\alpha) \cong P(\mathcal{N}_{0;1} \leqslant c'_\alpha) = \Phi(c'_\alpha)$.

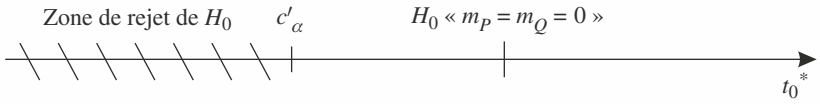


Figure 12.1

On prend la décision D_1 de rejeter H_0 au profit de H_1' lorsque $t_0^* \leqslant c'_\alpha$. Le niveau de signification observé $\alpha_c = P(T_0 \leqslant t_0^*) \cong P(\mathcal{N}_{0;1} \leqslant t_0^*)$.

Exemple

Un analyste financier souhaite comparer les rentabilités des entreprises situées dans deux secteurs d'activités distincts ***P*** et ***Q***, rentabilité évaluée par le ratio « *bénéfice/total de l'actif* ». À cette fin il dispose de deux échantillons :

– un échantillon de 40 entreprises issues du secteur ***P*** dont les ratios $x_1, x_2, \ldots, x_{40}$ ont pour valeur moyenne $\bar{x} = 0{,}025$ et pour écart-standard $s_x = 0{,}05$;

– un échantillon de 50 entreprises issues du secteur ***Q*** dont les ratios $y_1, y_2, \ldots, y_{50}$ ont pour valeur moyenne $\bar{y} = 0{,}045$ et pour écart-standard $s_y = 0{,}06$.

Pensant que la moyenne m_P des ratios dans ***P*** pourrait être inférieure ou égale à la moyenne m_Q des ratios dans ***Q***, il souhaite tester, avec un niveau de signification de 5 %, l'hypothèse de base H_0 « $m_P = m_Q$ » contre H_1' « $m_P < m_Q$ » (cas **[3]**).

Variable statistique utilisée. Les deux échantillons pouvant être considérés de grande taille puisque $n_P = 40$ et $n_Q = 50$ on utilise la *v.a.* $T_0 = (\bar{X} - \bar{Y}) / \sqrt{\frac{S_X^2}{n_P} + \frac{S_Y^2}{n_Q}}$ qui, sous H_0, suit sensiblement la loi normale standard : $T_0 \cong \mathcal{N}_{0;1}$.

Domaine de rejet de H_0 et règle de décision. Le domaine de rejet de H_0 est du type $]-\infty, c'_\alpha]$ où c'_α est défini par $\Phi(c'_\alpha) = \alpha$. Le niveau de signification α étant de 0,05 on constate que $c'_\alpha = -1.64$ et donc le domaine de rejet δ_α de H_0 est l'intervalle $]-\infty, -1{,}64]$.

La valeur t_0^* prise par T_0 est égale à $\dfrac{\bar{x}-\bar{y}}{\sqrt{\frac{S_x^2}{n_P}+\frac{S_y^2}{n_Q}}}$ soit $t_0^* = \dfrac{0{,}025-0{,}045}{\sqrt{\frac{0{,}05^2}{40}+\frac{0{,}06^2}{50}}} = -1{,}72$;

constatant que t_0^* est inférieure à $c'_\alpha = -1{,}64$ on décide de rejeter l'hypothèse H_0 au profit de H'_1 « $m_P < m_Q$ ».

REPÈRES : Généralisation

Si l'on dispose de deux grands échantillons issus respectivement de ***P*** et ***Q***, on peut vouloir tester $H_0(\omega)$ « $m_Q = \omega m_P$ où ω est une constante positive donnée » contre $H_1(\omega)$ « $m_Q < \omega m_P$ ». Pour cela on utilise la propriété suivante :

Si $H_0(\omega)$ est vraie, $T_0(\omega) = \dfrac{\omega\bar{X}-\bar{Y}}{\sqrt{\frac{\omega^2 S_X^2}{n_P}+\frac{S_Y^2}{n_Q}}} \cong \mathcal{N}_{0;1}$

On rejette $H_0(\omega)$ au profit de $H_1(\omega)$ « $\omega m_P > m_Q$ » avec un niveau de signification α lorsque la valeur prise par $T_0(\omega)$ vérifie $t^*_{0(\omega)} \geqslant c_\alpha$ où c_α est défini par $\Phi(c_\alpha) \cong 1-\alpha$. Inversement, pour tester $H_0(\omega)$ contre $H'_1(\omega)$ « $\omega m_P < m_Q$ » on rejette $H_0(\omega)$ au profit de $H'_1(\omega)$ si $t^*_{0(\omega)} \leqslant c'_\alpha$ où $\Phi(c'_\alpha) = \alpha$.

Remarque. Si n_P et n_Q sont grands et si les écarts-types σ_P et σ_Q ont des valeurs connues on utilise la statistique $T'_0 = \dfrac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_P^2}{n_P}+\frac{\sigma_Q^2}{n_Q}}}$. Sous l'hypothèse H_0, la statistique T'_0 suit sensiblement la loi normale standard : « $T'_0 \cong \mathcal{N}_{0;1}$ ». Les règles de décisions sont identiques à celles présentées pour les cas [1], [2], [3].

Cette dernière statistique s'utilise également avec des échantillons de *taille quelconque* issus de distributions « normales » lorsque les valeurs des écarts-type σ_P et σ_Q sont connues, autrement dit lorsque $X_1,\ldots,X_{n_P}$ est un échantillon *iid* d'une loi normale $\mathcal{N}(m_P,\sigma_P^2)$ et $Y_1,\ldots,Y_{n_Q}$ un échantillon *iid* d'une loi normale $\mathcal{N}(m_Q,\sigma_Q^2)$ avec σ_P et σ_Q connus. En effet on a alors « $T'_0 = \mathcal{N}_{0;1} \quad \forall n_P$ et n_Q ».

1.2 Test à partir de deux échantillons dont les distributions sont supposées être « normales » et avoir même écart-type : « $\sigma_P = \sigma_Q$ »

On dispose d'un échantillon $X_1,\ldots,X_{n_P}$ *iid* d'une loi normale $\mathcal{N}(m_P,\sigma_P^2)$ et d'un autre échantillon $Y_1,\ldots,Y_{n_Q}$ *iid* d'une autre loi normale $\mathcal{N}(m_Q,\sigma_Q^2)$, les valeurs de m_P,σ_P, m_Q et σ_Q étant inconnues (mais n_p et n_q éventuellement petits).

© Dunod. La photocopie non autorisée est un délit.

Variable statistique utilisée. Si on est sûr que $\sigma_Q = \sigma_P$ on utilise la propriété suivante « $T = \dfrac{(\bar{X} - \bar{Y}) - (m_P - m_Q)}{\sqrt{\dfrac{(n_P - 1)S_x^2 + (n_Q - 1)S_y^2}{n_P + n_Q - 2}} \times \sqrt{\dfrac{1}{n_P} + \dfrac{1}{n_Q}}} = \boldsymbol{t}_{n_P+n_Q-2}$ (variable de Student) ».

Sous l'hypothèse H_0 « $m_P - m_Q = 0$ » la variable statistique utilisée

$$T_0 = \frac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{(n_P - 1)S_x^2 + (n_Q - 1)S_y^2}{n_P + n_Q - 2}} \times \sqrt{\dfrac{1}{n_P} + \dfrac{1}{n_Q}}} = \boldsymbol{t}_{n_P+n_Q-2}$$

Après expérimentation on connaît les valeurs numériques $\bar{x}$ et s_x^2 prises par $\bar{X}$ et S_X^2 ainsi que les valeurs $\bar{y}$ et s_y^2 prises par $\bar{Y}$ et S_Y^2, aussi T_0 prend-il la valeur

$$t_0^* = \frac{(\bar{x} - \bar{y})}{\sqrt{\dfrac{(n_P - 1)s_x^2 + (n_Q - 1)s_y^2}{n_P + n_Q - 2}} \times \sqrt{\dfrac{1}{n_P} + \dfrac{1}{n_Q}}}$$

Règle de décision selon la forme de l'hypothèse alternative.

Même raisonnement et même règle de décision que dans le § 1.1, la variable $\boldsymbol{t}_{np+nq-2}$ se substituant à $\mathcal{N}_{0;1}$

Exemple

Un DRH souhaite tester l'hypothèse selon laquelle la rémunération moyenne des hommes (population $\boldsymbol{P}$) et des femmes (population $\boldsymbol{Q}$) d'une grande entreprise est la même. Prélevant au hasard douze hommes ($n_P = 12$) et dix femmes ($n_Q = 10$), il observe sur l'échantillon des hommes une rémunération moyenne $\bar{x} = 3\,000$ unités monétaires (*u.m.*) et un écart-type standard des rémunérations $s_x = 520$ et sur l'échantillon des femmes une rémunération moyenne $\bar{y} = 2\,000$ *u.m.* et un écart-type standard $s_y = 500$. La distribution des rémunérations masculines et féminines sont supposées sensiblement normales. Il est également admis que les deux distributions de rémunérations ont un même écart-type (*cf.* p. 208). On souhaite tester l'hypothèse H_0 « $m_P = m_Q$ » contre l'hypothèse alternative H_1 « $m_P > m_Q$ », cas [2] avec un niveau de signification $\alpha = 0{,}1$.

Variable statistique utilisée. Étant admis que $\sigma_Q = \sigma_P$ on sait que, sous l'hypothèse H_0 « $m_P - m_Q = 0$ », la variable statistique T_0 suit la loi de Student $\mathcal{S}t(20)$:

$$T_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{(n_P - 1)S_x^2 + (n_Q - 1)S_y^2}{n_P + n_Q - 2}} \times \sqrt{\dfrac{1}{n_P} + \dfrac{1}{n_Q}}} = \boldsymbol{t}_{12+10-2} = \boldsymbol{t}_{20}$$

Le domaine de rejet de H_0 est de type $[c_\alpha,\infty[$ où c_α est défini par $P(\boldsymbol{t}_{20} \geqslant c_\alpha) = \alpha = 0{,}1$ soit $c_\alpha = 1{,}33$.

Or $t_0^* = \dfrac{3\,000 - 2\,000}{\sqrt{\dfrac{(12-1)520^2 + (10-1)500^2}{12+10-2}} \times \sqrt{\dfrac{1}{12} + \dfrac{1}{10}}} = 4{,}56 > 1{,}33$ on rejette donc H_0.

En fait il aurait fallu au préalable s'assurer de la normalité des distributions des rémunérations, celles-ci étant fréquemment ajustées par des lois gamma.

1.3 Test à partir de deux échantillons issus de distributions « normales » et dont les écarts-types inconnus sont inégaux

Dans le cas où les variances sont inconnues et inégales on peut utiliser le test d'Aspin-Welch. Sous l'hypothèse H_0 « $m_P - m_Q = 0$ » la variable statistique utilisée suit sensiblement une loi de Student à ν degrés de liberté :

$$T_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{S_X^2}{n_P} + \dfrac{S_Y^2}{n_Q}}} \cong \boldsymbol{t}_\nu \quad \text{où} \quad \nu = \frac{(s_x^2/n_P + s_y^2/n_Q)^2}{(s_x^2/n_P)^2/(n_P-1) + (s_y^2/n_Q)^2/(n_Q-1)} \text{ est}$$

généralement arrondi à l'entier inférieur. Les règles de décisions sont les mêmes que précédemment.

2 Tests de comparaison d'écarts-types

Disposant de deux échantillons $X_1, X_2, \ldots, X_{n_P}$ et $Y_1, Y_2, \ldots, Y_{n_Q}$ *iid* respectivement extraits de $\boldsymbol{P}$ et $\boldsymbol{Q}$, on souhaite tester l'hypothèse d'égalité des écart-types des deux populations sachant que $S_X^2 = \dfrac{1}{n_P - 1}\sum_{i=1}^{n_P}(X_i - \bar{X})^2$ est un estimateur de σ_P^2 et $S_Y^2 = \dfrac{1}{n_Q - 1}\sum_{i=1}^{n_Q}(Y_i - \bar{Y})^2$ est un estimateur de σ_Q^2.

Les différents tests où H_0 désigne l'hypothèse « $\sigma_P = \sigma_Q$ » (ou de façon équivalente « $\sigma_P - \sigma_Q = 0$ ») diffèrent suivant la forme prise par l'hypothèse alternative :

[1] l'hypothèse alternative est $\bar{H}_0$ « $\sigma_P \neq \sigma_Q$ »

[2] l'hypothèse alternative est H_1 « $\sigma_P > \sigma_Q$ »
(ce test inclut le cas où H_0 est formulée de la façon suivante « $\sigma_P \leqslant \sigma_Q$ »)

[3] l'hypothèse alternative est H_1' « $\sigma_P < \sigma_Q$ »
(ce test inclut le cas où H_0 est formulée de la façon suivante « $\sigma_P \geqslant \sigma_Q$ »)

© Dunod. La photocopie non autorisée est un délit.

2.1 Test à partir de deux échantillons de grande taille

Les tests précédents peuvent être reformulés de la manière suivante :

[1]	[2]	[3]
H_0 « $\sigma_P^2 - \sigma_Q^2 = 0$ » contre $\bar{H}_0$ « $\sigma_P^2 - \sigma_Q^2 \neq 0$ »	H_0 « $\sigma_P^2 - \sigma_Q^2 = 0$ » contre H_1 « $\sigma_P^2 - \sigma_Q^2 > 0$ »	H_0 « $\sigma_P^2 - \sigma_Q^2 = 0$ » contre H_1' « $\sigma_P^2 - \sigma_Q^2 < 0$ »

La *v.a.* $(S_X^2 - S_Y^2)$ est un estimateur de $(\sigma_P^2 - \sigma_Q^2)$.

Variable statistique utilisée. Les échantillons extraits de $\boldsymbol{P}$ et $\boldsymbol{Q}$ étant de grande taille ($n_P \geqslant 30$ et $n_Q \geqslant 30$), on utilise la propriété suivante :

$$\text{« } T = \frac{S_X^2 - S_Y^2 - (\sigma_P^2 - \sigma_Q^2)}{\sqrt{\frac{\hat{\mu}_{4,X} - S_X^4}{n_P} + \frac{\mu_{4,Y} - S_Y^4}{n_Q}}} \cong \mathcal{N}_{0;1} \text{ où } \hat{\mu}_{4,X} = \frac{1}{n_P}\sum_{i=1}^{n_P}(X_i - \bar{X})^4,\ \hat{\mu}_{4,Y} = \frac{1}{n_Q}\sum_{i=1}^{n_Q}(Y_i - \bar{Y})^4 \text{ »}$$

Sous l'hypothèse H_0, la statistique utilisée $T_0 = \dfrac{S_X^2 - S_Y^2}{\sqrt{\frac{\hat{\mu}_{4,X} - S_X^4}{n_P} + \frac{\hat{\mu}_{4,Y} - S_Y^4}{n_Q}}} \cong \mathcal{N}_{0;1}$.

Règle de décision selon la forme de l'hypothèse alternative

[1] Test bilatéral avec l'hypothèse alternative H_0 « $\sigma_P^2 - \sigma_Q^2 \neq 0$ »

En raison de la pseudo symétrie de la loi que suit T_0, le domaine de rejet de H_0 est du type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$. Pour un niveau de signification α, $c_{\alpha/2}$ est défini par $\Phi(c_{\alpha/2}) = 1 - \alpha/2$. On rejette H_0 lorsque $t_0^* \leqslant -c_{\alpha/2}$ ou $t_0^* \geqslant +c_{\alpha/2}$.

Justification du choix de δ_α domaine de rejet de H_0. Si H_0 est vraie la variable utilisée T_0 tend à prendre des valeurs positives ou négatives mais proches de 0 car $(S_X^2 - S_Y^2) \xrightarrow{\text{pr}} \sigma_P^2 - \sigma_Q^2 = 0$. *A contrario*, si la valeur t_0^* prise par T_0 est éloignée de 0 on décide de rejeter H_0 au profit de $\bar{H}_0$.

Le niveau de signification observé $\alpha_c = 2 \times P_0(T_0 \geqslant |t_0^*|) = 2 \times P(\mathcal{N}_{0;1} \geqslant |t_0^*|)$

[2] Test unilatéral avec l'hypothèse alternative H_1 « $\sigma_P^2 > \sigma_Q^2$ »

Le domaine de rejet δ_α de H_0 est du type $[c_\alpha, \infty[$. Ayant choisi un niveau de signification α (par exemple ($\alpha = 5\,\%$), on cherche c_α défini par $\alpha = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0;1} \geqslant c_\alpha)$ et donc $\Phi(c_\alpha) \cong 1 - \alpha$. Lorsque, après analyse des échantillons, on constate que la valeur t_0^* prise par T_0 est supérieure à c_α, on rejette l'hypothèse H_0 au profit de l'hypothèse alternative H_1.

Le niveau de signification observé est $\alpha_c = P(T_0 \geqslant t_0^*) \cong P_0(\mathcal{N}_{0;1} \geqslant t_0^*)$.

Justification du choix de δ_α domaine de rejet de H_0. Si H_0 est fausse et donc H_1 « $\sigma_P^2 - \sigma_Q^2 > 0$ » est vraie, la variable utilisée T_0 tend à prendre de grandes valeurs positives car $(S_X^2 - S_Y^2) \xrightarrow{\text{pr}} \sigma_P^2 - \sigma_Q^2 > 0$ lorsque n_P et n_Q tendent vers l'infini.

[3] Test unilatéral avec l'hypothèse alternative H_1' « $\sigma_P^2 < \sigma_Q^2$ »

Le domaine de rejet de H_0 est du type $]-\infty, c'_\alpha]$ où est défini par
$\alpha = P_0(T_0 \leqslant c'_\alpha) \cong P(\mathcal{N}_{0;1} \leqslant c'_\alpha) = \Phi(c'_\alpha)$.
On rejette H_0 au profit de H_1 lorsque $t_0^* \leqslant c'_\alpha$.
Le niveau de signification observé $\alpha_c = P(T_0 \leqslant t_0^*) \cong P(\mathcal{N}_{0;1} \leqslant t_0^*)$.

Exemple

Un analyste financier d'une banque souhaite savoir si la dispersion de la taille des firmes est la même dans les secteurs d'activités distincts ***P*** et ***Q***. La taille des firmes étant appréciée par leur effectif, il dispose d'un échantillon de 60 firmes issues de ***P*** ($n_P = 60$) et observe leur effectif respectif $x_1, x_2, \ldots, x_{60}$ puis en déduit $s_x = 10$ et $\mu_{4,x} = \dfrac{1}{n_P}\sum_{i=1}^{n_P}(x_i - \bar{x})^4 = 20\,000$. Il prélève un échantillon de 70 firmes dans ***Q*** ($n_Q = 70$) et observant leur effectif y_i constate que $s_y = 20$ et $\mu_{4,y} = \dfrac{1}{n_Q}\sum_{i=1}^{n_Q}(y_i - \bar{y})^4 = 200\,000$. Il décide de tester, avec un niveau de signification de 5 %, l'hypothèse H_0 « $\sigma_P = \sigma_Q$ » contre $\bar{H}_0$ « $\sigma_P \neq \sigma_Q$ ».

Variable statistique utilisée. Sous l'hypothèse H_0, la statistique utilisée T_0 suit sensiblement la loi normale standard : $T_0 = \dfrac{S_X^2 - S_Y^2}{\sqrt{\dfrac{\hat{\mu}_{4,X} - S_X^4}{n_P} + \dfrac{\hat{\mu}_{4,Y} - S_Y^4}{n_Q}}} \cong \mathcal{N}_{0;1}$.

Le domaine de rejet de H_0 est de type [1] c'est-à-dire $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$ où $c_{\alpha/2} = 1{,}96$ car solution de $\Phi(c_{\alpha/2}) = 1 - \alpha/2 = 0{,}975$.

La valeur $t_0^* = \dfrac{10^2 - 20^2}{\sqrt{(20\,000 - 10^4)/60 + (200\,000 - 20^4)/70}} = -11{,}0$ prise par T_0 n'étant pas comprise entre $-1{,}96$ et $1{,}96$ on rejette H_0. Le niveau de signification observé $\alpha_c = 2P_0(T_0 \geqslant |-11{,}0|) \cong 2P(\mathcal{N}_{0;1} \geqslant 11{,}0) \cong 0$ donc avec un risque d'erreur négligeable on peut affirmer que H_0 est fausse.

2.2 Test avec deux échantillons dont les distributions sont « normales »

Dans ce cas, l'hypothèse de base H_0 « $\sigma_P = \sigma_Q$ » et les hypothèses alternatives $\bar{H}_0$ « $\sigma_P \neq \sigma_Q$ », H_1 « $\sigma_P > \sigma_Q$ », H_1' « $\sigma_P < \sigma_Q$ » peuvent être réécrites de la façon suivante :

© Dunod. La photocopie non autorisée est un délit.

[1]	[2]	[3]
H_0 « $\sigma_P^2/\sigma_Q^2 = 1$ » contre $\bar{H}_0$ « $\sigma_P^2/\sigma_Q^2 \neq 1$ »	H_0 « $\sigma_P^2/\sigma_Q^2 = 1$ » contre H_1 « $\sigma_P^2/\sigma_Q^2 > 1$ »	H_0 « $\sigma_P^2/\sigma_Q^2 = 1$ » contre H_1' « $\sigma_P^2/\sigma_Q^2 < 1$ »

Variable statistique utilisée. Si la distribution en X et la distribution en Y dont est issu chacun des deux échantillons peuvent être considérées comme normales, on utilise la propriété suivante qui est valable quelque soient les tailles des deux échantillons et en particulier lorsque n_P ou n_Q sont petits :

« $T = \dfrac{S_X^2}{S_Y^2} \times \dfrac{\sigma_Q^2}{\sigma_P^2}$ suit la loi de Fisher-Snédécor $F(n_P - 1, n_Q - 1)$ ».

Sous l'hypothèse H_0 , on constate que $T_0 = \dfrac{S_X^2}{S_Y^2} = F_{n_Q-1}^{n_P-1}$.

Après expérimentation T_0 prend la valeur $t_0^* = s_x^2/s_y^2$.

Règle de décision selon la forme prise par l'hypothèse alternative

Selon que t_0^* appartient où non à la zone de rejet de l'hypothèse H_0, on décide de rejeter ou non H_0.

[1] Test bilatéral avec l'hypothèse alternative H_0 « $\sigma_P^2/\sigma_Q^2 \neq 1$ »

Le domaine de rejet de H_0 est du type $[0, c_{1,\alpha/2}] \cup [c_{2,\alpha/2}, \infty[$.

$c_{1,\alpha/2}$ et $c_{2,\alpha/2}$ sont respectivement définis par les conditions $P(F_{n_Q-1}^{n_P-1} \leqslant c_{1,\alpha/2}) = \alpha/2$ et $P(F_{n_Q-1}^{n_P-1} \geqslant c_{2,\alpha/2}) = \alpha/2$.

Le niveau de signification observé $\alpha_c = 2 \text{ Min } [P(F_{n_Q-1}^{n_P-1} \leqslant t_0^*) ; P(F_{n_Q-1}^{n_P-1} \geqslant t_0^*)]$.

[2] Test unilatéral avec l'hypothèse alternative H_1 « $\sigma_P^2 > \sigma_Q^2$ »

Le domaine de rejet de H_0 est donc du type $[c_\alpha, \infty[$ où c_α est déterminé par la condition : $\alpha = P_0(T_0 \geqslant c_\alpha) = P(F_{n_Q-1}^{n_P-1} \geqslant c_\alpha)$.

En effet : sous H_0, $T \xrightarrow{pr} 1$, sous H_1, $T \xrightarrow{pr} (\sigma_P/\sigma_Q)^2 > 1$.

Le niveau de signification observé $\alpha_c = P_0(T_0 \geqslant t_0^*) = P(F_{n_Q-1}^{n_P-1} \geqslant t_0^*)$.

[3] Test unilatéral avec l'hypothèse alternative H_1' « $\sigma_P^2 < \sigma_Q^2$ »

La variable T_0 ne pouvant prendre que des valeurs positives, le domaine de rejet de H_0 est donc du type $[0, c'_\alpha]$ où c'_α est défini par $\alpha = P_0(T_0 \leqslant c'_\alpha) = P(F_{n_Q-1}^{n_P-1} \leqslant c'_\alpha)$

Exemple

Reprenons l'exemple du § 1.2, avec sur l'échantillon des hommes un écart-type standard des rémunérations $s_x = 520$ et sur l'échantillon des femmes un écart-type standard

$s_y = 500$. Les distributions des rémunérations masculines et féminines étant supposées normales, on souhaite tester H_0 « $\sigma_P = \sigma_Q$ » contre H_1 « $\sigma_P > \sigma_Q$ ».

Sous H_0, $T_0 = S_X^2/S_Y^2 = F_9^{11}$. Sous H_1, $T_0 = S_X^2/S_Y^2 \xrightarrow{\text{pr}} \sigma_P^2/\sigma_Q^2 > 1$, donc le domaine de rejet de H_0 est du type $[c_\alpha, \infty[$ où c_α est déterminé par la condition : $\alpha = 0{,}05 = P_0(T_0 \geqslant c_\alpha) = P(F_9^{11} \geqslant c_\alpha)$. On trouve $c_\alpha = 3{,}10$ et on constate que $t_0^* = s_x^2/s_y^2 = 520^2/500^2 = 1{,}08$ appartient au domaine d'acceptation de H_0. Le niveau de signification observé $\alpha_c = P(T_0 \geqslant t_0^*) \cong P(F_9^{11} \geqslant 1{,}08) = 0.46$ est d'ailleurs relativement élevé.

REPÈRES : Généralisation

On peut vouloir tester $H_0(\omega)$ « $\sigma_Q = \omega\sigma_P$ où ω est une constante donnée » contre $H_1(\omega)$ « $\sigma_Q < \omega\sigma_P$ ». Pour cela on utilise la variable $T_0(\omega) = \dfrac{\omega^2 S_X^2}{S_Y^2}$ qui sous H_0 suit la loi de Fisher-Snédécor $F(n_P - 1; n_Q - 1)$: $T_0(\omega) = F_{n_Q-1}^{n_P-1}$.

On rejette $H_0(\omega)$ au profit de $H_1(\omega)$ avec un niveau de signification α lorsque la valeur $t_{0(\omega)}^*$ prise par $T_0(\omega)$ vérifie $t_{0(\omega)}^* \geqslant c_\alpha$. Inversement, pour tester $H_0(\omega)$ contre $H_1'(\omega)$ « $\sigma_Q > \omega\sigma_P$ » on rejette $H_0(\omega)$ au profit de $H_1'(\omega)$ si $t_{0(\omega)}^* \leqslant c'_\alpha$.

3 Tests de comparaison de proportions

Il s'agit de comparer la proportion p_P d'éléments de $\boldsymbol{P}$ qui possèdent le caractère étudié à la proportion p_Q d'éléments de $\boldsymbol{Q}$ qui possèdent ce même caractère. À cette fin, on dispose

– d'un grand échantillon non exhaustif E_1 de taille n_P extrait de $\boldsymbol{P}$ ($n_P \geqslant 30$), dans lequel une proportion aléatoire F_1 possèdent le caractère considéré : $F_1 = N_1/n_P$, (N_1 étant le nombre d'éléments de E_1 ayant le caractère considéré)

– d'un grand échantillon non exhaustif E_2 de taille n_Q extrait de $\boldsymbol{Q}$ où ($n_Q \geqslant 30$), où $F_2 = N_2/n_Q$ est la proportion aléatoire d'individus ayant le caractère considéré (N_2 désignant le nombre d'éléments de E_2 qui possèdent le caractère considéré).

Les différents tests où H_0 est l'hypothèse « $p_P = p_Q$ » (ou de façon équivalente « $p_P - p_Q = 0$ ») diffèrent suivant la forme prise par l'hypothèse alternative. Dans le cas :

– **[1]** l'hypothèse alternative est $\bar{H}_0$ « $p_P \neq p_Q$ » et donc « $p_P - p_Q \neq 0$ »

– **[2]** l'hypothèse alternative est H_1 « $p_P > p_Q$ » c'est-à-dire « $p_P - p_Q > 0$ »

– **[3]** l'hypothèse alternative est H_1' « $p_P < p_Q$ » c'est-à-dire « $p_P - p_Q < 0$ »

$F_1 = N_1/n_P$ étant un estimateur de p_P et $F_2 = N_2/n_Q$ un estimateur de p_Q, on en déduit que $F_1 - F_2$ est un estimateur de $p_P - p_Q$.

© Dunod. La photocopie non autorisée est un délit.

Variable statistique utilisée. Si l'hypothèse H_0 « $p_P = p_Q$ (= *p valeur commune*) » est vraie, la *v.a.* $\bar{F} = \dfrac{n_P F_1 + n_Q F_2}{n_P + n_Q}$ est un estimateur convergent sans biais de p et on utilise la propriété « $T_0 = \dfrac{F_1 - F_2}{\sqrt{\bar{F}(1-\bar{F})} \times \sqrt{\frac{1}{n_P} + \frac{1}{n_Q}}} \cong \mathcal{N}_{0;1}$ lorsque n_P et n_Q sont grands ».

Règle de décision selon la forme de l'hypothèse alternative

Même règle de décision que dans le § 1.

[1] Test bilatéral avec l'hypothèse alternative $\bar{H}_0$ « $p_P \neq p_Q$ ».

Ayant choisi un niveau de signification α, le domaine de rejet de H_0 est du type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$ où $c_{\alpha/2}$ est défini par la condition $\Phi(c_{\alpha/2}) = 1 - \alpha/2$.

[2] Test unilatéral avec l'hypothèse alternative H_1 « $p_P > p_Q$ ».

Le domaine de rejet de H_0 est de type $[c_\alpha, \infty[$. Ayant choisi un niveau de signification α, c_α est déterminé par la condition $\alpha = P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0,1} \geqslant c_\alpha) = \alpha$, c'est-à-dire $\Phi(c_\alpha) = 1 - \alpha$.

[3] Test unilatéral avec l'hypothèse alternative H_1' « $p_P < p_Q$ ».

Le domaine de rejet de H_0 est donc du type $]-\infty, c'_\alpha]$ où c'_α est déterminé par la condition $\alpha = P_0(T_0 \leqslant c'_\alpha) \cong P(\mathcal{N}_{0;1} \leqslant c'_\alpha) = \Phi(c'_\alpha)$.

Remarque. Les populations ***P*** et ***Q*** étant binomiales on sait que : $m_P = p_P$, $m_Q = p_Q$, $S_X^2/n_P = \bar{X}(1-\bar{X})/(n_P - 1)$, $S_Y^2/n_Q = \bar{Y}(1-\bar{Y})/(n_Q - 1)$. On a donc la propriété [cf. § 1.1 p. 201] :

« $T = \dfrac{F_1 - F_2 - (p_P - p_Q)}{\sqrt{\dfrac{F_1(1-F_1)}{n_P - 1} + \dfrac{F_2(1-F_2)}{n_Q - 1}}} \cong \mathcal{N}_{0;1}$ lorsque n_P et n_Q grands ».

Si H_0 « $p_P = p_Q$ » est vraie, on constate que

$$T_0' = \frac{F_1 - F_2}{\sqrt{\dfrac{F_1(1-F_1)}{n_P - 1} + \dfrac{F_2(1-F_2)}{n_Q - 1}}} \cong \mathcal{N}_{0;1}$$

Cette statistique T_0' est également utilisée pour tester par H_0 « $p_P \leqslant p_Q$ » contre H_1 « $p_P > p_Q$ » ou H_0 « $p_P \geqslant p_Q$ » contre H_1' « $p_P < p_Q$ ». La valeur $t_0'^*$ prise par T_0' est à confronter à la règle de décision identique à la précédente.

REPÈRES : Généralisation

Si l'on souhaite tester $H_0(\omega)$ « $p_Q = \omega p_P$ *où* ω *est une constante positive donnée* », on utilise selon la même méthodologie, si n_P et n_Q grands,

$$T_0(\omega) = \frac{(\omega F_1 - F_2)}{\sqrt{\dfrac{\omega^2 F_1(1-F_1)}{n_P} + \dfrac{F_2(1-F_2)}{n_Q}}} \cong \mathcal{N}_{0;1}$$

On rejette $H_0(\omega)$ au profit de $H_1(\omega)$ « $\omega p_P > p_Q$ » avec un niveau de signification α lorsque la valeur $t^*_{0(\omega)}$ prise par $T_0(\omega)$ vérifie $t^*_{0(\omega)} \geqslant c_\alpha$. Inversement, pour tester $H_0(\omega)$ contre $H'_1(\omega)$ « $\omega p_P < p_Q$ », on rejette $H_0(\omega)$ au profit de $H'_1(\omega)$ si $t^*_{0(\omega)} \leqslant c_\alpha$.

Exemple

L'entreprise G se fournissant en composants ZF auprès des entreprises $\boldsymbol{P}$ et $\boldsymbol{Q}$ souhaite comparer la fiabilité des composants livrés. Sur un échantillon E_1 de 500 composants livrés par $\boldsymbol{P}$ et prélevés au hasard, cinq composants sont défectueux alors que sur un échantillon E_2 de 400 composants livrés par $\boldsymbol{Q}$ et prélevés au hasard, six sont défectueux. On teste l'hypothèse selon laquelle la proportion p_P de composants défectueux livrés par $\boldsymbol{P}$ est égale à la proportion p_Q de composants défectueux livrés par $\boldsymbol{Q}$, c'est-à-dire H_0 « $p_P = p_Q$ » contre $\bar{H}_0$ « $p_P \neq p_Q$ » avec un niveau de signification de 5 % .

Variable statistique utilisée. Sous l'hypothèse H_0,

$$T_0 = \frac{F_1 - F_2}{\sqrt{\bar{F}(1-\bar{F})} \times \sqrt{\frac{1}{n_P} + \frac{1}{n_Q}}} \cong \mathcal{N}_{0;1} \text{ où } \bar{F} = \frac{n_P F_1 + n_Q F_2}{n_P + n_Q}$$

Règle de décision. Le domaine de rejet de H_0 est du type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$ où $c_{\alpha/2} = 1{,}96$ car défini par la condition $\Phi(c_{\alpha/2}) = 1 - \alpha/2 = 0{,}975$.

Décision. F_1 et F_2 prennent les valeurs $f_1 = 5/500 = 0{,}01$, $f_2 = 6/400 = 0{,}015$. Donc $\bar{F}$ prend la valeur $\bar{f} = (0{,}01 \times 500 + 0{,}015 \times 400)/900 = 11/900$. La variable statistique T_0 prenant la valeur $t_0^* = -0{,}678$ qui appartient au domaine d'acceptation de H_0 puisque comprise entre $-1{,}96$ et $1{,}96$, on décide de ne pas rejeter H_0.

4 Test d'homogénéité de proportions

Soit respectivement $p_1, p_2, \ldots, p_h$ les proportions d'éléments qui possèdent le caractère considéré $\mathcal{C}$ dans les h populations binomiales $\boldsymbol{P_1}, \boldsymbol{P_2}, \ldots, \boldsymbol{P_h}$.

On extrait de chaque population $\boldsymbol{P_i}$ un échantillon E_i de taille ν_i. Le nombre aléatoire N_i d'éléments de cet échantillon qui possèdent ce caractère $\mathcal{C}$ suit la loi binomiale $\mathcal{B}(\nu_i ;\ p_i)$. Il prend la valeur n_i correspondant au nombre observé d'éléments qui, sur l'échantillon E_i, possèdent le caractère étudié.

À partir de la valeur n_i prise par chaque *v.a.* N_i, on se propose de tester :

$$H_0 \text{ « } p_1 = p_2 = \ldots p_h \ (= p \text{ valeur commune inconnue) » contre } \bar{H}_0.$$

© Dunod. La photocopie non autorisée est un délit.

– ***Statistique utilisée***. Notons $\nu = \sum_{i=1}^{h} \nu_i$ l'effectif cumulé des h échantillons, $\hat{p} = \left(\sum_{i=1}^{h} N_i \right) / \nu$ la proportion aléatoire d'éléments qui, sur l'ensemble des échantillons, possèdent le caractère considéré et $p^* = \left(\sum_{i=1}^{h} n_i \right) / \nu$ sa réalisation. Sous l'hypothèse H_0, « $Z = \sum_{i=1}^{h} \frac{(N_i - \nu_i \hat{p})^2}{\nu_i \hat{p}(1-\hat{p})} \cong \chi^2_{h-1}$ si $\nu_i p^* > 5 \quad \forall i$ ».

– ***Règle de décision***. Sous l'hypothèse alternative $\bar{H}_0$, la variable statistique Z tend à prendre des valeurs élevées. Le domaine de rejet de H_0 est donc de type $[c_\alpha, \infty[$ où c_α est tel que $P(\chi^2_{h-1} \geqslant c_\alpha) = \alpha$. On rejette H_0 si la valeur $z = \sum_{i=1}^{h} \frac{(n_i - \nu_i p^*)^2}{\nu_i p^*(1-p^*)}$ prise par Z est supérieure à c_α.

Le niveau de signification observé est $\alpha_c \cong P(\chi^2_{h-1} \geqslant z)$

Exemple

S'intéressant à la diffusion d'une innovation fiscale au sein des entreprises, on range les entreprises en trois catégories : celles de moins de dix salariés (population P_1), celles ayant entre dix et cinq cents salariés (population P_2) et celles de plus de cinq cents salariés (population P_3). On extrait de la population P_1 un échantillon E_1 de taille $\nu_1 = 40$, de P_2 un échantillon E_2 de taille $\nu_2 = 20$, de P_3 un échantillon E_3 de taille $\nu_3 = 14$ et on constate respectivement sur chacun de ces échantillons le nombre $n_1 = 20$, $n_2 = 13$, $n_3 = 12$ d'entreprises qui ont adopté l'innovation. Pour réaliser le test d'homogénéité H_0 « $p_1 = p_2 = p_3$ » contre $\bar{H}_0$ avec un niveau de signification $\alpha = 0{,}05$ on doit préalablement considérer la proportion d'entreprises p^* qui, sur l'ensemble des échantillons, possèdent le caractère considéré : $p^* = (20 + 13 + 12)/(40 + 20 + 14) \cong 0{,}6$.

Sous l'hypothèse H_0, $Z = \sum_{i=1}^{3} \frac{(N_i - \nu_i \hat{p})^2}{\nu_i \hat{p}(1-\hat{p})} \cong \chi^2_{3-1}$ puisque $\nu_i p^* > 5 \quad \forall i = 1,2,3$.

Le domaine de rejet de H_0 est de type $[c_\alpha, \infty[$ où c_α est tel que $P(\chi^2_2 \geqslant c_\alpha) = \alpha = 0{,}05$ d'où $c_\alpha = 5{,}99$. La valeur $z = (20 - 40 \times 0{,}6)^2/(40 \times 0{,}6 \times 0{,}4) + (13 - 20 \times 0{,}6)^2 /(20 \times 0{,}6 \times 0{,}4) + (12 - 14 \times 0{,}6)^2/(14 \times 0{,}6 \times 0{,}4) = 5{,}73$ appartenant à l'intervalle $[5{,}99, \infty[$ on ne rejette pas H_0.

Section 2

TESTS DE COMPARAISON DE DEUX DISTRIBUTIONS

Notant F la fonction de répartition sur la population $\boldsymbol{P}$ et G la fonction de répartition sur $\boldsymbol{Q}$, on teste H_0 « $\boldsymbol{P}$ et $\boldsymbol{Q}$ ont même distribution soit $F(t) = G(t) \quad \forall t$ ».

1 Test du khi-deux

Partant d'un échantillon $X_1,\ldots,X_n$ de taille n extrait de $\boldsymbol{P}$ et d'un échantillon $Y_1,\ldots,Y_{n'}$ de taille n' extrait de $\boldsymbol{Q}$, on teste H_0 « *X et Y suivent la même loi de probabilité* » contre $\bar{H}_0$ « *X et Y ont des distributions différentes* ».

La variable utilisée. Tester H_0 présuppose que X et Y ont le même domaine des valeurs Δ. On partage Δ en h classes adjacentes $C_1,C_2,\ldots,C_h$ et on compte le nombre n_k [resp. n'_k] de valeurs x_i [resp. y_j] qui appartiennent à C_k.

Classes	C_1	C_2	•	C_h	**Total**
Effectif des valeurs prises par X	n_1	n_2	•	n_h	n
Effectif des valeurs prises par Y	n'_1	n'_2	•	n'_h	n'

lire : n_1 valeurs x_i (parmi les n) et n'_1 valeurs y_j (parmi les n') appartiennent à la classe C_1, etc.

On utilise pour indicateur de proximité des distributions, la variable statistique

$$Z_{n,n'} = n \times n' \times \sum_{i=1}^{h} \frac{(N_i/n - N'_i/n')^2}{N_i + N'_i}$$

où N_i [resp. N'_i] désigne le nombre aléatoire d'éléments qui, sur l'échantillon de taille n [resp. n'] extrait de P [resp. Q], appartiennent à la classe C_i.

Sous l'hypothèse H_0 « *X et Y ont même distribution* », la *v.a.* $Z_{n,n'}$ suit sensiblement la loi khi-deux à $(h-1)$ degrés de liberté : si $\forall\, k$ on a n_k et $n'_k \geqslant 3$ alors $Z_{n,n'} \cong \chi^2_{h-1}$. [Afin de respecter la contrainte « n_k et $n'_k \geqslant 3 \quad \forall\, k$ » il est parfois nécessaire de regrouper des classes adjacentes].

Règle de décision. Si l'hypothèse alternative $\bar{H}_0$ est vraie, la *v.a.* $Z_{n,n'}$ tend à prendre des valeurs élevées. Le domaine de rejet de H_0 est donc de type $[c_\alpha,\infty[$ où prenant un risque d'erreur α, c_α est tel $\alpha = P_0(Z_{n,n'} \geqslant c_\alpha) \cong P(\chi^2_{h-1} \geqslant c_\alpha)$. On rejette H_0 lorsque la valeur $z = n \times n' \sum_{i=1}^{h} \frac{(n_i/n - n'_i/n')^2}{n_i + n'_i}$ prise par $Z_{n,n'}$ est supérieure à c_α. (*Cf.* exercice 2)

2 Test de Wilcoxon-Mann-Whitney

Notant F la fonction de répartition sur la population $\boldsymbol{P}$ et G la fonction de répartition sur $\boldsymbol{Q}$ on peut, à partir des valeurs numériques $x_1,\ldots,x_n$ et $y_1,\ldots,y_{n'}$ qui sont les réalisations de chaque échantillon, tester H_0 « $\boldsymbol{P}$ et $\boldsymbol{Q}$ *ont même distribution* soit $F(t) = G(t) \quad \forall t$ » contre respectivement :

© Dunod. La photocopie non autorisée est un délit.

[1] $\bar{H}_0$ « $G(t) \neq F(t)$ » soit $\boldsymbol{P}$ et $\boldsymbol{Q}$ ont des distributions différentes

[2] H_1 « $F(t) < G(t)$ » soit X *stochastiquement supérieur à* Y ce qui signifie que chaque p-quantile $\xi_p(X)$ (*cf.* figure 12.2) de la loi suivie par X est supérieur au p-quantile $\xi_p(Y)$ de la loi de Y).

[3] H_1 « $F(t) > G(t)$ » soit X *stochastiquement inférieur à* Y.

On dispose pour cela d'un échantillon $X_1,\ldots,X_n$ *iid* d'une loi continue F (c'est-à-dire dont F est la fonction de répartition) et d'un échantillon $Y_1,\ldots,Y_{n'}$ *iid* d'une loi continue G, avec $n \leqslant n'$ selon la convention généralement adoptée qui consiste à attribuer la lettre X à l'échantillon qui a la plus petite taille.

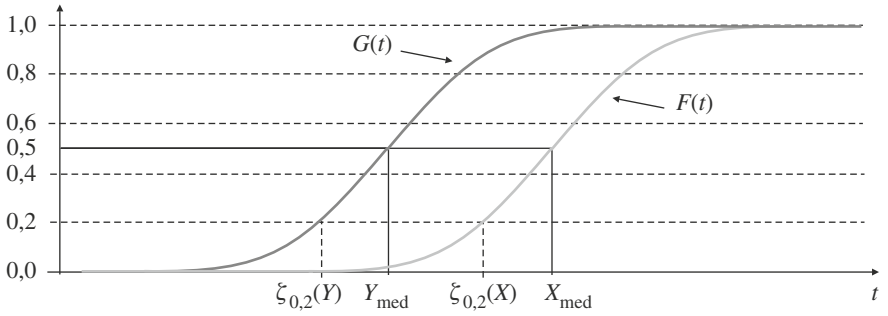


Figure 12.2 – *Fonctions de répartition*

– **La statistique** $W_{n,n'}$ **de Wilcoxon**. Après regroupement des valeurs $x_1,\ldots,x_n$ et $y_1,\ldots,y_{n'}$ prises par les échantillons $X_1,\ldots,X_n$ et $Y_1,\ldots,Y_{n'}$, on classe par ordre croissant ces $n+n'$ valeurs numériques afin d'obtenir les rangs r_h des observations x_h (avec $h = 1,2,...,n$) et les rangs s_k des observations y_k (avec $k = 1,2,\ldots,n'$).

La somme w des rangs occupés par les valeurs $x_1,\ldots,x_n$ ($w = \sum_{h=1}^{n} r_h$) est la réalisation de la statistique $W_{n,n'}$ de Wilcoxon définie comme étant la somme des rangs $R_1,\ldots,R_n$ occupés par les variables aléatoires $X_1,\ldots,X_n$: $W_{n,n'} = \sum_{h=1}^{n} R_h$.

$W_{n,n'}$ statistique linéaire de rangs a pour domaine de valeurs
$\Delta_W = \{n(n+1)/2,\, n(n+1)/2+1, n(n+1)/2+2,\ldots, n(n+1)/2+nn'\}$.

– **Distribution de la statistique** $W_{n,n'}$ **sous l'hypothèse** H_0. Sous l'hypothèse de base H_0 « *X et Y suivent une même loi continue* », la distribution de $W_{n,n'}$ est indépendante de cette loi et possède les propriétés suivantes :

i) $E_0(W_{n,n'}) = n(n+n'+1)/2$; $V_0(W_{n,n'}) = nn'(n+n'+1)/12$

ii) elle est symétrique autour de $n(n+n'+1)/2$, aussi

$$P_0(W_{n,n'} = h) = P_0(W_{n,n'} = n(n+n'+1) - h) \quad \forall\, h \in \Delta_W\ ;$$

iii) elle est tabulée pour de faibles valeurs de n et n' [*cf.* p. 380],

iv) pour n et n' supérieurs à 25 on utilise l'approximation normale

$$\frac{W_{n,n'} - n(n+n'+1)/2}{\sqrt{nn'(n+n'+1)/12}} \cong \mathcal{N}_{0;1} \text{ avec correction de continuité.}$$

Règle de décision suivant la forme prise par l'hypothèse alternative

[1] Test avec l'hypothèse alternative $\bar{H}_0$ « $F \neq G$ » (test bilatéral)

Le domaine de rejet δ_α de H_0 est de type $[n(n+1)/2, c'_{\alpha/2}] \cup [c_{\alpha/2}, n(n+1)/2 + nn']$ où $c'_{\alpha/2}$ est le plus grand entier tel que $P_0(W_{n,n'} \leqslant c'_{\alpha/2}) \leqslant \alpha/2$ et $c_{\alpha/2}$ est le plus petit entier tel que $P_0(W_{n,n'} \geqslant c_{\alpha/2}) \leqslant \alpha/2$. En raison de la symétrie de la distribution de $W_{n,n'}$, sous H_0 on a $c'_{\alpha/2} + c_{\alpha/2} = n(n+n'+1)$.

On rejette H_0 au profit de $\bar{H}_0$ si la valeur w prise par $W_{n,n'}$ appartient au domaine de rejet δ_α. Le niveau de signification observé est
$\alpha_c = 2 \times \text{Min}\{P_0(W_{n,n'} \leqslant w),\ P_0(W_{n,n'} \geqslant w)\}$.

[2] Test avec l'hypothèse alternative H_1 « $F(t) < G(t)$ »

Le domaine de rejet de H_0 est du type $[c_\alpha, n(n+1)/2 + nn']$. Prenant un niveau de signification α, on cherche le plus petit entier c_α tel que $P_0(W_{n,n'} \geqslant c_\alpha) \leqslant \alpha$. Si $w \geqslant c_\alpha$ on rejette H_0 au profit de H_1. Le niveau de signification observé $\alpha_c = P_0(W_{n,n'} \geqslant w)$.

[3] Test avec l'hypothèse alternative H'_1 « $F(t) > G(t)$ »

Le domaine de rejet de H_0 est $[n(n+1)/2, c'_\alpha]$. Pour un niveau de signification α, on cherche le plus grand entier c'_α tel que $P_0(W_{n,n'} \leqslant c'_\alpha) \leqslant \alpha$. Si $w \leqslant c'_\alpha$, on rejette H_0 au profit de H'_1. Le niveau de signification observé est $\alpha_c = P_0(W_{n,n'} \leqslant w)$.

Exemple

Pour comparer la rentabilité de firmes situées dans deux secteurs ***P*** et ***Q***, on dispose d'un échantillon de 3 firmes prélevées dans le secteur P auxquelles correspondent les *v.a.* X_1, X_2, X_3 et d'un échantillon de 4 firmes prélevées dans le secteur Q auxquelles correspondent les *v.a.* Y_1, Y_2, Y_3, Y_4. Les relevés exprimés en % sont $x_1 = 2{,}1$; $x_2 = 1{,}3$; $x_3 = -1{,}5$; $y_1 = 2{,}6$; $y_2 = -2{,}0$; $y_3 = 3{,}1$; $y_4 = 1{,}8$. Les distributions sur les populations ***P*** et ***Q***, bien que nécessairement discrètes sont considérées être correctement interpolées par des distributions continues, les éventuelles valeurs ex aequo étant classées par utilisation de la table de nombres au hasard.

On teste H_0 « $F(t) = G(t)$, autrement dit les distributions de la rentabilité X des firmes du secteur ***P*** et de la rentabilité Y des firmes du secteur ***Q*** sont sensiblement les mêmes » contre H'_1 « $F(t) > G(t)$ c'est-à-dire *X est stochastiquement inférieur à Y* » cas [3]. Pour cela, il convient de classer par ordre de valeurs croissantes les relevés des deux échantillons regroupés :

© Dunod. La photocopie non autorisée est un délit.

Rang	1	2	3	4	5	6	7
Valeur	$y_2 = -2.0$	$x_3 = -1{,}5$	$x_2 = 1{,}3$	$y_4 = 1{,}8$	$x_1 = 2{,}1$	$y_1 = 2{,}6$	$y_3 = 3{,}1$

On constate que x_1,x_2 et x_3 occupent respectivement les rangs 5, 3 et 2 : soit $r_1 = 5$, $r_2 = 3$, $r_3 = 2$. La statistique $W_{3,4}$ prend donc la valeur $w = 5 + 3 + 2 = 10$. Pour un niveau de signification $\alpha = 10\,\%$ on cherche le plus grand entier c'_α tel que $P_0(W_{3,4} \leqslant c'_\alpha) \leqslant 0{,}10$. On lit p. 380 $c'_\alpha = 7$. La *v.a.* $W_{3,4}$ prenant une valeur $w = 10$ supérieure à 7, on conclut au non rejet de l'hypothèse H_0 « $G = F$ ».

Remarques. Lorsque les lois F ou G sont discrètes, les observations peuvent présenter des ex aequo. Pour remédier à cette situation, on recourt à une méthode de départition des ex aequo par usage de la table de nombres au hasard et ainsi on définit le rang de chaque observation. Puis on utilise le test de Wilcoxon tel qu'il a été présenté dans le cadre des lois continues.

Soit, par exemple, un échantillon de X de taille 4 : $x_1 = 2$, $x_2 = 1$, $x_3 = 5$, $x_4 = 2$ et un échantillon de Y de taille 5 : $y_1 = 2$, $y_2 = 6$, $y_3 = 3$, $y_4 = 7$ et $y_5 = 8$. On a le rangement $x_2 < x_1 = x_4 = y_1 < y_3 < x_3 < y_2 < y_4 < y_5$. On substitue aux valeurs égales, des valeurs différenciées à l'aide de la table de nombres au hasard. En utilisant une colonne à deux chiffres de la table de nombres au hasard, on lit successivement : 50, 84, 22, 68, …. Choisissant un entier naturel h arbitrairement grand, aux ex aequo $x_1 = x_4 = y_1 = 2$ on substitue respectivement $x'_1 = 2 + 50 \times 10^{-h}$, $x'_4 = 2 + 84 \times 10^{-h}$ et $y'_1 = 2 + 22 \times 10^{-h}$. On obtient alors le rangement suivant : $x_2 < y'_1 < x'_1 < x'_4 < y_3 < x_3 < y_2 < y_4 < y_5$ pour lequel $W_{4,5}$ prend la valeur 14.

En présence d'ex æquo, les logiciels utilisent la méthode des rangs moyens.

Dans l'exemple précédent les *ex aequo* x_1, x_2 et y_1 occupent les rangs 2, 3 et 4 dont le rang moyen est $(2 + 3 + 4)/3 = 3$ donc x_1 a pour rang $r_1 = 3$, x_4 a pour rang $r_4 = 3$ et par suite $w = 3 + 1 + 6 + 3 = 13$. Ce procédé n'est valide que pour les échantillons de grande taille.

Section **3**

TESTS AVEC EXCEL ET SPSS

Dans le cadre d'une étude régionale sur le chômage, on a sélectionné au hasard 15 diplômés issus d'une filière universitaire professionnalisée P et 17 titulaires d'un baccalauréat professionnel Q et recensé ci-dessous le temps (évalué en semaines) qui leur a été nécessaire pour avoir un premier emploi après l'obtention de leur diplôme.

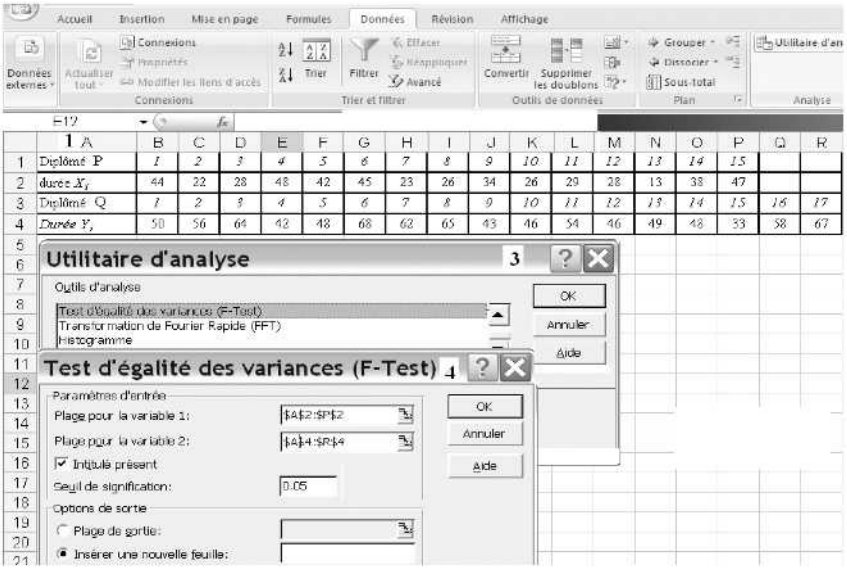
Diplômé P	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
Durée X_i	44	22	28	48	42	45	23	26	34	26	29	28	13	38	47		
Diplômé Q	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Durée Y_i	50	56	64	42	48	68	62	65	43	46	54	46	49	48	33	58	67

À partir de ces résultats peut-on considérer que le temps Y mis par le titulaire d'un bac professionnel est équivalent ou différent à celui X mis par un diplômé de la filière universitaire ?

1 Traitements statistiques avec Excel

L'outil qui, sous Excel, permet de tester H_0 « $m_P = m_Q$ » contre $\bar{H}_0$ « $m_P \neq m_Q$ » utilise une variable de Student . Les distributions de X et de Y sont supposées normales, soit $X \rightsquigarrow \mathcal{N}(m_P, \sigma_P^2)$ et $Y \rightsquigarrow \mathcal{N}(m_Q, \sigma_Q^2)$ avec σ_P et σ_Q inconnus . Aussi convient-il de s'assurer que $\sigma_P = \sigma_Q$ avant de réaliser le test de comparaison correspondant.

1.1 On teste avec Excel H_0 « $\sigma_P^2 - \sigma_Q^2 = 0$ » contre $\bar{H}_0$ « $\sigma_P^2 - \sigma_Q^2 \neq 0$ »



Procédure.

1. On entre les données du premier échantillon dans la ligne 2 et les données du second échantillon dans la ligne 4 (*cf.* ci-dessus).

2. On se place dans une cellule quelconque non utilisée et l'on sélectionne (Données) puis on clique sur (Utilitaire d'analyse)

3. Dans le menu déroulant **Utilitaire d'analyse** on sélectionne (Test d'égalité des variances (F-test)) et on clique sur (OK)

4. Dans le menu test d'égalité des variances, on désigne dans la matrice 1 les cellules correspondant aux données du premier échantillon soit ($A$2 : $P$2) et dans la matrice 2 celles du deuxième échantillon soit ($A$4 : $R$4) puisque l'on indique que l'intitulé est présent dans la plage des cellules.

© Dunod. La photocopie non autorisée est un délit.

Test d'égalité des variances (F-Test)

	Durée X_i	**Durée** Y_i
Moyenne	32,87	52,88
Variance	112,70	99,74
Observations	15,00	17,00
Degré de liberté	14,00	16,00
F	1,13	
$P(F \geqslant f)$ **unilatéral**	0,40	
Valeur critique pour F **(unilatéral)**	2,37	

On constate les valeurs des moyennes et des variances standard de deux échantillons : $\bar{x} = 32{,}87$, $\bar{y} = 52{,}88$, $s_x^2 = 112{,}7$ avec $n_P = 15$ et $s_y^2 = 99{,}74$ avec $n_Q = 17$.

On constate que sous l'hypothèse H_0 la *v.a.* $T_0 = S_X^2/S_Y^2$ suit la loi de Fisher-Snédécor avec $(n_P - 1) = 14$ et $(n_Q - 1) = 16$ degrés de liberté. T_0 prenant la valeur $t_0^* = 1{,}13$, le niveau de signification observé est de 0,40 pour un test unilatéral (de $2 \times 0{,}4 = 0{,}8$ pour un test bilatéral). Autrement dit, le risque de rejeter à tort H_0 est d'au moins 80 % et est supérieur à un risque d'erreur raisonnable de 5 % ($\alpha = 0{,}05$), aussi ne peut-on pas rejeter H_0. L'ordinateur présélectionne l'hypothèse alternative du test unilatéral la plus vraisemblable eu égard aux résultats de l'échantillon, soit ici $\sigma_P^2 > \sigma_Q^2$ puisque l'estimation s_x^2 est supérieure à s_y^2.

Une fois vérifié que $\sigma_P = \sigma_Q$, on peut utiliser sous EXCEL la variable $\boldsymbol{t}_{n_p+n_{q-2}}$ de Student pour tester H_0 « $m_P = m_Q$ » contre $\bar{H}_0$ « $m_P \neq m_Q$ ».

1.2 On teste avec Excel H_0 « $m_P - m_Q = 0$ » contre $\bar{H}_0$ « $m_P - m_Q \neq 0$ »

Procédure. Même procédure que précédemment.

1. On sélectionne (Données) puis (Utilitaire d'analyse).

2. Dans le menu **utilitaire d'analyse**, on sélectionne (Test d'égalité des espérances : deux observations de variances égales) puis (OK) (remarque si les variances sont inégales, on sélectionne (Test d'égalité des espérances : deux observations de variances différentes) soit le test **d'Aspin-Welch** présenté au 1.3 de la section 1)

3. Dans le menu **test d'égalité des espérances** on désigne dans plage pour la variable 1 les coordonnées des données du premier échantillon ($A$2 :$P$2) et dans plage pour la variable 2 celles du second ($A$4 :$R$4).

On obtient le tableau suivant

Test d'égalité des espérances : deux observations de variances égales

	Durée X_i	**Durée** Y_i
Moyenne	32,87	52,88
Variance	112,70	99,74
Observations	15,00	17,00
Variance pondérée	105,78	
Différence hypothétique des moyennes	0,00	
Degré de liberté	30,00	
Statistiques t	– 5,49	
unilatéral	2,89 E-06	
Valeur critique de t **(unilatéral)**	1,70	
bilatéral	5,78 E-06	
Valeur critique de t **(bilatéral)**	2,04	

Commentaires.

$\bar{X} = (X_1 + \ldots + X_{15})/15$ estimateur de m_P prend pour valeur (32,87),
$\bar{Y} = (Y_1 + \ldots + Y_{17})/17$ estimateur de m_Q prend pour valeur (52,88) .

Loi suivie par la statistique associée à l'estimateur :

« sous H_0 : $T_0 = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{(n_P - 1)S_x^2 + (n_Q - 1)S_y^2}{n_P + n_Q - 2}} \times \sqrt{\dfrac{1}{n_P} + \dfrac{1}{n_Q}}} = \boldsymbol{t}_{n_P + n_Q - 2}$

où $n_P + n_Q - 2 = 15 + 17 - 2 = \boxed{30}$ »

La valeur t_0^* prise par T_0 est égale à (– 5,49).

Test bilatéral (ou hypothèse alternative $\bar{H}_0$ « $m_P \neq m_Q$ »). Le niveau de signification observé du test bilatéral apparaît $2 \times P(T_0 \geqslant |t_0^*|)$ bilatéral $(= 5{,}78 \times 10^{-6})$. Il y a donc une probabilité négligeable de $5{,}78195 \times 10^{-6}$ de commettre une erreur si on décide de rejeter H_0.

Pour un risque d'erreur de 5 % le domaine de rejet du test bilatéral est de type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$ où $c_{\alpha/2}$ est défini par $P(-c_{\alpha/2} < \boldsymbol{t}_{n_P + n_Q - 2} < c_{\alpha/2}) = 1 - \alpha = 0{,}95$ et l'ordinateur donne $c_{\alpha/2} = 2{,}04$.

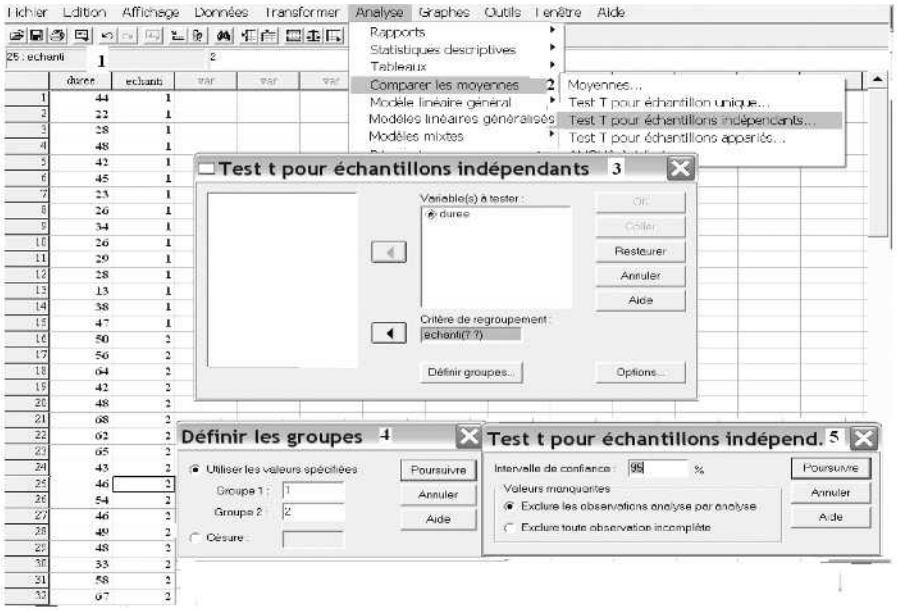
Test unilatéral. L'ordinateur retient pour hypothèse alternative H_1 « $m_P > m_Q$ » (cas [2]) si la valeur t_0^* prise par T_0 (appelée **Statistique** t) est supérieure à 0. Dans le cas contraire l'hypothèse alternative est H_1' « $m_P < m_Q$ » (cas [3]).

© Dunod. La photocopie non autorisée est un délit.

2 Traitements statistiques avec SPSS

2.1 Test de H_0 « $m_P = m_Q$ » contre $\bar{H}_0$ « $m_P \neq m_Q$ »

Est reprise la même étude que dans le § 1 des filières respectivement notées P et Q. On utilise le test T de Student afin de tester H_0 « $m_P = m_Q$ » contre $\bar{H}_0$ « $m_P \neq m_Q$ », les distributions de X et Y étant supposées normales.



Procédure.

1. On entre les observations sur la durée dans une seule colonne et on fait figurer l'échantillon correspondant 1 ou 2 dans une autre colonne.

2. On pointe sur (Analyse) puis sur (Comparer les moyennes) et (Test *T* pour échantillons indépendants).

3. Dans le menu **test-*T* pour échantillon…** on retient la **variable à tester** *duree* et la variable *echant* pour **variable de regroupement**.

4. Cliquant sur (Définir groupes) on obtient le menu dans lequel on sélectionne les échantillons à comparer, ici on met (1) dans groupe 1 puis (2) dans groupe 2.

5. Cliquant sur (Options) on obtient le menu dans lequel on sélectionne le risque d'erreur ou inversement le niveau de confiance.

Résultats

	Échantillon	N	Moyenne	Ecart-standard de l'échantillon	Erreur standard moyenne
Durée	1	$n_P = 15$	32,86 $(= \bar{x})$	10,6291 $(= s_x)$	2,74 $(= \sqrt{s_x^2/n_P})$
	2	$n_Q = 17$	52,88 $(= \bar{y})$	9,99 $(= s_y)$	2,42 $(= \sqrt{s_y^2/n_Q})$

Test-t pour égalités des moyennes

		t	ddl	Sig. bilatérale	Différence moyenne	Différence écart-type	Intervalle de confiance 95 % de la différence	
							inférieure	supérieure
Durée	Hypothèse de variances égales	−5,493	30	$5{,}78E-06$	−20,02	3,64	−27,46	−12,57
	Hyp. de variances inégales	−5,47	28,95	$6{,}87E-06$	−20,02	3,66	−27,50	−12,53

Commentaires des résultats

Le logiciel teste systématiquement H_0 « $m_P = m_Q$ » contre $\bar{H}_0$ « $m_P \neq m_Q$ ».

$\bar{X} = (X_1 + \ldots + X_{15})/15$ estimateur de m_P prend pour valeur (32,86) et $\bar{Y} = (Y_1 + \ldots + Y_{17})/17$ estimateur de m_Q prend pour valeur (52,88) et $\bar{X} - \bar{Y}$ estimateur de $m_P - m_Q$ prend pour valeur (−20,02).

– *Loi suivie par la statistique associée à l'estimateur.* SPSS suppose qu'il s'agit d'une distribution normale et les écarts-types étant proches on utilise, sous H_0, la statistique $T_0 = (\bar{X} - \bar{Y})/\sqrt{\dfrac{(n_P-1)S_x^2 + (n_Q-1)S_y^2}{n_P + n_Q - 2}}\sqrt{\dfrac{1}{n_P} + \dfrac{1}{n_Q}} = \boldsymbol{t}_{n_P+n_Q-2}$

où $n_P + n_Q - 2 = 15 + 17 - 2 = 30$.

La valeur t_0^* prise par T_0 est égale à (−5,49). Pour un risque d'erreur $\alpha = 0{,}05$, le domaine de rejet de H_0 est $]-\infty, -27{,}46] \cup [-12{,}57, \infty[$. −20,02 appartenant au domaine de rejet de H_0, on rejette H_0 avec un risque d'erreur de 5 %. La probabilité de commettre une erreur en rejetant l'hypothèse H_0 ou niveau de signification observé α_c est égal à sig. $(= 5{,}78 \times 10^{-6}) = 2 \times P(\boldsymbol{t}_{30} > |-5{,}493|) = (2{,}89 \times 10^{-6})$

– Dans le cas où les variances seraient inégales la statistique est celle d'Aspin-Welch

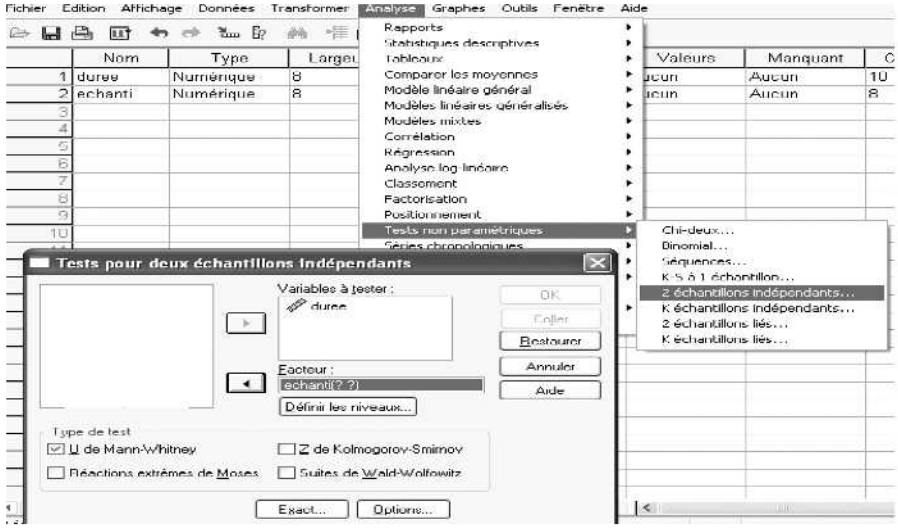
$$T_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/n_P + S_Y^2/n_Q}} = \boldsymbol{t}_\nu \quad \text{où} \quad \nu = \frac{(s_x^2/n_P + s_y^2/n_Q)^2}{(s_x^2/n_P)^2/(n_P-1) + (s_y^2/n_Q)^2/(n_Q-1)}$$

$= 28{,}95$. La valeur t_0^* prise par T_0 est égale à (−5,47) et la probabilité de commettre une erreur en rejetant l'hypothèse H_0 est égale à sig. $(= 6{,}87 \times 10^{-6})$. On rejette donc H_0 avec un risque d'erreur raisonnable.

© Dunod. La photocopie non autorisée est un délit.

2.2 Test de H_0 « $F = G$ » contre $\bar{H}_0$ « $G \neq F$ » avec Wilcoxon-Mann-Whitney

Il s'agit de tester avec le logiciel **SPSS** l'hypothèse H_0 « $F = G$ » contre $\bar{H}_0$ « $G \neq F$ » à l'aide du test de Wilcoxon-Mann-Whitney, aucune hypothèse n'étant formulée concernant les distributions de X et de Y.



Procédure.

1. On clique sur (Analyse) puis (Tests non paramétriques) puis (2 échantillons indépendants.)

2. Dans le menu **Test pour deux échantillons indépendants** on sélectionne la **variable à tester duree** et la variable de regroupement (les échantillons) **echanti.**

3. Cliquant sur (Définir niveaux) on met 1 pour le groupe 1 et pour le groupe 2 et on clique sur (Poursuivre) et (Ok) .

4. En cliquant sur (Options) on peut opter pour un test unilatéral ou bilatéral.

On obtient les résultats suivants

	ECHANTIL	N	Rang moyen	Somme des rangs
DUREE	1	15	9,37	140,50
	2	17	22,79	387,50
	Total	32		

Test	
	DUREE
U **de Mann-Whitney**	20,50
W **de Wilcoxon**	140,50
Z	$-4,04$
Signification asymptotique (bilatérale)	$5,26E-05$
Signification exacte	$9,48E-06$

Commentaires. Le plus petit échantillon est l'échantillon 1 des diplômés P aussi $n = 15$, l'échantillon des diplômés Q ayant la taille $n' = 17$.

La statistique W de Wilcoxon correspond à la somme des rangs du plus petit échantillon (ici de l'échantillon 1 des diplômés P) soit $W_{15,17} = \sum_{i=1}^{15} R_i$ qui prend pour valeur $\boxed{140{,}5}$.

La *v.a.* $W_{15,17}$ prenant une valeur $w = 140{,}5$ (est pris en compte le rang moyen dans le cas d'ex æquo), la probabilité de commettre une erreur en rejetant l'hypothèse H_0 est égale à sig. $= 9{,}48 \times 10^{-6}$.

La variable de Mann-Whitney $U = W - n(n+1)/2$ prend la valeur 20,5.

Lorsque n et n' grand on utilise l'approximation normale $Z = \dfrac{W_{n,n'} - n(n+n'+1)/2}{\sqrt{nn'(n+n'+1)/12}} \cong \mathcal{N}_{0;1}$, la valeur prise par Z est $\boxed{-4{,}04}$ soit un niveau de signification observé $\boxed{5{,}26 \times 10^{-5}}$.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 357)

Exercice 1

Pour pouvoir choisir entre deux types de conditionnement possibles, le service a constitué deux échantillons de personnes. Il est demandé aux individus des deux échantillons d'indiquer le prix qu'ils considèrent comme normal pour le produit.

Au premier échantillon est proposé le produit dans son premier type de conditionnement. Sur cet échantillon de taille 50, on a trouvé un prix normal moyen égal à 42,6 pour un écart-standard égal à 4,62 et un moment centré d'ordre 4 égal à 625 .

Au deuxième échantillon de personnes est proposé le produit dans son deuxième type de conditionnement. Sur cet échantillon de taille 60 on a trouvé un prix normal moyen égal à 37,8, un écart-standard égal à 4,45 et un moment centré d'ordre 4 égal à 600.

1. Au seuil de 10 %, existe-t-il une différence significative entre les moyennes des prix selon le type de conditionnement ?

2. Au seuil de 10 %, existe-t-il une différence significative entre les écart-types des prix selon le type de conditionnement ?

Exercice 2

Un sondage a été réalisé dans le Hall de la gare de Libourne auprès de voyageurs qui prennent régulièrement le train pour se rendre à leur travail. La question posée était la suivante : « Combien avez-vous acheté de quotidiens *P-M* au cours de la semaine précédente ? ».

50 personnes ont été interrogées et les réponses obtenues figurent dans le tableau ci-après :

Nombre de *P-M* achetés	0	1	2	3	4	5	6	**Total**
Effectifs	5	6	9	8	6	6	10	50

Suite à une campagne publicitaire par affichage l'on a réalisé une nouvelle enquête auprès de 60 personnes et obtenu les résultats suivants :

Nombre de *P-M* achetés	0	1	2	3	4	5	6	**Total**
Effectifs	15	16	9	6	4	4	6	60

Au cours des deux semaines étudiées aucune perturbation particulière n'est intervenue du type jours fériés, grève, vacances scolaires… L'objet du sondage est de savoir si la publicité a eu une influence significative sur le comportement d'achat de ce type de clientèle. Pour cela il faut tester avec le test de comparaison du Khi-deux l'hypothèse H_0 selon laquelle la distribution du nombre aléatoire X de quotidiens achetés par un client avant

la campagne de publicité est identique à celle du nombre aléatoire Y de quotidiens achetés par un client après la campagne de publicité (prendre un risque d'erreur de première espèce de 5 %).

Exercice 3

On veut savoir si la fonction score utilisée par la banque pour décider d'octroyer ou non un crédit à la consommation sépare bien les clients défaillants des non défaillants. Le responsable du risque crédit dispose d'un échantillon de 9 clients défaillants et d'un échantillon de 7 non défaillants. Il connaît pour ces clients la valeur individuelle de leur score au moment de l'étude de leur demande de crédit :

Non-défaillants	1	2	3	4	5	6	7		
Score	0,41	0,17	0,10	0,40	0,73	0,55	– 0,50		
Défaillants	1	2	3	4	5	6	7	8	9
Score	– 0,3	– 0,18	– 0,13	0,11	0,70	0,31	– 0,32	– 0,20	– 0,30

Désignant par Y le score obtenu par un client défaillant et par X le score d'un client non-défaillant (le plus petit échantillon) on demande de tester H_0 « X et Y ont même distribution » contre H_1 « $F < G$ c'est-à-dire X est stochastiquement supérieur à Y » avec le test de Wilcoxon-Mann-Withney (prendre un risque d'erreur de première espèce de 10 %).

Exercice 4

Prélevant respectivement 130 et 89 dossiers de déclaration de revenus de 2 catégories sociales C_1 et C_2 dans une région donnée, on décide de vérifier s'il y a une même proportion de fraudeurs dans ces deux catégories sociales.

Les résultats obtenus sont présentés dans le tableau ci-dessous.

Catégories sociales	C_1	C_2
Nombre de dossiers étudiés	120	80
Nombre de fraudeurs	10	7

1. Tester l'hypothèse H_0 « il y a une même proportion de fraudeurs dans chaque catégorie sociale » contre $\bar{H}_0$ (prendre un risque d'erreur de première espèce de $\alpha = 5\,\%$).

2. Étendant l'analyse à une troisième catégorie sociale C_3 on prélève 100 dossiers de cette catégorie et on observe 15 fraudeurs. Tester l'hypothèse H_0 « il y a une même proportion de fraudeurs dans chaque catégorie sociale » contre $\bar{H}_0$ (prendre $\alpha = 5\,\%$).

QCM. $x_1 \ldots, x_{n1}$ est la réalisation d'un échantillon *iid* de $\mathcal{N}(m_1; \sigma_1^2), y_1 \ldots, y_{n2}$ est la réalisation d'un échantillon *iid* da $\mathcal{N}(m_2; \sigma_2^2)$. Pour tester $\sigma_1 = \sigma_2$ on utilise la loi :

① normale ; ② de Student ; ③ khi-deux ; ④ de Fisher-Snedécor.

© Dunod. La photocopie non autorisée est un délit.

13 COUPLES ALÉATOIRES ET TESTS D'INDÉPENDANCE

Un couple aléatoire de réels, noté dans le chapitre (X,Y), est le résultat numérique d'une expérience envisagée, par exemple la taille X et la rentabilité Y d'une entreprise prélevée au hasard. Pour déceler un lien éventuel entre deux caractères sur une population, il faut disposer d'un échantillon considéré comme tiré au hasard et utiliser des procédures développées dans ce chapitre. Ainsi prélevant un échantillon d'entreprises du secteur S on peut déterminer si la rentabilité d'une firme est liée ou non à sa taille.

Section 1 ■ Lois bivariées discrètes
Section 2 ■ Lois bivariées continues
Section 3 ■ Test d'indépendance par la méthode du khi-deux
Section 4 ■ Mesures d'association entre deux variables
Section 5 ■ Traitement sous Excel et SPSS

Section 1

LOIS BIVARIÉES DISCRÈTES

1 Distribution de probabilité

La loi que suit un couple (X,Y) où X peut prendre les valeurs $\{x_1^*,x_2^*,\ldots,x_m^*\}$ et Y les valeurs $\{y_1^*,y_2^*,\ldots,y_n^*\}$ est caractérisée :

– par son domaine de définition $\Delta_{XY} = \{(x_1^*,y_1^*),(x_1^*,y_2^*),\ldots,(x_2^*,y_1^*),\ldots,(x_m^*,y_n^*)\}$

– par les $m \times n$ nombres p_{ij} tels que $P(X = x_i^* \text{ et } Y = y_j^*) = p_{ij}$ où $p_{ij} \geqslant 0$ et $\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} = 1$.

Elle peut être présentée sous forme de tableau. La case située à l'intersection de la i-ème colonne et la j-ème ligne représente l'événement $(X = x_i^* \text{ et } Y = y_j^*)$, le nombre p_{ij} figurant dans cette case est sa probabilité de réalisation.

Tableau 13.1 – *Distribution de probabilité du couple*

Y \ X	x_1^*	x_2^*	…	x_i^*	…	x_m^*	**Total**
y_1^*	p_{11}	p_{21}	…	p_{i1}	…	p_{m1}	$p_{\bullet 1}$
y_2^*	p_{12}	p_{22}	…	p_{i2}	…	p_{m2}	$p_{\bullet 2}$
…	…	…	…		…	…	…
y_j^*	p_{1j}	p_{2j}	…	$\boldsymbol{p_{ij}}$	…		
…	…	…	…	…	…	…	…
y_n^*	p_{1n}	p_{2n}	…	…	…	p_{mn}	$p_{\bullet n}$
Total	$p_{1\bullet}$	$p_{2\bullet}$	…	…	…	$p_{m\bullet}$	1

Lire : $P(X = x_1^* \text{ et } Y = y_1^*) = p_{11}$, $P(X = x_1^* \text{ et } Y = y_2^*) = p_{12}$, etc.

– Les première et dernière lignes du tableau caractérisent la loi de X appelée distribution marginale de X :

$$\Delta_X = \{x_1^*,x_2^*,\ldots,x_m^*\},\ P(X = x_i^*) = p_{i\bullet} = p_{i1} + p_{i2} + \ldots + p_{in}$$

$$E(X) = \sum_{i=1}^{m} x_i^* p_{i\bullet}\ ;\ E(X^2) = \sum_{i=1}^{m} (x_i^*)^2 p_{i\bullet}\ ;\ V(X) = E(X^2) - E(X)^2$$

© Dunod. La photocopie non autorisée est un délit.

– Les première et dernière colonnes du tableau caractérisent la loi de Y ou distribution marginale de Y :

$$\Delta_Y = \{y_1^*, y_2^*, \ldots, y_n^*\},\ P(Y = y_j^*) = p_{\bullet j} \text{ où } p_{\bullet j} > 0 \text{ et } \sum_{j=1}^{n} p_{j\bullet} = 1\,.$$

$$E(Y) = \sum_{j=1}^{n} y_j^* p_{\bullet j},\ E(Y^2) = \sum_{j=1}^{n} (y_j^*)^2 p_{\bullet j},\ V(Y) = E(Y^2) - E(Y)^2$$

– La loi de probabilité conditionnelle de X pour $Y = y_j$ est caractérisée par $P(X = x_i / Y = y_j) = p_{ij}/p_{\bullet j}$. En particulier $E(X/Y = y_j) = \sum_{i=1}^{m} x_i^* p_{ij}/p_{\bullet j}$

– La loi de probabilité conditionnelle de Y pour $X = x_i$ est caractérisée par $P(Y = y_j / X = x_i) = p_{ij}/p_{i\bullet}$. En particulier $E(Y/X = x_i) = \sum_{j=1}^{n} y_j^* p_{ij}/p_{i\bullet}$

Propriété. $E(X) = \sum_{j=1}^{n} E(X/Y = y_j) \times (P(Y = y_j)$;

$$E(Y) = \sum_{i=1}^{m} E(Y/X = x_i) \times (P(X = x_i)$$

Exemple

Considérons la distribution de probabilité d'un couple (X,Y) sous forme de tableau

Y \ X	1	2	3	Total
0	$p_{11} = 0{,}2$	$p_{21} = 0{,}2$	$p_{31} = 0{,}2$	$p_{\bullet 1} = 0{,}6$
1	$p_{12} = 0$	$p_{22} = 0{,}4$	$p_{32} = 0$	$p_{\bullet 2} = 0{,}4$
Total	$p_{1\bullet} = 0{,}2$	$p_{2\bullet} = 0{,}6$	$p_{3\bullet} = 0{,}2$	1

Lire : $P(X = 1 \text{ et } Y = 0) = p_{11} = 0{,}2$, $P(X = 1 \text{ et } Y = 1) = p_{12} = 0$, etc.

Loi de probabilité marginale de X :

$\Delta_X = \{1,2,3\}$ avec $P(X=1)=0{,}2$, $P(X=2)=0{,}6$, $P(X = 3) = 0{,}2$.

Loi de probabilité marginale de Y :

$\Delta_Y = \{0,1\}$ avec $P(Y = 0) = 0{,}6$, $P(Y = 1) = 0{,}4$.

On constate que

$$E(X) = \sum_{i=1}^{m} x_i^* p_{i\bullet} = 2,\ E(X^2) = \sum_{i=1}^{m} (x_i^*)^2 p_{i\bullet} = 4{,}4,\ V(X) = E(X^2) - E(X)^2 = 0.4,$$

$$E(Y) = \sum_{j=1}^{n} y_j^* p_{\bullet j} = 0{,}4\,,\ E(Y^2) = 0{,}4,\ V(Y) = E(Y^2) - E(Y)^2 = 0{,}4 - 0{,}4^2 = 0{,}24$$

$E(X/Y = 0) = (1 \times 0{,}2 + 2 \times 0{,}2 + 3 \times 0{,}2)/0{,}6 = 2$;

$E(X/Y = 1) = (1 \times 0 + 2 \times 0{,}4 + 3 \times 0)/0{,}4 = 2$.

2 Moments centrés et matrice des variance-covariance

Les moments. r et s désignant deux entiers naturels, on définit :

– le moment d'ordre r en X et s en Y noté

$$m_{r,s} = E(X^r Y^s) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i^*)^r (y_j^*)^s p_{ij}$$

ainsi, $E(XY) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i^* y_j^* p_{ij}$ (on pose $r = 1$ et $s = 1$) ;

$$E(X) = m_{1,0}\,,\, E(X^2) = m_{2,0}\,,\, E(Y) = m_{0,1}\,,\, E(Y^2) = m_{0,2}$$

– le moment centré d'ordre r en X et s en Y noté

$$\mu_{r,s} = E[(X - m_{1,0})^r (Y - m_{0,1})^s] = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i^* - m_{1,0})^r (y_j^* - m_{0,1})^s p_{ij}$$

avec $V(X) = \mu_{2,0} = \sum_{i=1}^{m}(x_i^* - m_{1,0})^2 p_{i\bullet}$, $V(Y) = \mu_{0,2} = \sum_{j=1}^{n}(y_j^* - m_{0,1})^2 p_{\bullet j}$

– la covariance

$$\mathrm{Cov}(X,Y) = \mu_{1,1} = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i^* - m_{1,0})(y_j^* - m_{0,1}) p_{ij} = E(XY) - E(X) \times E(Y)$$

La covariance satisfait aux propriétés suivantes :

i) $|\mathrm{Cov}(X,Y)|^2 \leqslant V(X) \times V(Y)$;

ii) $\mathrm{Cov}(aX + b, cY + d) = \mathrm{Cov}(aX, cY) = ac\,\mathrm{Cov}(X,Y)$ $\forall$ les réels a, b, c et d ;

iii) $V(\alpha X + \beta Y) = \alpha^2 V(X) + \beta^2 V(Y) + 2\alpha\beta\,\mathrm{Cov}(X,Y)$, $\forall$ les réels α et β.

– le coefficient de corrélation linéaire $r(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)} \times \sqrt{V(Y)}}$ qui mesure le degré de dépendance entre X et Y. Il satisfait aux propriétés suivantes :

j) il est compris entre -1 et $+1$, soit $-1 \leqslant r(X,Y) \leqslant 1$

jj) si $r(X,Y) = 1$ [resp. $= -1$] alors il existe deux réels a et b avec $a > 0$ [resp. < 0] tels que $Y = aX + b$. Le domaine $\Delta_{X,Y}$ est alors constitué de points portés par la droite d'équation $y = ax + b$.

Remarque. D'une façon générale considérant la *v.a.* $Z = \Psi(X,Y)$ où $\Psi(x,y)$ est une fonction définie sur $\Delta_{X,Y}$ et à valeurs réelles on a $E(Z) = E[\Psi(X,Y)] = \sum_i \sum_j \Psi(x_i, y_j) p_{ij}$.

Centre et matrice des covariances. Le centre $(m_{1,0}, m_{0,1})$ et la matrice W des covariances de la loi bivariée que suit le couple (X,Y) sont définis par :

$$(m_{1,0}, m_{0,1}) = (E(X), E(Y)) \;;\; W = \begin{bmatrix} V(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & V(Y) \end{bmatrix}$$

© Dunod. La photocopie non autorisée est un délit.

Ainsi dans l'exemple précédent on constate que

$$E(XY) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i^* y_j^* p_{ij} = (1 \times 0) \times 0{,}2 + (1 \times 1) \times 0 + \ldots$$
$$+ (3 \times 1) \times 0{,}0 = 0.8$$

$\mathrm{Cov}(X,Y) = E(XY) - E(X) \times E(Y) = 0$, $r(X,Y) = \mathrm{Cov}(X,Y)/(\sigma(X) \times \sigma(Y)) = 0$

Le centre $(m_{1,0}, m_{0,1}) = (2,\ 0{,}4)$; la matrice des covariances $W = \begin{bmatrix} 0{,}4 & 0 \\ 0 & 0{,}24 \end{bmatrix}$.

Section 2

LOIS BIVARIÉES CONTINUES

Le lecteur intéressé par les lois bivariées continues pourra se référer à l'annexe page 367 où sont définis la répartition des masses, les lois que suivent séparément X et Y, les moments et en particulier $\mathrm{Cov}(X,Y)$, $r(X,Y)$. Il y trouvera également les propriétés essentielles concernant les lois normales bivariées.

Section 3

TEST D'INDÉPENDANCE PAR LA MÉTHODE DU KHI-DEUX

Connaissant n couples de valeurs numériques $(x_1, y_1), \ldots, (x_n, y_n)$ prises par (X,Y), on s'interroge sur la possible indépendance des *v.a.* X et Y.

1 Distribution d'un couple aléatoire sur une population

Soit une expérience dont le résultat est un couple aléatoire de réels (X,Y), la probabilité pour que « $X \leqslant t$ et $Y \leqslant u$ » est appelée *fonction de répartition* du couple aléatoire (X,Y). Ainsi le nombre de télévisions $X(e)$ et le revenu mensuel $Y(e)$ d'un ménage e choisi au hasard dans la population $\boldsymbol{P} = \{e_1, e_2, \ldots, e_\nu\}$ constitue un couple aléatoire. Pour chaque couple de réels t et u, considérant les « sous populations » $\boldsymbol{P}_{t,u}$ des ménages dont simultanément le nombre de postes de télévisions est $\leqslant t$ et le revenu mensuel est $\leqslant u$ on a « $(X(e) \leqslant t$ et $Y(e) \leqslant u)$ si et seulement si $e \in \boldsymbol{P}_{t,u}$ ». Le prélèvement de e ayant lieu au hasard on a

$P(X \leqslant t \text{ et } Y \leqslant u) = \mathrm{card} \cdot \boldsymbol{P}_{t,u} / \mathrm{card} \cdot \boldsymbol{P}$.

2 Indépendance

Les *v.a.* X et Y sont dites indépendantes lorsque quelque soient les réels t et u on a : $P(X \leqslant t \text{ et } Y \leqslant u) = P(X \leqslant t) \times P(Y \leqslant u)$.

Dans le cas d'une loi bivariée discrète, on a en particulier :

$$\text{« } X \text{ et } Y \text{ sont indépendantes} \Leftrightarrow p_{ij} = p_{i\bullet} \times p_{\bullet j} \; \forall i, j \text{ ».}$$

Théorème. Si X et Y indépendants alors le coefficient de corrélation linéaire est nul : $r(X,Y) = 0$, la réciproque étant généralement fausse sauf si (X,Y) suit une loi normale bivariée. Ainsi dans l'exemple précédent $r(X,Y) = 0$ mais les variables X et Y ne sont pas indépendantes puisque en particulier $P(X = 1 \text{ et } Y = 1) \neq P(X = 1) \times P(Y = 1)$ car $0 \neq 0{,}2 \times 0{,}4$.

3 Test d'indépendance par la méthode du khi-deux

Cette méthode permet de tester H_0 « X et Y indépendants » contre $\overline{H}_0$ quelque soit la nature de la loi bivariée, discrète ou continue .

3.1 Tableau croisé

Le domaine Δ_X des valeurs possibles de X est partagé en h classes adjacentes notées $C_{1\bullet}, C_{2\bullet}, \ldots, C_{h\bullet}$. Le domaine Δ_Y des valeurs possibles de Y est partagé en k classes adjacentes notées $C_{\bullet 1}, \ldots, C_{\bullet k}$. Les n couples de valeurs numériques $(x_1, y_1), \ldots, (x_n, y_n)$ prises par (X,Y) sont répartis dans les $h \times k$ classes notées $C_{ij} = C_{i\bullet} \times C_{\bullet j}$. Le nombre de couples qui appartiennent à C_{ij} sera noté n_{ij} (nombre de couples (x_ℓ, y_ℓ) tels que $x_\ell \in C_{i\bullet}$ et $y_\ell \in C_{\bullet j}$). Le nombre de valeurs prises par X qui appartiennent à la classe $C_{i\bullet}$ sera noté $n_{i\bullet} : n_{i\bullet} = \sum_{j=1}^{k} n_{ij}$. De même, $n_{\bullet j} = \sum_{i=1}^{h} n_{ij}$ désigne le nombre de valeurs prises par Y qui appartiennent à la classe $C_{\bullet j}$.

Tableau 13.2 – *Répartition des effectifs entre les classes*

Y \ X	$C_{1\bullet}$	$C_{2\bullet}$	...	$C_{h\bullet}$	**Total**
$C_{\bullet 1}$	n_{11}	n_{21}	...	n_{h1}	$n_{\bullet 1}$
$C_{\bullet 2}$	n_{12}	n_{22}	...	n_{h2}	$n_{\bullet 2}$
...	...	...	...	...	...
$C_{\bullet k}$	n_{1k}	n_{2k}	...	n_{hk}	$n_{\bullet k}$
Total	$n_{1\bullet}$	$n_{2\bullet}$	...	$n_{h\bullet}$	n

© Dunod. La photocopie non autorisée est un délit.

3.2 Procédure de test

On teste l'hypothèse H_0 « X et Y indépendants c'est-à-dire $p_{ij} = p_{i\bullet} \times p_{\bullet j}\ \forall i,j$ » contre « X et Y ne sont pas indépendants ».

– La statistique Z utilisée. Soit N_{ij} le nombre aléatoire de couples qui, dans un échantillon de taille n appartiennent à la classe C_{ij} et soit n_{ij} sa valeur numérique prise sur l'échantillon. $N_{i\bullet}$ désigne le nombre aléatoire de couples (X,Y) dont la valeur de X appartient à la classe $C_{i\bullet}$, de même $N_{\bullet j}$ désigne le nombre aléatoire de couples (X,Y) où Y appartient à la classe $C_{\bullet j}$.

Après expérimentation $N_{i\bullet}$ et $N_{\bullet j}$ prennent respectivement pour valeur $n_{i\bullet}$ et $n_{\bullet j}$.

Si H_0 est vraie, la statistique $Z = \sum_{i=1}^{h}\sum_{j=1}^{k} \dfrac{(N_{ij} - N_{i\bullet} \times N_{\bullet j}/n)^2}{N_{i\bullet} \times N_{\bullet j}/n}$ suit sensiblement la loi $\chi^2(h-1)(k-1)$: $Z \cong \chi^2_{(h-1)(k-1)}$

La *v.a.* Z prend la valeur $z = \sum_{i=1}^{h}\sum_{j=1}^{k} \dfrac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$ où n_{ij} est le nombre de couples de l'échantillon appartenant à la classe C_{ij} et $n^*_{ij} = (n_{i\bullet} \times n_{\bullet j})/n$ est l'effectif dit « *théorique* » par analogie avec l'hypothèse d'indépendance lorsque le relevé statistique porte sur toute la population.

L'approximation $Z \cong \chi^2_{(h-1)(k-1)}$ n'est toutefois acceptable que si l'effectif « théorique » n^*_{ij} de chaque classe est supérieur ou égal à 5 : $n^*_{ij} \geqslant 5\ \forall i,j$. Lorsque les effectifs « théoriques » de certaines classes sont inférieurs à 5, on regroupe des classes adjacentes dans la partition de Δ_X ou de Δ_Y de façon à accroître les effectifs des classes.

– Règle de décision. Sous l'hypothèse alternative $\overline{H}_0$, la variable statistique Z tend à prendre des valeurs élevées. Le domaine de rejet de H_0 est de type $[c_\alpha, \infty[$ où c_α est déterminé par la condition $P_0(Z \geqslant c_\alpha) \cong P(\chi^2_{(h-1)(k-1)} \geqslant c_\alpha) = \alpha$. On rejette H_0 si $z \geqslant c_\alpha$. Le niveau de signification observé $\alpha_c = P_0(Z \geqslant z) \cong P(\chi^2_{(h-1)(k-1)} \geqslant z)$ est le risque minimal d'erreur si l'on décide de rejeter H_0. Lorsque α_c est inférieur au risque de première espèce α fixé on rejette H_0.

Exemple

Une compagnie d'assurances automobile se demande s'il y a indépendance entre X l'âge de l'assuré (X exprimé en nombre d'années) et Y le nombre d'accidents déclarés par ledit assuré au cours de l'année. Pour cela on considère le couple aléatoire (X,Y) où X peut prendre n'importe quelle valeur entre 18 et 95 ans : $\Delta_X = [18,95]$ et Y n'importe quelle valeur entière naturelle : $\Delta_Y = \mathbb{N}$. Afin de tester H_0 « X et Y indépendants » contre $\overline{H}_0$ avec un niveau de signification de 5 % on prélève dans le fichier de la compagnie un échantillon de 100 couples de valeurs numériques prises par (X,Y) : $(x_1,y_1) = (19,2)$, $(x_2,y_2) = (23,1),\ldots,(x_{100},y_{100}) = (76,0)$.

Δ_X est partagé en 3 classes ($h = 3$) et Δ_Y en 2 classes ($k = 2$) ainsi que l'indique le tableau ci-dessous. Les 100 couples de valeurs sont donc répartis entre 6 classes C_{ij} présentées ci-dessous

		$C_{1\bullet} : 18 \leqslant X < 25$	$C_{2\bullet} : 25 \leqslant X \leqslant 60$	$C_{3\bullet} : X > 60$	**Total** $n_{\bullet j}$
$C_{\bullet 1}$: $Y = 0$	***effectifs observés*** n_{ij} ***effectifs théoriques*** n^*_{ij}	$n_{11} = 12$ $n^*_{11} = 14{,}4$	$n_{21} = 40$ $n^*_{21} = 36$	$n_{31} = 20$ $n^*_{31} = 21{,}6$	$n_{\bullet 1} = 72$
$C_{\bullet 2}$: $Y \geqslant 1$	***effectifs observés*** n_{ij} ***effectifs théoriques*** n^*_{ij}	$n_{12} = 8$ $n^*_{12} = 5{,}6$	$n_{22} = 10$ $n^*_{22} = 14$	$n_{32} = 10$ $n^*_{32} = 8{,}4$	$n_{\bullet 2} = 28$
	Total $n_{i\bullet}$	$n_{1\bullet} = 20$	$n_{2\bullet} = 50$	$n_{3\bullet} = 30$	$n = 100$

Les effectifs théoriques : $n^*_{11} = n_{\bullet 1} \times n_{1\bullet}/100 = 72 \times 20/100 = 4{,}4, \ldots, n^*_{32} = n_{\bullet 2} \times n_{3\bullet}/100 = 8{,}4$.

Statistique utilisée. Tous les effectifs théoriques $n^*_{ij} = n^*_{i\bullet} \times n^*_{\bullet j}/n$ étant supérieurs à 5, sous l'hypothèse H_0 on a la propriété $Z = \sum_{i=1}^{3} \sum_{j=1}^{2} \dfrac{(N_{ij} - N_{i\bullet} \times N_{\bullet j}/n)^2}{N_{i\bullet} \times N_{\bullet j}/n} \cong \chi^2_2$.

Règle de décision. Le domaine de rejet de H_0 est nécessairement de type $[c_\alpha, \infty[$ où c_α est défini par $P_0(Z \geqslant c_\alpha) \cong P(\chi^2_2 \geqslant c_\alpha) = \alpha = 0{,}05$ soit $c_\alpha = 5{,}99$. La *v.a.* Z prend pour valeur $z = (12 - 14{,}4)^2/14{,}4 + (8 - 5{,}6)^2/5{,}6 + (40 - 36)^2/36 + (10 - 14)^2/14 + (20 - 21{,}6)^2/21{,}6 + (10 - 8{,}4)^2/8{,}4 = 3{,}439$. La valeur z étant inférieure à 5,99 on décide de ne pas rejeter l'hypothèse d'indépendance des variables X et Y.

REPÈRES : Cas particulier d'un tableau 2 × 2

Sous l'hypothèse H_0 « X et Y indépendants » et à condition que $np_{ij} \geqslant 4$ (en pratique $n_{ij} \geqslant 5$) $\forall\, i, j$ on peut utiliser l'approximation $T_0 = \dfrac{\sqrt{n}(N_{11}N_{22} - N_{12}N_{21})}{\sqrt{N_{1\bullet}N_{2\bullet}N_{\bullet 1}N_{\bullet 2}}} \cong \mathcal{N}_{0;1}$ pour tester H_0. T_0 après expérimentation prend la valeur $t^*_0 = \dfrac{\sqrt{n}(n_{11}n_{22} - n_{12}n_{21})}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}$.

Règle de décision

[1] Si l'hypothèse alternative est $\overline{H}_0$, le domaine de rejet de H_0 est de type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$ où $P_0(T_0 \geqslant c_{\alpha/2}) = P(\mathcal{N}_{0;1} \geqslant c_{\alpha/2}) = \alpha/2$ soit $\phi(c_{\alpha/2}) = 1 - \alpha/2$.

[2] Si l'hypothèse alternative est l'hypothèse de concordance de valeurs H_1 « $p_{11}p_{22} - p_{12}p_{21} > 0$ » le domaine de rejet est $[c_\alpha, \infty[$ où c_α est déterminé par la condition $P_0(T_0 \geqslant c_\alpha) \cong P(\mathcal{N}_{0;1} \geqslant c_\alpha) = \alpha$ on rejette H_0 si $t^*_0 \geqslant c_\alpha$

[3] Si l'hypothèse alternative est l'hypothèse de discordance de valeurs H'_1 « $p_{11}p_{22} - p_{12}p_{21} < 0$ » le domaine de rejet est de type $]-\infty, c'_\alpha]$ où c'_α est tel que $P_0(T_0 \leqslant c'_\alpha) \cong P(\mathcal{N}_{0;1} \leqslant c'_\alpha) = \alpha$.

© Dunod. La photocopie non autorisée est un délit.

Section 4

MESURES D'ASSOCIATION ENTRE DEUX VARIABLES

1 Test de Spearman

1.1 Le coefficient de corrélation $\rho(X,Y)$ de Spearman

Le coefficient de corrélation $\rho(X,Y)$ de Spearman sert à mesurer le degré de dépendance qui lie X et Y. Il est le coefficient de corrélation linéaire du couple aléatoire $(F(X),G(Y))$ où F et G désignent respectivement les fonctions de répartition de X et de Y : $\rho(X,Y) = r(F(X),G(Y))$.

Il satisfait aux propriétés suivantes :

i) $-1 \leqslant \rho(X,Y) \leqslant 1$;
ii) X et Y indépendants $\Rightarrow \rho(X,Y) = 0$;
iii) $\rho(X,Y) = 12\ E(F(X) \times G(Y)) - 3$.
iv) $\rho(X,Y) = 1$ [resp. -1] si et seulement si il existe une fonction φ strictement croissante [resp. décroissante] telle que $Y = \varphi(X)$.
v) si φ et Ψ sont deux fonctions strictement croissantes, alors $\rho(\varphi(X),\Psi(Y)) = \rho(X,Y)$
vi) Si (X,Y) suit une loi normale bivariée $\mathcal{N}_2(m; W)$, le coefficient de corrélation linéaire $r(X,Y)$ est lié à $\rho(X,Y)$ par la relation : $r(X,Y) = 2\sin(\pi \times \rho(X,Y)/6)$.

1.2 La variable ρ_s de Spearman

Soit n couples de valeurs numériques $(x_1,y_1),\ldots,(x_n,y_n)$ prises par un couple aléatoire (X,Y) qui suit une loi continue. Au couple (X_i,Y_i) qui prend pour valeur (x_i,y_i) on associe le couple aléatoire d'entiers (R_i,S_i) qui prend pour valeur (r_i,s_i) où r_i désigne le rang de x_i dans $x_1,x_2,\ldots,x_n$ et s_i le rang de y_i dans $y_1,y_2,\ldots,y_n$.

La statistique de Spearman est le coefficient de corrélation linéaire entre R et S soit la *v.a.* $\rho_s = \mathrm{Cov}(R,S)/(\sigma(R) \times \sigma(S))$. Elle est un estimateur convergent de $\rho(X,Y)$ et prend après expérimentation la valeur correspondant au coefficient de corrélation linéaire des rangs ρ^*.

$$\rho^* = \mathrm{Cov}(r,s)/\sigma(r) \times \sigma(s) = \frac{12}{n^3 - n} \sum_{i=1}^{n} \left(r_i - \frac{n+1}{2} \right) \left(s_i - \frac{n+1}{2} \right)$$
$$= 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (r_i - s_i)^2$$

En effet $\overline{r} = \overline{s} = \dfrac{n+1}{2}$, $V(r) = V(s) = \dfrac{n^3 - n}{12}$.

La valeur de ρ^* est comprise entre -1 et 1 soit $-1 \leqslant \rho^* \leqslant 1$:

$\rho^* = 1$ correspond au cas de concordance parfaite ($r_i = s_i\ \forall\, i = 1,\ldots,n$, autrement dit quelque soient les couples (x_i,y_i) et (x_j,y_j) si $x_i < x_j$ alors $y_i < y_j$ et réciproquement) ;

$\rho^* = -1$ correspond au cas de discordance parfaite (si $x_i < x_j$ alors $y_i > y_j$ et réciproquement).

Exemple

On dispose d'un échantillon de 5 couples de valeurs numériques correspondant à la taille x en millions d'euros et à la rentabilité économique y en % des firmes d'un secteur : (0,9, 1,8) ; (1,1, 3,6) ; (0,5, 1,5) ; (1,2, 0,8) ; (1,9, 3,5) et l'on se propose de calculer la valeur prise par ρ_S. Pour cela, réécrivons les 5 couples de résultats selon les valeurs croissantes de la première composante x_i :

Rangs r_i des x_i	1	2	3	4	5
x_i	0,5	0,9	1,1	1,2	1,9
y_i	1,5	1,8	3,6	0,8	3,5
Rangs s_i des y_i	2	3	5	1	4

Avec ce nouvel ordonnancement, on a $r_1 = 1, r_2 = 2, \ldots, r_5 = 5$. Dans le classement par ordre croissant des valeurs y_i : $0{,}8 < 1{,}5 < 1{,}8 < 3{,}5 < 3{,}6$ on constate que $s_1 = 2$, $s_2 = 3$, $s_3 = 5$, $s_4 = 1$, $s_5 = 4$ donc

$$d^2 = \sum_{i=1}^{5} (r_i - s_i)^2 = (2-1)^2 + (3-2)^2 + (5-3)^2 + (1-4)^2 + (4-5)^2 = 16\,.$$

Aussi $\rho^* = 1 - 6 \times d^2/(5^3 - 5) = 0{,}2\,.$

1.3 Distributions de ρ_S sous l'hypothèse H_0 d'indépendance

Sous l'hypothèse de base « X et Y sont indépendants » :

– $E_0(\rho_S) = 0$, $V_0(\rho_S) = 1/(n-1)$ (E_0 et V_0 signifiant espérance et variance sous H_0),

– la distribution de ρ_S est symétrique autour de 0 : $P_0(\rho_S \geqslant t) = P_0(\rho_S \leqslant -t)\ \forall\, t$,

– pour $4 \leqslant n \leqslant 19$ la distribution de ρ_S est tabulée (*cf.* p. 381),

– pour $19 \leqslant n \leqslant 30$ on utilise la variable $T = \rho_s\sqrt{n-2}/\sqrt{1-\rho_S^2}$ qui suit sensiblement la loi de Student à $n-2$ degrés de liberté : $T \cong t_{n-2}$,

– pour $n > 30$ on a l'approximation normale : $\sqrt{n-1}\rho_S \cong \mathcal{N}_{0;1}$.

1.4 Procédure de test

Soit n couples de valeurs numériques $(x_1, y_1), \ldots, (x_n, y_n)$ prises par (X, Y). Partant de cet échantillon, la variable ρ_S permet de tester H_0 « X et Y indépendantes et donc $\rho(X,Y) = 0$ » contre l'hypothèse :

[1] $\overline{H}_0$ « $\rho(X,Y) \neq 0$ » et donc X et Y non indépendantes

[2] H_1 « $\rho(X,Y) > 0$ » autrement dit les valeurs prises par X et Y ont tendance à être *concordantes*

© Dunod. La photocopie non autorisée est un délit.

[3] H_1' « $\rho(X,Y) < 0$ » autrement dit les valeurs prises par X et Y ont tendance à être *discordantes*

Règle de décision.

[1] Dans le cas où l'hypothèse alternative est $\overline{H}_0$, le domaine d'acceptation de H_0 est du type $]-c_{\alpha/2},c_{\alpha/2}[$ où $c_{\alpha/2}$ est défini par $P_0(\rho_S \geqslant c_{\alpha/2}) \leqslant \alpha/2$. On rejette H_0 au profit de $\overline{H}_0$ lorsque la valeur ρ^* prise par ρ_S n'appartient pas à l'intervalle $]-c_{\alpha/2},c_{\alpha/2}[$. Le niveau de signification observé α_c est tel que $\alpha_c/2 = P_0(\rho_S \geqslant |\rho^*|)$.

[2] Le domaine de rejet de H_0 au profit de H_1 est de type $[c_\alpha,1[$ où la valeur c_α est définie par $P_0(\rho_S \geqslant c_\alpha) \leqslant \alpha$ et peut être lue page 381. On rejette H_0 lorsque la valeur numérique ρ^* prise par ρ_S est supérieure à c_α. En effet, il convient de remarquer que si l'hypothèse alternative H_1 est vraie, ρ_S tend à prendre des valeurs strictement positives puisque ρ_S converge en probabilité vers $\rho(X,Y) > 0$. Le niveau de signification observé α_c est tel que $\alpha_c = P_0(\rho_S \geqslant \rho^*)$.

[3] Lorsque l'hypothèse alternative est H_1' « $\rho_S(X,Y) < 0$ » le domaine de rejet de H_0 est donc de type $]-1,c_\alpha'[$ où c_α' est tel que $P_0(\rho_S \leqslant c_\alpha') \leqslant \alpha$. En raison de la symétrie de la distribution, on a $c_\alpha' = -c_\alpha$ où c_α est défini par $P_0(\rho_S \geqslant c_\alpha) \leqslant \alpha$. On rejette H_0 au profit de H_1' lorsque la valeur numérique ρ^* est inférieure à c_α'. Le niveau de signification observé $\alpha_c = P_0(\rho_S \leqslant \rho^*)$.

Exemple

Partant de l'exemple précédent où X désigne la taille ou chiffre d'affaires et Y la rentabilité économique en % des firmes d'un secteur on se propose de tester l'hypothèse d'indépendance H_0 « X et Y indépendants » contre l'hypothèse de concordance H_1 « $\rho(X,Y) > 0$ » avec un risque d'erreur de 5 %. Sous l'hypothèse H_0 , la distribution de ρ_S est tabulée pour $n = 5$ (*cf.* p. 381)).

Règle de décision. Cas [2]. Le domaine de rejet de H_0 est de type $[c_\alpha,1[$ où c_α est tel que $P_0(\rho_S \geqslant c_\alpha) \cong \alpha = 0{,}05$. Par lecture de table on lit $c_\alpha = 0{,}9$. On prend la décision de ne pas rejeter H_0 au profit de H_1 puisque $\rho^* = 0{,}2 < 0{,}9$.

Remarque. Présence d'*ex æquo*. Lorsque la loi H que suit (X,Y) n'est pas continue, les observations $x_1,x_2,\ldots,x_n$ d'une part et $y_1,y_2,\ldots,y_n$ d'autre part peuvent présenter des *ex-aequo*. Une méthode simple consiste à départager les *ex-æquo* (*cf.* p. 216) à l'aide de la table de nombre au hasard puis utiliser la procédure développée ci-dessus.

2 Test de Bloomqvist

Toute distribution discrète pouvant être interpolée par une distribution continue, sauf précision contraire, la loi $H(x,y) = P(X \leqslant x \text{ et } Y \leqslant y)$ que suit le couple (X,Y) est supposée continue.

2.1 Le coefficient de corrélation médial

Notons $\xi_{0.5}$ [resp. $\eta_{0.5}$] la valeur médiane de la loi que suit X [resp. Y].

On a une concordance parfaite de position par quadrant lorsque

$P[Y > \eta_{0.5}/X > \xi_{0.5}] = 1$ et donc $P[Y < \eta_{0.5}/X < \xi_{0.5}] = 1$, c'est-à-dire si

$$m^+(X,Y) = P[(X - \xi_{0.5})(Y - \eta_{0.5}) > 0] = 1\,.$$

De même on a une discordance parfaite de position lorsque le poids de la distribution est entièrement concentrée dans les second et quatrième quadrant, c'est-à-dire si $m^-(X,Y) = P[(X - \xi_{0.5})(Y - \eta_{0.5}) < 0] = 1$.

Le coefficient de corrélation médial $m(X,Y) = m^+(X,Y) - m^-(X,Y)$ est un indicateur de concordance ou de discordance selon qu'il est positif ou négatif.

Il possède les propriétés suivantes :

i) $-1 \leqslant m(X,Y) \leqslant 1$

ii) X et Y indépendants $\Rightarrow m(X,Y) = 0$

iii) ϕ et ψ désignant des fonctions croissantes de $\mathbb{R}$ dans $\mathbb{R}$, on a $m(\phi(X),\psi(Y)) = m(X,Y)$

2.2 Le test de Bloomqvist

À partir d'un échantillon de n couples de valeurs numériques (x_1,y_1), …, (x_n,y_n) prises par (X,Y), la statistique Q_n de Bloomqvist permet de tester :

[cas 1] l'hypothèse H_0 d'indépendance contre l'hypothèse H_1 « $m(X, Y) > 0$ » de concordance de position par quadrant ou plus généralement « $m(X,Y) \leqslant 0$ » contre « $m(X,Y) > 0$ ».

[cas 2] l'hypothèse H_0 d'indépendance contre l'hypothèse H_1' « $m(X,Y) < 0$ » de discordance de position par quadrant.

– **Statistique Q_n**. Un fois déterminées la médiane empirique $m_{e,x}$ des n valeurs $x_1,\ldots,x_n$ et celle $m_{e,y}$ des n valeurs $y_1,\ldots,y_n$ et on compte le nombre n^* de couples (x_i,y_i) tels que $x_i > m_{e,x}$ et $y_i > m_{e,y}$.

Le quotient $q = n^*/n$ est la valeur prise par la variable Q_n qui est un estimateur convergent de $m^+(X,Y)/2 = (m(X,Y) + 1)/4$

– **Distribution de $N_n^* = nQ_n$ sous l'hypothèse H_0 d'indépendance.**

Sous l'hypothèse H_0 « X et Y indépendantes » la distribution de $N_n^* = nQ_n$ est une loi hypergéométrique dont les paramètres dépendent de la parité de n.

Si $n = 2\nu$ est pair, N_n^* suit la loi $\mathcal{H}(n,\nu;0,5)$ loi hypergéométrique de paramètres $(n,\nu;0,5)$:

le domaine des valeurs $\Delta = \{0,1,2,\ldots,\nu\}$; $P_0(nQ_n = k) = C_\nu^k \times C_\nu^{\nu-k}/C_{2\nu}^\nu$ $\forall k \in \Delta$. La loi est symétrique et $E_0(Q_n) = 0.25$; $V_0(Q_n) = 1/16(n-1)$

© Dunod. La photocopie non autorisée est un délit.

Si $n = 2\nu + 1$ est impair, N_n^* suit la loi $\mathcal{H}(n,\nu;\nu/n)$; son domaine des valeurs est $\Delta = \{0,1,2,\ldots,\nu\}$ et $P_0(nQ_n = k) = C_\nu^k \times C_{\nu+1}^{\nu-k}/C_{2\nu+1}^{\nu}$ $\forall k \in \Delta$

On a $E_0(Q_n) = 0.25 \times (1 - 1/n)^2$; $V_0(Q_n) = (1 - 1/n^2)^2/16(n-1)$.

Règle de décision. Dans le [cas 1] où l'on teste H_0 contre l'hypothèse H_1 de concordance de position de quadrant, on cherche le plus petit entier n_α tel que $P_0(nQ_n \geqslant n_\alpha) \leqslant \alpha$ (niveau nominal de signification) puis on rejette H_0 si la valeur n^* prise par nQ_n est supérieure ou égale à n_α.

Dans le [cas 2] où l'on teste H_0 contre l'hypothèse H_1' de discordance, on cherche le plus grand entier n_α tel que $P_0(nQ_n \leqslant n_\alpha) \leqslant \alpha$ puis on rejette H_0 si la valeur n^* prise par nQ_n est inférieure ou égale à n_α.

Exemple

Une équipe de recherche en économie industrielle se demandent si les fabricants de petite taille disposent de par leur plus grande flexibilité et de par leur position dans des segments de marché étroits, d'un avantage concurrentiel vis à vis des firmes de grande taille. *A contrario*, les entreprises de grande taille pourraient bénéficier d'économie d'échelle et d'un plus grand pouvoir de négociation vis à vis des fournisseurs, ce qui les rendrait plus rentables. Pour réaliser l'analyse, l'équipe a à sa disposition le *ROA* (*mesure de rentabilité*) et le nombre de salariés d'un échantillon aléatoire de *16* entreprises d'un même secteur.

Obs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Roa	0,11	–10,7	3,2	0,47	–0,6	–1,2	–0,85	–0,23	2,09	4,36	4,06	– 0,29	1,6	–5,7	2,58	5,26
effectif	213	5	19	162	71	32	57	130	22	21	18	47	38	60	32	60

Pensant qu'il ne peut exister de relation fonctionnelle liant la rentabilité à la taille mesurée par l'effectif de l'entreprise, le choix de la statistique de Bloomqvist s'est imposé car elle permet de dégager des tendances en probabilité par rapport à un indicateur de tendance centrale, à savoir ici, les valeurs médianes de l'effectif et de la rentabilité. On constate que les médianes empiriques des rentabilités et effectifs sont respectivement *0,29 = (0,11+0,47)/2* et *42,5* ; les *16* couples de valeur se répartissent ainsi

effectif ↓ rentabilité →	< 0, 29	> 0,29
> 42,5	6	2
< 42,5	2	6

La statistique $N^* = 16Q_{16}$ prend la valeur 2. Pour tester l'hypothèse d'indépendance H_0 contre celle de discordance « $m(X,Y) < 0$ » on évalue le niveau de signification observé $P_0(N^* \leqslant 2) = (C_8^0 \times C_8^8 + C_8^1 \times C_8^7 + C_8^2 \times C_8^6)/C_{16}^8$ soit *6,6 %*. On constate une tendance en probabilité pour qu'une entreprise qui a un fort effectif soit peu rentable et vice-versa. Autrement dit, peu d'entreprises de ce secteur ont à la fois un effectif supérieur à la médiane des effectifs du secteur et également une rentabilité supérieure à la médiane des rentabilités du secteur.

REPÈRES : Une version adaptée du test du quadrant

Le praticien peut s'intéresser à l'éventuelle concordance de position entre les valeurs de X comprises dans l'intervalle interquartile $(\xi_{0,25}; \xi_{0,75})$ de sa distribution F et les valeurs de Y supérieures à la valeur médiane $\eta_{0,5}$ de sa distribution G. L'indicateur proposé est alors :

$m_P(X, Y) = 2P(\xi_{0,25} \leqslant X \leqslant \xi_{0,75} \text{ et } Y > \eta_{0,5}) - 2P(\xi_{0,25} \leqslant X \leqslant \xi_{0,75} \text{ et } Y < \eta_{0,5})$

Il vérifie les propriétés suivantes :

i) $-1 \leqslant m_P(X, Y) \leqslant 1$.

ii) Si X et Y sont indépendants, $P(Y > \eta_{0,5} / \xi_{0,25} \leqslant X \leqslant \xi_{0,75}) = 0{,}5$ soit $m_P(X, Y) = 0$.

iii) S'il y a concordance $P(Y > \eta_{0,5} / \xi_{0,25} \leqslant X \leqslant \xi_{0,75}) > 0{,}5$ soit $m_P(X, Y) > 0$.

iv) La concordance parfaite correspond à $P(Y > \eta_{0,5} / \xi_{0,25} \leqslant X \leqslant \xi_{0,75}) = 1$ soit $m_P(X, Y) = 1$ et $P(X < \xi_{0,25} \text{ et } Y > \eta_{0,5}) = P(X > \xi_{0,75} \text{ et } Y > \eta_{0,5}) = P(\xi_{0,25} \leqslant X \leqslant \xi_{0,75} \text{ et } Y < \eta_{0,5}) = 0$.

Variable statistique utilisée. Soit n couples de valeurs numériques (x_1, y_1), … , (x_n, y_n) prises par (X, Y). On détermine les premier et troisième quartile empirique $q_{1,x}$ et $q_{3,x}$ des n valeurs $x_1, \ldots, x_n$ ainsi que la médiane empirique $m_{e,y}$ des n valeurs $y_1, \ldots, y_n$ et on compte le nombre n^{**} de couples (x_i, y_i) tels que $q_{1,x} \leqslant x_i \leqslant q_{3,x}$ et $y_i > m_{e,y}$.

Sous l'hypothèse H_0 « X et Y sont indépendants » la statistique $N_n^{**} = nQ_n^*$ qui prend la valeur n^{**} suit la loi hypergéométrique $\mathcal{H}(n; m_o, p)$, l'expression du paramètre m_o dépendant du reste de la division euclidienne de n par 4 : $n = 4n' + h$ avec $0 \leqslant h \leqslant 3$.

On a $m_o = 2n' + h$ et $p = 0{,}5$ si n pair, $p = [n/2]/n$ si n est impair :

$$P_0(N_n^{**} = h) = \frac{C_{[n/2]}^{h} C_{n-[n/2]}^{m_o - h}}{C_n^{m_o}} \quad \forall\, h = 0, 1, 2, \ldots, [n/2]$$ où $[n/2]$ désigne la partie entière de $n/2$. (*Cf.* P-Ch. Pupion *Thèse* 1997)

Règle de décision. Pour tester l'hypothèse de base H_0 « X et Y sont indépendants » contre l'hypothèse de concordance de position H_1 « $m_P(X, Y) > 0$ » avec un niveau de signification α on cherche le plus petit entier h_α tel que $P_0(nQ_n^* \geqslant h_\alpha) \leqslant \alpha$ et l'on rejette l'hypothèse de base au profit de l'hypothèse de concordance lorsque n^{**} est supérieur ou égal à h_α.

Pour tester H_0 contre H_1' « $m_P(X, Y) < 0$ » on cherche le plus grand entier h_α tel que $P_0(nQ_n^* \leqslant h_\alpha) \leqslant \alpha$ et l'on rejette l'hypothèse de base au profit de l'hypothèse de discordance lorsque n^{**} est inférieur ou égal à h_α.

Ainsi, dans l'exemple précédent, si x désigne la taille et que l'on se demande si les toutes grandes et les toutes petites seraient plus rentables (variable y) que les autres il convient de calculer les premier et troisième quartile de x : $q_{1,x} = 21{,}5 = (21 + 22)/2$, $q_{3,x} = 65{,}5 = (60 + 71)/2$, $m_{e,y} = (y_{(3)} + y_{(4)})/2 = 0{,}29$ et l'on constate que $N_6^{**} = 6Q_6^*$ prend la valeur 4.

Le niveau de signification observé est $P_0(N^* \leqslant 4) = (C_8^0 \times C_8^8 + C_8^1 \times C_8^7 + \ldots + C_8^4 \times C_8^4)/C_{16}^8$ soit 69 %.

© Dunod. La photocopie non autorisée est un délit.

effectif ↓ rentabilité →	< 0, 29	> 0,29
$x \leqslant 21{,}5$	1	**3**
$21{,}5 < x < 65{,}5$	4	4
$x \geqslant 65{,}5$	3	**1**

Section 5

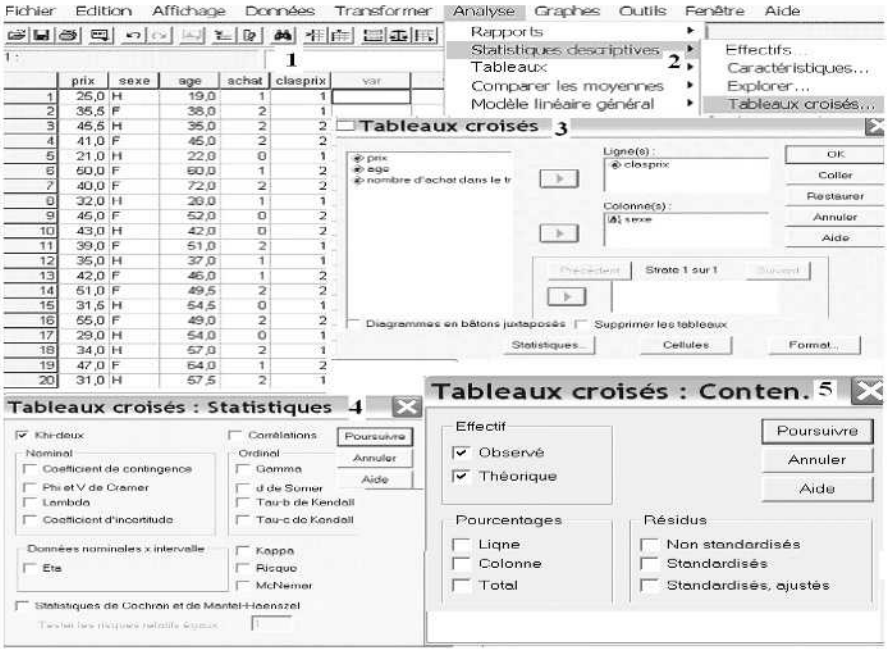
TRAITEMENT SOUS EXCEL ET SPSS

Une société fait un test auprès de 20 individus tirés au hasard et leur demande quel est le prix qu'ils considèrent comme normal pour le parfum qui leur est présenté. Souhaitant affiner son analyse, le service d'études se demande s'il y a un lien entre le prix psychologique annoncé par le client, son âge, son sexe et le nombre de flacons de parfum acheté au cours du semestre.

1 Traitements statistiques avec SPSS

1.1 Test du Khi-deux

On teste H_0 « prix et sexe indépendants » contre $\overline{H}_0$ par la méthode du Khi-deux. Pour cela, il faut regrouper au préalable, les prix psychologiques en classes. Ici les prix ont été répartis en deux classes selon qu'ils sont inférieurs ou supérieurs à 40.



Procédure.

1. Nous avons réparti les prix en deux classes selon qu'ils sont inférieurs ou supérieurs à 40 et défini ainsi une nouvelle variable **clasprix** (=1 si prix < 40 et = 2 dans le cas contraire)

2. Pointer sur (Analyse) puis (Statistiques Descriptives) puis (Tableaux croisés)

3. Dans le menu tableaux croisés on sélectionne les deux variables dont on veut tester l'indépendance et dont les résultats sont regroupés en classe : **clasprix** et **sexe**

4. Cliquant sur (Statistiques) on sélectionne le test du Khi-deux (Khi-deux) et clique sur (Poursuivre)

5. Dans la rubrique (Cellules) on sélectionne (effectif observé) et (effectif théorique)

Résultat Tableau croisé CLASPRIX*SEXE

			SEXE		Total
			F	H	
CLASPRIX	1	Effectif	2	8	10
		Effectif théorique	5	5	10
	2	Effectif	8	2	10
		Effectif théorique	5	5	10
Total		Effectif	10	10	20

Tous les effectifs théoriques sont $\geqslant 5$

Test du Khi-deux

	Valeur	ddl	Sig (bilatéral)
Khi-deux de Pearson	7,2	1	0,007
Nombre d'observations	20		

Interprétation. Le premier tableau est un tableau croisé où figurent la variable *sexe* (que l'on peut numériser en $X = 1$ pour une femme, $X = 0$ pour un homme soit $h = 2$) et la partition en deux classes des prix psychologiques ($k = 2$). Le domaine des valeurs du couple aléatoire est scindé en $h \times k = 4$ classes. Dans le tableau figurent les effectifs « théoriques » $n^*_{ij} = (n_{i\bullet} \times n_{\bullet j})/n$, soit par exemple $n^*_{11} = (n_{1\bullet} \times n_{2\bullet})/n = 10 \times 10/20 = 5$.

Statistique. Sous l'hypothèse H_0 d'indépendance entre *sexe* et *prix psychologique* soit « $p_{ij} = p_{i\bullet} \times p_{\bullet j}$ », la statistique

$$Z = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(N_{ij} - N_{i\bullet} \times N_{\bullet j}/n)^2}{N_{i\bullet} \times N_{\bullet j}/n} \cong \chi^2_{(h-1)(k-1)} = \chi^2_1$$

car $h = 2$ et $k = 2$. Z prend la valeur $z = 7{,}2$.

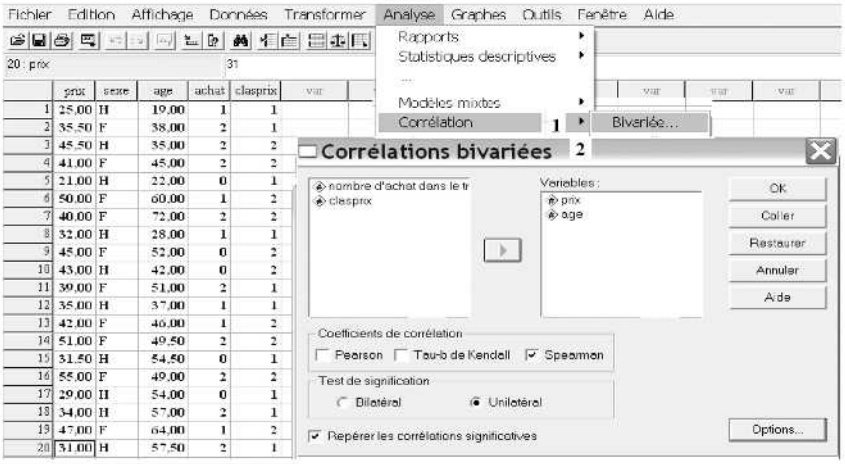
Règle de décision. Le niveau de signification observé du test ou risque d'erreur associé à la décision de rejeter H_0 est égal à 0,007 : $\alpha_c = P_0(Z > 7{,}2) \cong \quad P(\chi^2_1 > 7{,}2) = 0{,}007$. Aussi peut-on rejeter avec un risque d'erreur négligeable l'hypothèse d'indépendance. Le

© Dunod. La photocopie non autorisée est un délit.

prix psychologique est lié au sexe de la personne interrogée. Le logiciel fournit également le niveau de signification observé du test exact de Fisher dans le cas d'un tableau 2×2.

1.2 Test de Spearman

On peut tester à l'aide du test de Spearman l'hypothèse H_0 « *prix* et *âge* sont indépendants » contre l'hypothèse H_1 « les valeurs prises par *prix* et *âge* ont tendance à être concordantes » ou H_1' « les valeurs prises par *prix* et *âge* ont tendance à être discordantes ».



	prix	sexe	age	achat	classprix
1	25,00	H	19,00	1	1
2	35,50	F	38,00	2	1
3	45,50	H	35,00	2	2
4	41,00	F	45,00	2	2
5	21,00	H	22,00	0	1
6	50,00	F	60,00	1	2
7	40,00	F	72,00	2	2
8	32,00	H	28,00	1	1
9	45,00	F	52,00	0	2
10	43,00	H	42,00	0	2
11	39,00	F	51,00	2	1
12	35,00	H	37,00	1	1
13	42,00	F	46,00	1	2
14	51,00	F	49,50	2	2
15	31,50	H	54,50	0	1
16	55,00	F	49,00	2	2
17	29,00	H	54,00	0	1
18	34,00	H	57,00	2	1
19	47,00	F	64,00	1	2
20	31,00	H	57,50	2	1

Procédure.

1. Pointer sur (Analyse) puis (Corrélation) puis (Bivariée).

2. Dans le menu Corrélations bivariées on sélectionne les variables prix et âge et on opte pour le test de (Spearman) avec un test (Unilatéral).

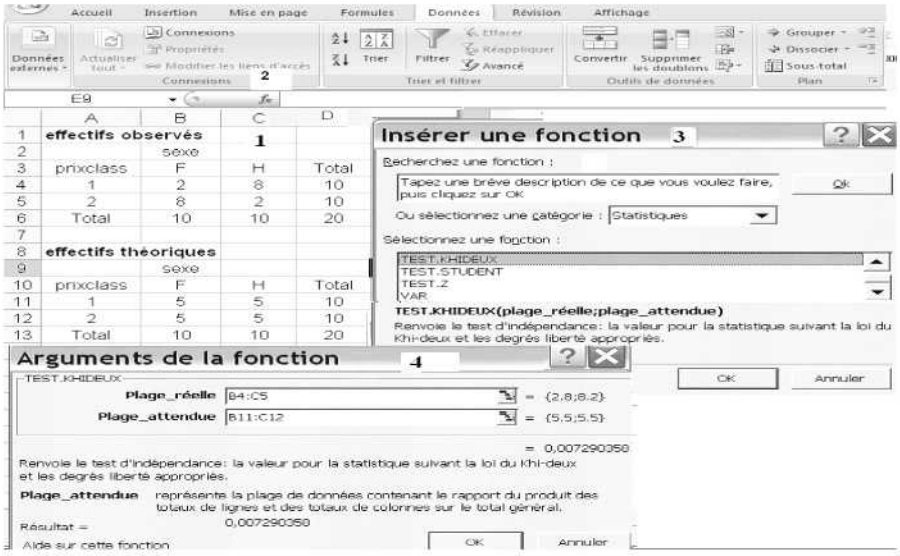
Résultats

Corrélations			AGE
Rhô de Spearman	PRIX	Coefficient de corrélation	0,262
		Sig. (unilatérale)	0,133
		N	20

À chaque couple (x_i, y_i) où x_i est le prix et y_i l'âge associés à la i-ème observation on associe le couple d'entiers (r_i, s_i) où r_i désigne le rang de x_i dans $x_1, x_2, \ldots, x_{20}$ et s_i le rang de y_i dans $y_1, y_2, \ldots, y_{20}$. La variable de Spearman ρ_S prend pour valeur $\rho^* = \text{Cov}(r,s)/\sigma(r) \times \sigma(s) = 0{,}262$ qui est le coefficient de corrélation linéaire entre les rangs des observations ***r*** définis sur la variable prix et les rangs ***s*** définis sur la variable âge. Dans le cas de test unilatéral, c'est l'ordinateur qui sélectionne l'hypothèse alternative la plus plausible, soit ici l'hypothèse H_1 « les valeurs prises par *prix* et *âge* ont tendance à être concordantes » puisque l'estimation empirique ρ_S^* est positive. L'ordinateur calcule le niveau de signification observé du test $\alpha_c = P_0(\rho_S \geqslant 0{,}262) = 0{,}133$. Pour un niveau de signification $\alpha = 5\ \%$ on a $\alpha_c > \alpha$ et on conclut au non rejet de H_0.

2 Traitements statistiques avec EXCEL

Pour effectuer un test d'indépendance du Khi-deux entre X le sexe et Y le prix il faut au préalable faire les tableaux croisés des données et des effectifs « théoriques » obtenus par le calcul suivant « *effectif théorique de la classe* = (*total des effectifs de la ligne* × *total effectifs de la colonne*) / *taille de l'échantillon* ». La taille de l'échantillon correspond à la somme totale des effectifs.



Procédure.

1. Réaliser le tableau croisé des effectifs observés et en dessous le tableau croisé des effectifs théoriques (ici effectif théorique de prixclas = 1 et sexe F est $10 \times 10/20 = 5 \ldots$).

2. On clique sur (Fx) et on sélectionne (Fonction).

3. Dans le menu **Insérer une fonction** on sélectionne dans catégorie (Statistiques) et on déroule pour sélectionner la fonction (Test Khi-deux) et OK.

4. Dans le menu **Argument de la fonction** on sélectionne dans (plage_réelle) les références des cellules contenant les effectifs observés et dans (plage_attendue) les références des cellules contenant les effectifs théoriques.

Il apparaît dans le coin droit le niveau de signification du test soit 0,00729. Autrement dit, avec un risque d'erreur négligeable on décide de rejeter H_0.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 360)

Exercice 1

Un chercheur en économie industrielle s'intéressant au secteur Q se demande si les grandes firmes disposent de par leur taille, d'un avantage concurrentiel vis-à-vis des firmes de petite taille. Pour vérifier son intuition, le chercheur a à sa disposition X le *ROA* (le *return on assets*) et Y l'effectif d'un échantillon de 8 entreprises fabriquant des chaussures.

Observations	1	2	3	4	5	6	7	8
X	0,11	– 10,79	3,18	0,47	– 0,60	– 1,18	– 0,85	– 0,23
Y	213	5	19	162	71	32	57	130

Tester à l'aide du rhô de Spearman et avec un niveau de signification de 5 % l'hypothèse H_0 « X et Y indépendants » contre H_1 « les valeurs prises par X et Y ont tendance à être concordantes ».

Exercice 2

Une entreprise souhaite lancer un nouveau produit. Elle confie l'étude de marché à une société de sondage. Lors du sondage, on demande à 400 clients potentiels, s'ils sont prêts à acheter ce nouveau produit. Les résultats en fonction de l'âge des personnes interrogées sont les suivants :

Opinion / **Âge**	**oui**	**non**	**ne peut pas se prononcer**
< 30 ans	65	27	8
de 30 à 45 ans	50	19	11
de 45 à 60 ans	35	24	11
plus de 60 ans	50	80	20

L'opinion est elle indépendante de l'âge ? (Prendre un risque d'erreur de première espèce de 5 %)

14 TESTS D'AJUSTEMENT

Dans ce chapitre figurent différents procédés qui permettent de tester si la distribution d'un phénomène aléatoire est régie par une loi déterminée F ou tout au moins appartient à une famille de lois F_θ.

Section 1 ■ Test d'ajustement de Kolmogorov-Smirnov
Section 2 ■ Test d'ajustement du Khi-deux
Section 3 ■ Tests d'appartenance à une distribution normale
Section 4 ■ Outliers ou recherche de valeurs discordantes
Section 5 ■ Traitements avec Excel et SPSS

À partir d'un échantillon de n valeurs numériques $x_1, x_2, \ldots, x_n$ prises par une *v.a.* X, il s'agit de tester l'hypothèse H_0 « X suit une loi donnée F » contre $\overline{H}_0$ « X ne suit pas la loi F ». Autrement dit, $G(t) = P(X \leqslant t)$ désignant la fonction de répartition de X et $F(t)$ étant une fonction de répartition donnée et connue, il s'agit de tester « X suit la loi F, c'est-à-dire $G(t) = F(t)$ » contre $\overline{H}_0$ « la distribution G de X est distincte de F : $G(t) \neq F(t)$ ».

Exemple

Le résultat X d'une expérience envisagée est une valeur entière positive ou nulle. Dix expériences réalisées dans des conditions identiques ont donné les résultats suivants : 1, 0, 2, 0, 6, 0, 1, 1, 0, 6. À partir de ces résultats, on veut tester l'hypothèse H_0 « X suit la loi de Poisson de paramètre $\lambda = 1{,}7$ » contre « X ne suit pas la loi de Poisson de paramètre $\lambda = 1{,}7$ ».

© Dunod. La photocopie non autorisée est un délit.

On peut également vouloir tester la nature de la loi que suit X, c'est-à-dire l'hypothèse H_0' « X suit une loi qui appartient à une famille donnée F_θ » contre l'hypothèse « la loi que suit X n'est pas de type F_θ ». Dans l'exemple précédent, on peut ainsi tester H_0' « X suit une loi de Poisson (dont la valeur du paramètre λ n'est pas connue) » contre $\overline{H}_0'$ « la loi que suit X n'est pas une loi de Poisson ».

Notations communes à ce chapitre. F désigne la fonction de répartition de la loi qu'est supposé suivre X sous H_0 : $F(t) = P_0(X \leqslant t)$. $G_n^*(t)$ désigne la fonction de répartition empirique des n valeurs numériques $x_1, x_2, \ldots, x_n$ prises par X : $G_n^*(t) =$ (nombre de valeurs x_i telles que $x_i \leqslant t)/n$. (*Cf.* chapitre 1).

Section 1

TEST D'AJUSTEMENT DE KOLMOGOROV-SMIRNOV

1 Ajustement sur une loi continue

Il sert à tester H_0 « X suit la loi continue F » contre $\overline{H}_0$ « la distribution continue G de la loi que suit X est distincte de F : $G \neq F$ ».

Variable statistique utilisée. La méthode de Kolmogorov-Smirnov consiste à déterminer le plus grand écart absolu d qui sépare $G_n^*(t)$ de $F(t)$:

$$d = \sup_t |G_n^*(t) - F(t)|$$

On obtient d en ordonnant les valeurs prises par l'échantillon par ordre de valeurs croissantes : $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, puis en utilisant la propriété « $d = \max\{d^+, d^-\}$ où $d^+ = \max_i\{(i/n) - F(x_{(i)})\}$ et $d^- = \max_i\{F(x_{(i)}) - (i-1)/n\}$ pour $i = 1, 2, \ldots, n$ ». La variable aléatoire qui prend la valeur d sera notée D.

Distribution de D sous H_0. Sous l'hypothèse H_0,

i) la distribution de D ne dépend pas de la loi que suit X. Elle est tabulée pour des valeurs de n n'excédant pas 100 (*cf.* p. 386).

ii) Pour $n > 35$, on peut utiliser l'approximation $P_0(D \geqslant t) \cong 2e^{-2nt^2}$. Autrement dit, en choisissant $d_\alpha = [\ln(2/\alpha)/2n]^{1/2}$, on a $P_0(D \geqslant d_\alpha) \cong \alpha$.

Procédure d'utilisation. Pour un risque d'erreur α, on cherche d_α tel que $P_0(D \geqslant d_\alpha) \cong \alpha$. Si la valeur d prise par D est supérieure à d_α, on rejette l'hypothèse H_0 « X suit la loi continue F » au profit de $\overline{H}_0$. Le niveau de signification observé du test est $\alpha_c = P_0(D \geqslant d)$; si α_c est inférieur à α on rejette H_0.

Exemple

On dispose de 7 valeurs prises par une *v.a.* X et qui ont été ordonnées par ordre de valeurs croissantes : 90,87 ; 92,55 ; 96,20 ; 98,98 ; 100,42 ; 101,58 ; 106,82 . Les 7 valeurs

x_i ayant été ordonnées par ordre de valeurs croissantes on a $x_{(1)} = 90{,}87$, $x_{(2)} = 90{,}27, \ldots x_{(7)} = 106{,}82$ et $G_7^*(x_i) = i/7$. On se propose de tester H_0 « X suit la loi normale $\mathcal{N}(98{,}5; 8^2)$ » contre $\overline{H}_0$ en utilisant le test d'ajustement de Kolmogorov-Smirnov. Sous l'hypothèse nulle, on a $F(x_i) = \Phi[(x_i - 98{,}5)/8]$ et donc le tableau suivant où $d_i^+ = i/n - F(x_i)$ et $d_i^- = F(x_i) - (i-1)/n$ avec $n = 7$:

x_i	90,87	92,55	96,2	98,98	100,42	101,58	106,82
$F(x_i)$	0,17	0,229	0,387	0,524	0,595	0,65	0,851
i/n	0,143	0,286	0,429	0,571	0,714	0,857	1,000
$d_i^+ = i/n - F(x_{(i)})$	-0,027	0,057	0,042	0,047	0,119	0,207	0,149
$(i-1)/n$	0	0,143	0,286	0,429	0,571	0,714	0,857
$d_i^- = F(x_{(i)}) - (i-1)/n$	0,170	0,086	0,101	0,095	0,024	– 0,064	– 0,006

On constate que la variable D prend pour valeur $d = \max\{d_i^+, d_i^-\} = 0{,}207$.

Prenant un risque d'erreur de première espèce $\alpha = 0{,}05$ on constate par lecture de tableau (*cf.* p. 386) $P_0(D > 0{,}486) \leqslant 0{,}05$. Dans l'exemple considéré la valeur $d = 0{,}207$ étant inférieure à 0,486, on ne peut pas rejeter H_0.

2 Ajustement sur une loi discrète

La statistique de Kolmogorov-Smirnov permet également de tester l'hypothèse H_0 « X suit la loi discrète F où F est caractérisée par son support $\{x_1^*, x_2^*, \ldots, x_k^*\}$ (où $x_1^* < x_2^* < \ldots < x_k^*$ avec éventuellement k infini) et $P(X = x_i^*) = p_i$ ».

Dans ce cas on considère la *v.a.* D' qui prend la valeur $d' = \sup_i |G_n(x_i^*) - F(x_i^*)|$ où G_n désigne la fonction de répartition empirique de l'échantillon. Ayant l'inégalité $P_0(D' \geqslant d_\alpha) \leqslant P_0(D \geqslant d_\alpha)$ on cherche d_α tel que $P_0(D \geqslant d_\alpha) \cong \alpha$ et on rejette H_0 lorsque $d' \geqslant d_\alpha$.

Section 2

TEST D'AJUSTEMENT DU KHI-DEUX

Il permet de tester H_0 « X suit la loi discrète ou continue F » contre $\overline{H}_0$ « la distribution G de la loi que suit X est distincte de F : $G \neq F$ ».

Pour cela, on partage le domaine Δ_X des valeurs possibles de la variable aléatoire X en h classes $C_1, C_2, \ldots, C_h$. Les n valeurs $x_1, x_2, \ldots, x_n$ prises par X étant réparties entre ces h classes, on considère les effectifs respectifs $n_1, n_2 \ldots, n_h$ des classes $C_1, \ldots, C_h$.

Classes	C_1	C_2	...	C_h
Effectifs	n_1	n_2	...	n_h

© Dunod. La photocopie non autorisée est un délit.

La statistique utilisée. Notons N_i le nombre aléatoire de valeurs de X qui, sur un échantillon de taille n, appartiendront à la classe C_i. Sous l'hypothèse H_0 « X suit la loi F », la fonction de répartition $F(x) = P(X \leqslant x)$ est connue et on détermine les probabilités $p_i = P_0(X \in C_i)$: $p_1 = P_0(X \in C_1)$, $p_2 = P_0(X \in C_2), \ldots,$ $p_h = P_0(X \in C_h)$ et les effectifs théoriques correspondant $n \times p_i$.

Sous l'hypothèse H_0, la *v.a.* $Z = \sum_{i=1}^{h} \frac{(N_i - np_i)^2}{np_i}$ qui prend la valeur $z = \sum_{i=1}^{h} \frac{(n_i - np_i)^2}{np_i}$ après expérimentation, suit approximativement une loi khi-deux à $(h-1)$ degrés de liberté lorsque $np_i \geqslant 5 \; \forall i = 1, \ldots, h$: $Z \cong \chi^2_{h-1}$ (si nécessaire il faut regrouper des classes adjacentes de telle façon que la condition $np_i \geqslant 5$ soit vérifiée pour chaque i).

Procédure de test. Sous l'hypothèse alternative $\overline{H}_0$, la variable statistique Z tend à prendre des valeurs élevées. Le domaine de rejet de H_0 est de type $[c_\alpha, \infty[$. Pour un niveau de signification α, on cherche c_α tel que $P_0(Z \geqslant c_\alpha) \cong P(\chi^2_{h-1} \geqslant c_\alpha) = \alpha$. Lorsque la valeur z est supérieure à c_α on rejette H_0.

Exemple

À partir de l'échantillon de 100 valeurs prises par X et présentées ci-dessous,

Valeurs de X	0	1	2	3	Total
Effectifs n_i	25	50	20	5	100

on se propose de tester l'hypothèse H_0 « X suit la loi de Poisson de paramètre $\lambda = 1$ ». Pour cela on répartit les 100 valeurs entre les 4 classes que sont naturellement $(X = 0)$, $(X = 1)$, $(X = 2)$ et $(X \geqslant 3)$ puis on calcule successivement les probabilités $p_i = P_0(X \in C_i)$, les effectifs théoriques np_i et les termes $(n_i - np_i)^2/np_i$. Ainsi, $p_1 = P_0(X \in C_1) = P(X = 0 / X \sim> \mathcal{P}(1)) = \mathrm{e}^{-1} \times 1^0/(0!) = 0{,}368$ d'où $np_1 = 100 \times 0{,}368$ et $(25 - 36{,}8)^2/36{,}8 = 3{,}784$; $p_2 = P_0(X \in C_2) = P(X = 1 / X \sim> \mathcal{P}(1)) = \mathrm{e}^{-1} \times 1^1/(1!) = 0{,}368$ d'où $np_2 = 100 \times 0{,}368$ et $(n_2 - np_2)^2/np_2 = (50 - 36{,}8)^2/36{,}8 = 4{,}735$ etc. :

Valeurs de X	0	1	2	$X \geqslant 3$	Total
$p_i = P_0(X \in C_i)$	0,368	0,368	0,184	0,080	1
Effectif théorique np_i	36,80	36,80	18,40	8,00	100
Effectif réel n_i	25	50	20	5	100
$(n_i - np_i)^2/np_i$	3,784	4,735	0,139	1,125	9,783

La statistique Z prend la valeur $z = 9{,}783$. Sous H_0, la *v. a.* Z suit sensiblement la loi khi-deux à $(4-1) = 3$ degrés de liberté. Prenant un risque d'erreur de première espèce

$\alpha = 0{,}05$ on cherche c_α tel que $0{,}05 = P_0(Z \geqslant c_\alpha) \cong P(\chi^2_3 \geqslant c_\alpha)$. On lit page 377 $c_\alpha = 7{,}82$. La valeur $z = 9{,}783$ prise par Z étant supérieure à 7,82 on décide de rejeter H_0. Le niveau de signification observé du test est $\alpha_c = P_0(Z \geqslant z) \cong P(\chi^2_3 \geqslant 9{,}78) = 0{,}02$ inférieur à $\alpha = 0{,}05$.

Remarque. Plaçons-nous sous l'hypothèse « X suit une loi F_θ. où le vecteur paramètre $\theta = (\theta_1, \ldots, \theta_p)$ est inconnu. Lorsque l'estimateur $\overline{\theta}$ est obtenu par la méthode du maximum de vraisemblance à partir des n valeurs de l'échantillon, on établit que $Z \xrightarrow{\text{loi}} \chi^2_{(h-1-p)}$ lorsque $n \to \infty$ (on dit que l'on perd un degré de liberté par paramètre estimé). Quand les valeurs de l'échantillon sont regroupées par classes, l'estimateur est déduit des effectifs des h classes et on établit que $Z \xrightarrow{\text{loi}} \chi^2_{(h-1-p)} + \sum_{i=1}^{p} a_i \chi_1^{2^{(i)}}$ où les variables khi-deux sont indépendantes et les constantes a_i vérifient $0 \leqslant a_i \leqslant 1$. Autrement dit $Z \cong \chi^2_{(h-1-p)} + \sum_{i=1}^{p} a_i \chi_1^{2^{(i)}} \leqslant \chi^2_{(h-1)}$. Si on cherche le domaine de rejet de H_0 de type $[c_\alpha, \infty[$ on cherche donc c_α tel que $P(\chi^2_{(h-1)} \geqslant c_\alpha) = \alpha$ car $P_0(Z > c_\alpha) \leqslant P(\chi^2_{(h-1)} > c_\alpha)$. Dans l'exemple précédent où $p = 1$, on estime le paramètre λ de la loi supposée être de Poisson par la moyenne de l'échantillon $\overline{x} = 0{,}95$ (puisque $\overline{X}$ est un estimateur de $m = \lambda$) et on teste l'hypothèse « X suit une loi de Poisson dont la valeur estimée du paramètre est 0,95 » en prenant cette estimation pour paramètre.

Section **3**

TESTS D'APPARTENANCE À UNE DISTRIBUTION NORMALE

Il s'agit de tester à partir d'un échantillon de taille n l'hypothèse H_0 « X suit une loi normale $\mathcal{N}(m, \sigma^2)$ où les valeurs de m et σ sont inconnues » contre $\overline{H}_0$. Pour cela il est conseillé d'utiliser prioritairement les deux tests de symétrie ci-après puis le test du kurtosis.

1 Tests de symétrie

1.1 Test du Skewness

On utilise la variable $\gamma_{1,n} = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{3/2}}$ qui, sous l'hypothèse H_0, a une distribution indépendante de m et σ et converge en probabilité vers 0. Sous H_0 on a $P_0(\gamma_{1,n} > t) = P_0(\gamma_{1,n} < -t)\ \forall t$

© Dunod. La photocopie non autorisée est un délit.

Si la valeur b_1 prise par $\gamma_{1,n}$ est trop éloignée de 0, autrement dit si « $P(|\gamma_{1,n} \geqslant (|b_1|)$ a une valeur trop faible » (par exemple $\leqslant 5\ \%$) on décide de rejeter H_0 (*Cf.* table des valeurs critiques page 387).

1.2 Test de symétrie des rangs signés de Wilcoxon

Soit un n-échantillon $X_1, X_2, \ldots, X_n$ *iid* d'une loi continue F. À partir de ce n-échantillon, privilégiant par exemple X_n, on construit le pseudo-échantillon Y_1, $Y_2, \ldots, Y_{n-1}$ défini comme suit :

$$Y_i = X_i - \overline{X} - (X_n - \overline{X})/(1 + n^{1/2}) \text{ où } \overline{X} = (X_1 + X_2 + \ldots + X_n)/n$$

qui est tel que $E(Y_i) = 0$, $V(Y_i) = \sigma^2$ et $\mathrm{Cov}(Y_i, Y_j) = 0$.

Sous l'hypothèse H_0, Y_1, $Y_2, \ldots, Y_{n-1}$ peut être considéré comme un échantillon *iid* de la loi $N(0, \sigma^2)$ qui est symétrique autour de 0. Il suffit alors d'appliquer le test de rangs signés de Wilcoxon $W^+_{n-1}(Y)$ à ces $(n-1)$ variables pour s'assurer ou non de la symétrie (*cf.* p. 194).

Exemple

Un analyste financier, lisant une étude réalisée par un journal économique, se demande, à la lecture des ratios d'indépendance financière de 20 entreprises, si l'intervalle de confiance de la valeur moyenne de ce ratio proposé pour l'ensemble des entreprises de ce secteur est valide.

Entreprise	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
valeur du ratio en %	71	63	62	56	45	41	32	31	27	25	22	15	7	9	10	11	16	18	19	20

L'échantillon étant de petite taille, la construction de cet intervalle de confiance repose sur l'hypothèse que la distribution du ratio est normale, alors même que la valeur du ratio est nécessairement comprise entre 0 et 1. Ceci laisse perplexe notre lecteur qui décide de tester l'hypothèse H_0 « la distribution du ratio est sensiblement normale ».

Entreprise	1	2	3	4	…	19	20	Total
valeur du ratio en %	*71*	*63*	*62*	*56*	…	*19*	*20*	600
$(x_i - \overline{x})^2$	1 681	1 089	1 024	676		121	100	7 436
$(x_i - \overline{x})^3$	68 921	35 937	32 768	17 576		–13 331	–1 000	112 788
$(x_i - \overline{x})^4$	2 825 761	1 185 921	1 048 576	456 976		14 641	10 000	6 486 380

La moyenne $\overline{x} = 30$, aussi la valeur empirique du Skewness est-elle $\gamma_{1,20} = (112788/20)/(7436/20)^{1,5}$ étant égale à 0,786, avec un risque d'erreur inférieur à 10 % on peut rejeter H_0 (valeur critique sur la table 0,777 *cf.* p. 387).

Pour confirmer ce résultat discutable, puisque 0,786 est proche de 0,77 on utilise le test de symétrie des rangs. En privilégiant $x_{20} = 20$, on détermine les 19 valeurs $y_i = x_i - \overline{x} - (x_{20} - \overline{x})/(1 + 20^{1/2}) = x_i - 30 + 1{,}82$ puis on détermine la valeur w prise par W_{19}^+ à partir du tableau ci-dessous, les valeurs de y_i sont, pour simplifier, arrondies à la valeur entière la plus proche :

y_i	42,8	34,8	33,8	27,8	16,8	12,8	3,8	2,8	–1,2	–3,2
Rang de $\lvert y_i \rvert$	19	18	17	16	11	9	4	2	1	3

y_i	–6,2	–13,2	–21,2	–19,2	–18,2	–17,2	–11,2	–10,2	–9,2
Rang de $\lvert y_i \rvert$	5	10	15	14	13	12	8	7	6

w la somme des rangs r_i^a des observations à valeur positives est égale à :

$$w = 19 + 18 + 17 + 16 + 11 + 9 + 4 + 2 = 96\,.$$

Sous l'hypothèse de symétrie et puisque $n \geqslant 15$, on peut utiliser l'approximation normale avec correction de continuité : pour tout $h \in \Delta_{W^+}$, $P_0(W_n^+ \leqslant h) \cong \Phi[\theta(h)]$ où $\theta(h) = [h + 0.5 - n(n+1)/4]/[n(n+1)(2n+1)/24]^{1/2}$

On constate que $P_0(W_{19}^+ \geqslant 96) \cong 1 - P_0(W_{19}^+ \leqslant 95) = 1 - \Phi[\theta(95)] = 1 - \Phi(0{,}02)$ $\cong 0{,}492$. Avec un niveau de signification observé de 98 %, on accepte l'hypothèse de symétrie.

2 Test du kurtosis

On utilise la variable $B_{2,n} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4 \Big/ \left[\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2$ qui, sous l'hypothèse H_0, a une distribution indépendante de m et σ et converge en probabilité vers 3 : « $B_{2,n} \xrightarrow{\text{pr}} 3$ lorsque $n \longrightarrow \infty$ ».

Si la valeur b_2 prise par $B_{2,n}$ est trop éloignée de 3, autrement dit si « $b_2 > 3$ et $P_0(B_{2,n} \geqslant b_2)$ a une valeur trop faible » ou si « $b_2 < 3$ et $P_0(B_{2,n} \leqslant b_2)$ a une valeur trop faible » (par exemple $\leqslant 5$ %) on décide de rejeter H_0.

Cf. distribution de $B_{2,n}$ page 000.

Exemple précédent (suite)

La valeur empirique de la variable de kurtosis $B_{2,n}$ étant égale à 2,35 (soit un coefficient de kurtosis $2{,}35 - 3 = -0{,}65$), le niveau de signification observé est égal à $2P_0(B_{2,20} > 2{,}35) > 0{,}2 \times 2$, on décide de ne pas rejeter H_0.

© Dunod. La photocopie non autorisée est un délit.

3 Statistique de Kolmogorov-Smirnov

Elle permet également de tester l'hypothèse H_0' contre $\overline{H}_0'$. Les valeurs des paramètres m et σ étant respectivement estimés par $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ on considère l'écart $d = sup_t|G_n^*(t) - \Phi[(t - \overline{x})/s]|$ où Φ désigne la fonction de répartition de la loi normale centrée réduite.

Les valeurs prises par l'échantillon étant ordonnées de façon croissante : $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, on considère

$d = \max\{d^+, d^-\}$ où

$d^+ = \max_i\{(i/n) - \Phi[(x_{(i)} - \overline{x})/s]\}$ et $d^- = \max_i\{\Phi[(x_{(i)} - \overline{x})/s] - (i-1)/n\}$.

Règle de décision. On rejette H_0 avec un niveau de signification $\alpha = 15\ \%$ [resp. 10 %, 5 %, 1 %] lorsque

$$(n^{1/2} + 0{,}85 \times n^{-1/2} - 0{,}01)d > 0{,}775 \quad \text{[resp. 0,819 ; 0,895 ; 1,035].}$$

Section 4

OUTLIERS OU RECHERCHE DE VALEURS DISCORDANTES

Soit la réalisation $x_1, x_2, \ldots, x_n$ d'un n-échantillon. On émet l'hypothèse H_0 selon laquelle $X_1, X_2, \ldots, X_n$ est un échantillon *iid* d'une distribution inconnue. Lorsque une ou plusieurs valeurs extrêmes paraissent discordantes, par exemple $x_{(1)}$ trop petite ou $x_{(n)}$ trop grande, on doit s'interroger sur la validité de l'hypothèse H_0.

En présence d'un échantillon de grande taille on peut aisément détecter les valeurs discordantes. Par contre, lorsque l'on ne dispose que d'un échantillon de petite taille, il faut au préalable déterminer la famille F de lois à laquelle semble appartenir l'échantillon, une fois amputé des valeurs qui paraissent discordantes.

Notations. $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$; $s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$

1 Test de valeur discordante d'un échantillon dont la nature de la distribution est inconnue

1.1 Cas d'un échantillon de grande taille

Si la valeur $x_{(n)}$ prise par une *v.a.* X_i semble trop grande on s'intéresse à la valeur $1 - \Phi\left[\sqrt{n}(x_{(n)} - \overline{x}_{n-1})/s_{n-1}\right]$ où Φ désigne la fonction de répartition de la loi normale standard. $\overline{x}_{n-1}$ et s_{n-1} sont la moyenne et l'écart-type de la série amputée de $x_{(n)}$.

Lorsque cette probabilité a une valeur négligeable on peut considérer que $x_{(n)}$ est une valeur discordante, autrement dit l'élément e_i dont $x_{(n)}$ est la mesure n'appartient pas à la population étudiée.

Justification. $\sqrt{n-1}\,\dfrac{X_i - \overline{x}_{n-1}}{s_{n-1}} \cong \mathcal{N}_{0;1}$

$$\Rightarrow P(X_i \geqslant x_{(n)} = x_i) \cong 1 - \Phi\left[\sqrt{n}(x_{(n)} - \overline{x}_{n-1})/s_{n-1}\right]$$

Lorsque la valeur $x_{(1)}$ paraît discordante car trop petite, on se ramène au cas précédent en considérant $-x_1, -x_2, \ldots, -x_n$.

1.2 Cas d'un échantillon de petite taille

Si la valeur prise par une *v.a.* semble trop grande, l'échantillon sera indexé pour que cette valeur soit x_n. Sous l'hypothèse « $X_1, X_2, \ldots, X_n$ est un échantillon *iid* d'une loi continue mais inconnue F », les $[n/2]$ *v.a.* $Y_1 = X_1 - X_n$, $Y_2 = X_2 - X_{n-1}$, $Y_3 = X_3 - X_{n-3}$, …, $Y_{[n/2]} = X_{[n/2]} - X_{n-[n/2]}$ peuvent être considérées comme étant un échantillon *iid* d'une loi symétrique de moyenne nulle. La variable $W_n^+(Y_1, Y_2, \ldots, Y_{[n/2]})$ des rangs signés de Wilcoxon doit alors suivre la loi tabulée page 384. Si W_n^+ prend une valeur trop faible on acceptera l'hypothèse selon laquelle x_n est une valeur discordante (*Cf.* test p. 194).

2 Tests de valeurs discordantes sur une distribution supposée normale $\mathcal{N}(m, \sigma^2)$

Soit la réalisation $x_1, x_2, \ldots, x_n$ d'un n-échantillon. On émet l'hypothèse H_0 selon laquelle $X_1, X_2, \ldots, X_n$ est bien un échantillon *iid* d'une distribution normale $N(m,\sigma^2)$ où les valeurs de m et σ sont inconnues.

Cas d'une valeur discordante. Si compte tenu des valeurs numériques $\overline{x}$ et s, $x_{(n)}$ paraît être une valeur discordante car trop grande, on considère la valeur numérique $t = (x_{(n)} - \overline{x})/s$ prise par la statistique $T_{(n)} = (X_{(n)} - \overline{X})/S$.

© Dunod. La photocopie non autorisée est un délit.

Lorsque le niveau de signification observé $P_0(T_{(n)} \geqslant t)$ a une valeur trop faible on rejette l'hypothèse H_0 « $X_1, X_2, \ldots, X_n$ est un échantillon *iid* d'une loi normale ».

Un procédé classique de simulation permet d'obtenir une estimation de $P_0(T_{N.(n)} \geqslant t)$. Plus simplement on peut utiliser la majoration suivante :

$$P_0(T_{(n)} \geqslant t) \leqslant nP\left(\boldsymbol{t}_{n-2} > \sqrt{\frac{n(n-2)t^2}{(n-1)^2 - nt^2}}\right), \text{ où } \boldsymbol{t}_{n-2} \text{ var. de Student à } (n-2)$$

degrés de liberté, cette inégalité étant une égalité lorsque $t^2 \geqslant (n-1)(n-2)/2n$.

Lorsque la valeur $x_{(1)}$ paraît discordante car trop petite, on se ramène au cas précédent en considérant $-x_1, -x_2, \ldots, -x_n$ et donc $T_{(1)} = (\overline{X} - X_{(1)})/S$.

Exemple

On dispose de 10 valeurs numériques, classées par ordre croissant, qui sont la réalisation d'un échantillon supposé *iid* d'une loi continue dont le support est $]-\infty;\infty[$: 4,45 , 16,35 , 26,47 , 28,98 , 30,77, 32,60 , 34,20 , 36,66 , 48,33 , 74,19. On se propose de tester H_0 « on dispose bien d'un échantillon *iid* d'une loi normale » contre $\overline{H}_0$ après avoir remarqué que la valeur $x_{(10)}$ semble discordante. La statistique $T_{(10)}$ prend la valeur $(74{,}19 - 33.30)/18{,}565 \cong 2.20$. On constate que $P_0(T_{.(10)} \geqslant 2.20) < 10P(\boldsymbol{t}_8 > 3.45)$ $\cong 4.3$ %. Avec un risque d'erreur négligeable on rejette donc l'hypothèse H_0.

Cas de deux valeurs extrêmes discordantes. Lorsque les valeurs $x_{(1)}$ et $x_{(n)}$ paraissent discordantes, on considère la valeur numérique
$t = (x_{(n)} - x_{(1)})/s$ prise par la statistique $T_{(1)(n)} = (X_{(n)} - X_{(1)})/S$.

Lorsque le niveau de signification observé $P_0(T_{(1)(n)} \geqslant t)$ a une valeur trop faible on rejette l'hypothèse H_0 « $X_1, X_2, \ldots, X_n$ est bien un échantillon *iid* d'une loi normale ». Pour obtenir une estimation de $P_0(T_{(1)(n)} \geqslant t)$ on peut utiliser la majoration

$$P_0(T_{(1)(n)} \geqslant t) \leqslant n(n-1)P\left(\boldsymbol{t}_{n-2} > \sqrt{\frac{(n-2)t^2}{2n-2-t^2}}\right),$$

cette inégalité étant une égalité lorsque $t^2 \geqslant 3(n-1)/2$.

Ce test est recommandé lorsque sous l'hypothèse alternative, la distribution des X_i est symétrique.

Cas de *k* valeurs discordantes. Lorsque k valeurs $x_{(n-k+1)}, \ldots, x_{(n-1)}, x_{(n)}$ (avec $k \geqslant 2$) paraissent discordantes, déterminer la valeur t prise par
$T_{(n-k+1)-(n)} = (X_{(n-k+1)} + \ldots + X_{(n-1)} + X_{(n)} - k\overline{X})/S$ puis évaluer le niveau de signification observé $P_0(T_{(n-k+1)-(n)} \geqslant t)$ en utilisant la majoration

$$P_0(T_{(n-k+1)-(n)} \geqslant t) \leqslant \frac{n!}{k!(n-k)!}\, P\left(\boldsymbol{t}_{n-2} > \sqrt{\frac{n(n-2)t^2}{k(n-k)(n-1) - nt^2}}\right),$$

cette inégalité étant une égalité lorsque $t^2 \geqslant k^2(n-1)(n-k-1)/(nk+n)$.

Ce test est recommandé pour tester H_0 contre H_1 « k observations proviennent d'une distribution commune $\mathcal{N}(m+a,\sigma^2)$ où $a>0$ ».

Le choix du nombre k de discordants peut être réalisé en considérant le plus grand des écarts $(x_{(n)}-x_{(n-1)})$, $(x_{(n-1)}-x_{(n-2)})$,… noté $(x_{(\ell+1)}-x_{(\ell)})$ puis en considérant comme discordants les $k=n-\ell$ valeurs $x_{(i)}$.

Section 5

TRAITEMENTS AVEC SPSS ET EXCEL

1 Traitements statistiques avec SPSS

Un établissement bancaire spécialisé dans les crédits aux entreprises a élaboré une fonction score. Le service d'études dispose de données statistiques concernant 10 clients d'un secteur donné et souhaite savoir si le score des firmes du secteur suit une distribution normale.

Clients défaillants	1	2	3	4	5	6	7	8	9	10
Score *X*	0,01	0,33	0,25	0,21	0,24	0,19	0,18	0,17	0,14	0,10

À partir de ces données on se propose de tester H_0 « X suit une loi normale » contre $\overline{H}_0$ « la distribution G de la loi que suit X n'est pas une distribution normale » avec le test de Kolmogorov-Smirnov.

Procédure.

1. On pointe sur (Analyse) puis (Tests non paramétriques) puis ($K-S$ à 1 échantillon).

2. Dans le menu **Test Kolmogorov-Smirnov** on sélectionne la variable x et la distribution à tester (Gaussienne)

© Dunod. La photocopie non autorisée est un délit.

3. Pour obtenir la valeur du skewness et du kurtosis il faut cliquer sur (Statistiques descriptives) puis sur (Effectif) et sélectionner dans (Statistiques) applatissement et asymétrie (*cf.* chapitre 1 p 000).

Résultat du test de Kolmogorov-Smirnov à un échantillon

		X
N		10,00
Paramètres normaux	Moyenne	0,18
	Écart-type	0,09
Différences les plus extrêmes	Absolue	0,15
	Positive	0,12
	Négative	– 0,15
Z de Kolmogorov-Smirnov		0,46
Signification asymptotique (bilatérale)		0,98

N désigne la taille de l'échantillon. La moyenne de l'échantillon $\overline{x} \cong 0{,}18$ et l'écart-type standard de l'échantillon $s \cong 0{,}09$ sont les deux estimations des paramètres m et σ de la loi supposée normale.

La valeur prise par la statistique D de Kolmogorov-Smirnov est $d = \max\{d^+, d^-\} = 0{,}15$ où $d^+ = \max_i\{(i/n) - \Phi[(x_{(i)} - \overline{x}/s)]\} = 0{,}12$ et $d^- = \max_i\{\Phi[(x_{(i)} - \overline{x}/s)] - (i-1)/n\} = 0{,}15$ pour $i = 1, \ldots, n$. La *v.a.* $Z = n^{1/2}D$ prend la valeur $0{,}145 \times 10^{1/2} = 0{,}46$.

Le niveau de signification observé $\alpha_c = P_0(Z \geqslant 0{,}46) = 0{,}98$ est le risque minimum de rejeter à tort l'hypothèse H_0. On accepte donc H_0.

2 Traitements sous Excel

Pour effectuer un test d'ajustement du Khi-deux, il faut au préalable faire le tableau avec les classes et les effectifs observés, et faire le tableau des effectifs théorique puis faire comme au § 2 de la section 5 du chapitre 13.

Exercices (corrections page 361)

Exercice 1

Les Transports FT exploitent une plate forme de messagerie dans la région parisienne. La ramasse a lieu dans l'après-midi et les colis sont regroupés sur le quai de départ en fonction de la plate forme de destination : Bordeaux, Nice, Nancy, Poitiers, Toulouse.

À 19 heures, du lundi au vendredi, un poids lourd de 50 tonnes part pour chaque destination desservie. La répartition des colis par destination étant aléatoire, il arrive donc que le nombre de colis excède les capacités du véhicule sur la destination concernée, certain colis restant à quai. Cet incident peut se produire pour un nombre aléatoire de destinations.

Le relevé suivant, portant sur 30 jours, donne le nombre x_i de destinations saturées et le nombre n_i de fois où cet événement s'est réalisé.

Nombre de destinations saturées x_i	0	1	2	3
Nombre de jours n_i	12	12	4	2

1. Pourquoi peut-on penser à une loi de Poisson pour rendre compte de ces données ?

2. Peut-on ajuster la distribution par une loi de Poisson de paramètre $\lambda = 0{,}8$ (calculer la probabilité théorique correspondant à chaque nombre de destinations saturées) ? Tester cette hypothèse avec un risque d'erreur de 5 % en utilisant le test du Khi-deux.

Exercice 2

On a demandé à 12 acheteurs potentiels d'indiquer quel est le prix qu'ils considèrent comme normal pour le bien qui leur est présenté.

Clients	1	2	3	4	5	6	7	8	9	10	11	12
Prix *X*	20	22	25	15	18	19	21	20,5	19,5	18,5	21,5	23

Tester H_0 « X suit une loi normale » contre $\overline{H}_0$ « la distribution G de la loi que suit X n'est pas une distribution normale » en utilisant le test de Kolmogorov-Smirnov avec un risque d'erreur de première espèce de 5 %.

© Dunod. La photocopie non autorisée est un délit.

15 ANALYSE DE VARIANCES

Dans ce chapitre on étudie d'abord le cas où la mesure X du caractère considéré peut être modifiée par un facteur A susceptible d'être en situation A_1, ou $A_2,\ldots$, ou A_k, puis on envisage le cas où deux facteurs A et B peuvent agir simultanément sur X.

Section 1 ■ Analyse de variances à un facteur
Section 2 ■ Analyse de variances à deux facteurs
Section 3 ■ Test de concordance de Kendall
Section 4 ■ Traitement sous Excel et SPSS

Section 1

ANALYSE DE VARIANCES À UN FACTEUR

Envisageons le cas où la mesure X du caractère considéré peut être modifiée par un facteur A susceptible d'être en situation A_1 ou $A_2,\ldots$ ou A_k, l'influence du facteur étant caractérisée par la relation :

$$X_i = m_i + \varepsilon_i$$

où X_i est la mesure du caractère lorsque $A = A_i$, m_i est une constante de valeur inconnue et $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_k$ sont des variables centrées indépendantes qui sont supposées suivre une même loi centrée ($E(\varepsilon_i) = 0\ \forall\, i = 1,2,\ldots,k$).

Pour tester l'hypothèse « les situations ou *modalités* $A_1, A_2, \ldots, A_k$ du facteur A ont une même influence sur la mesure du caractère » c'est-à-dire « $m_1 = m_2 = \ldots = m_k$ » on peut utiliser le test Jonckheere-Terpstra ou celui de Kruskal-Wallis ou encore le test d'analyse de variances de Fisher qui est plus puissant mais présuppose que $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$ suivent une même loi normale centrée ($\varepsilon_i \sim> \mathcal{N}(0, \sigma_0^2)$)

Exemple

Un dirigeant de magasin à succursales multiples souhaite connaître l'impact de différents types de promotions envisagées sur le chiffre d'affaires. Concevant 3 types de campagnes de promotion P_1, P_2, P_3 ayant des coûts sensiblement égaux, il assigne à 10 magasins tests ces campagnes de promotion selon la répartition suivante : 3 pour P_1, 3 pour P_2 et 4 pour P_3. Le relevé du taux de croissance du chiffre d'affaires de chacun des 10 magasins pour la période des promotions est présenté ci-dessous, ce taux de croissance δ exprimé en % est calculé par référence au chiffre d'affaires de la période précédente de même durée:

	Taux de croissance δ			
Promotion P_1	2,1	4,0	3,5	
Promotion P_2	4,5	3,6	1,8	
Promotion P_3	2,5	2,2	3,1	3,8

Au vu de ces résultats, il convient de tester l'hypothèse H_0 selon laquelle les promotions ont la même influence sur le taux δ d'accroissement du chiffre d'affaires contre l'hypothèse alternative $\overline{H}_0$.

1 Test de Fisher

Pour chaque modalité A_i (où $i = 1,2,\ldots,k$) du facteur A on dispose d'un échantillon *iid* de taille n_i : $X_{i1}, X_{i2}, \ldots, X_{in_i}$, et de la moyenne aléatoire sur l'échantillon i : $\overline{X}_i = \dfrac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$. Autrement dit on dispose de k échantillons de tailles respectives $n_1, n_2, \ldots, n_k$ correspondant chacun à une modalité du facteur A et on calcule les moyennes aléatoires $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$ de ces échantillons ainsi que la moyenne générale $\overline{X} = \dfrac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$ où $n = n_1 + n_2, \ldots + n_k$.

Modèle et formulation de l'hypothèse. Considérant que les modalités A_i du facteur A peuvent influer sur les moyennes de distribution et non sur la variance, on part de la décomposition « $X_{ij} = m_i + \varepsilon_{ij} \; \forall \, j = 1,2,\ldots,n_i$ » où l'on suppose que $\varepsilon_{ij} \sim> \mathcal{N}(0, \sigma_0^2)$, et donc $X_{ij} \sim> \mathcal{N}(m_i, \sigma_0^2) \; \forall \, j = 1,2,\ldots,n_i$ et on teste l'hypothèse H_0 « $m_1 = m_2 = \ldots = m_k$ » contre $\overline{H}_0$.

© Dunod. La photocopie non autorisée est un délit.

Variable statistique utilisée.

La statistique $F = \dfrac{\dfrac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2}{\dfrac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2}$ ou son expression équivalente $F = \dfrac{V_E/(k-1)}{V_R/(n-k)}$ où V_E la variance expliquée (variance des moyennes) est la variance due au facteur A et V_R la variance résiduelle est la moyenne des variances. Lorsque l'hypothèse H_0 est vérifiée, la statistique F suit la loi de Fisher-Snédécor $F(k-1,n-k) : F = F_{n-k}^{k-1}$.

Règle de décision. On cherche w_α tel que $P(F_{n-k}^{k-1} \geqslant w_\alpha) = \alpha$ (risque d'erreur) et on rejette H_0 au profit de $\overline{H}_0$ lorsque la valeur w prise par la statistique F est supérieure à w_α. Le niveau de signification observé est $\alpha_c = P(F_{n-k}^{k-1} \geqslant w)$.

Exemple

Il s'agit d'analyser le problème présenté dans l'exemple introductif. Notons X_{ij} le taux de croissance du chiffre d'affaires du j-ème magasin soumis à une promotion de type $P_i (i = 1,2 \text{ ou } 3)$. Si on admet que X_{ij} fluctue de façon normale autour de sa valeur moyenne m_i qui caractérise l'impact de la promotion P_i on peut appliquer le test associé à l'analyse de variance en utilisant le tableau suivant :

					$\overline{x}_i$	n_i	$\sum_j (x_{ij} - \overline{x}_i)^2$	$n_i(\overline{x}_i - \overline{x})^2$
A_1 « promotion P_1 »	x_{11}=2,1	x_{12}=4,0	x_{13}=3,5		3,2	3	1,94	0,02
A_2 « promotion P_2»	x_{21}=4,5	x_{22}=3,6	x_{23}=1,8		3,3	3	3,78	0,11
A_3 « promotion P_3 »	x_{31}=2,5	x_{32}=2,2	x_{33}=3,1	x_{34}=3,8	2,9	4	1,5	0,18
Total						10	7,22	0,31

$n = 3 + 3 + 4 = 10$, $k = 3$, $n_1 = 3$, $n_2 = 3$, $n_3 = 4$,

$\overline{x} = (2,1 + 4,0 + \ldots + 3,8)/10 = (3 \times 3,2 + 3 \times 3,3 + 4 \times 2,9)/10 = 3,11$

Aussi on constate que la statistique F prend la valeur $w = \dfrac{\dfrac{1}{3-1} \times 0,31}{\dfrac{1}{10-3} \times 7,22} \cong 0,15$.

Règle décision. Avec un risque d'erreur de 5 %, on cherche w_α tel que $0,05 = P_0(F \geqslant w_\alpha) = P(F_7^2 \geqslant w_\alpha)$. On lit (*cf.* p. 379) $w_\alpha = 4,73$. La valeur w prise par F étant inférieur à 4,73, on accepte donc l'hypothèse selon laquelle les différents types de promotions ont le même impact : $m_1 = m_2 = m_3$.

2 Test par les rangs de Jonckheere-Terpstra

Ce test, qui ne présuppose pas la normalité des variables ε_{ij}, peut être utilisé pour tester l'hypothèse H_0 « $m_1 = m_2 = \ldots = m_k$ » contre $\overline{H}_0$ « il existe au moins deux échantillons i_0 et i'_0 tel que $m_{i_0} \neq m'_{i_0}$ ».

La statistique J utilisée. Pour une étude comparative des échantillons deux à deux, par exemple aux deux échantillons i et j, où $i < j$: $\{x_{i1},\ldots,x_{in_i}\}$ et $\{x_{j1},\ldots,x_{jn_j}\}$ on associe le nombre u_{ij} de couples (x_{ih},x_{jk}) tels que $x_{ih} > x_{jk}$. Renouvelant cette opération autant que nécessaire pour que tous les échantillons soient comparés deux à deux, on somme les nombres u_{ij} et on obtient la valeur j^* prise par la statistique J de Jonckheere-Terpstra : $j^* = \sum u_{ij}$ avec $1 \leqslant i < j \leqslant k$. La statistique J de Jonckheere-Terpstra est définie comme étant la somme des variables de Mann-Whitney U_{ij} qui prennent pour valeur u_{ij} : $J = \sum U_{ij}$ avec $1 \leqslant i < j \leqslant k$.

Distribution de J sous H_0. Sous l'hypothèse de base H_0, « $m_1 = m_2 \ldots = m_k$ »

i) certaines valeurs critiques de la distribution de J_n sont tabulées lorsque le nombre de facteurs k est faible et que les effectifs n_i des échantillons sont petits (*cf.* p. 385) ;

ii) lorsque $n_i \geqslant 5\ \forall\, i = 1,\ldots,k$ on utilise l'approximation normale avec correction de continuité :

$$[J - E_0(J)]/\sigma_0(J) \cong \mathcal{N}_{0;1}$$

où $E_0(J) = (n^2 - \sum_{i=1}^{k} n_i^2)/4$, $V_0(J) = \dfrac{1}{72}\left(n^2(3+2n) - \sum_{i=1}^{k} n_i^2(3+2n_i)\right)$

Règle de décision. On cherche l'entier θ_α tel que $P_0(J \geqslant \theta_\alpha) \cong \alpha$ puis on rejette H_0 lorsque la valeur j^* prise par J est supérieure ou égale à θ_α.

Exemple

Considérons l'exemple introductif du § 1.

Statistique. On réalise une étude comparative des échantillons deux à deux :

– aux premier et deuxième échantillons d'effectifs respectifs 3 et 3, on associe l'ensemble des couples réalisables {(2,1, 4,5), (2,1, 3,6), (2,1, 1,8), (4,0, 4,5), (4,0, 3,6), (4,0, 1,8), (3,5, 4,5), (3,5, 3,6), (3,5, 1,8)} et on constate qu'il y a 4 couples où la valeur issue du premier échantillon est supérieure à celle du deuxième échantillon : $u_{12} = 4$;

– aux premier et troisième échantillons, on associe la variable de Mann-Whitney U_{13} qui prend pour valeur $u_{13} = 7$ puisqu'il y a 7 couples où la valeur issue du premier échantillon est supérieure à celle du troisième échantillon ;

– aux deuxième et troisième échantillons, on associe la valeur $u_{23} = 7$.

On en déduit que J prend pour valeur $j^* = (4 + 7 + 7) = 18$.

© Dunod. La photocopie non autorisée est un délit.

Règle de décision. La distribution de J étant tabulée sous l'hypothèse de base H_0 « $m_1 = m_2 = m_3$ » on cherche sur la table statistique avec $n_1 = 3$, $n_2 = 3$ et $n_3 = 4$, l'entier θ_α tel que $P_0(J \geqslant \theta_\alpha) \cong 10\ \%$ soit $\theta_\alpha = 24$. La valeur j^* prise par J étant inférieure à 24 on ne peut rejeter l'hypothèse de base.

3 Test par les rangs de Kruskal-Wallis

Comme précédemment ce test ne présuppose pas la normalité de la loi suivi par les ε_{ij}.

Il convient alors de tester H_0 « $m_1 = m_2 = \ldots = m_k$ » contre $\overline{H}_0$.

La variable statistique KW utilisée. Classant par ordre croissant l'ensemble des n valeurs x_{ij} prises par les X_{ij} on attribue à X_{ij} son rang noté R_{ij}. Pour chaque échantillon i, on détermine alors la somme $R_{i\bullet} = \sum_{j=1}^{n_i} R_{ij}$ des rangs de ses éléments constitutifs puis la moyenne de ses rangs $\overline{R}_{i\bullet} = R_{i\bullet}/n_i$. La variable de Kruskal-Wallis est définie à une constante multiplicative près comme la variance des moyennes $\overline{R}_{i\bullet}$ des rangs des échantillons dont la moyenne générale est égale à $(n+1)/2$, la pondération étant celle des effectifs des échantillons :

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\overline{R}_{i\bullet} - \frac{n+1}{2} \right)^2$$

Distribution de KW sous H_0. Sous l'hypothèse H_0

i) on a $E_0(KW) = k - 1$

et $$V_0(KW) = 2(k-1) - \frac{2n(2k^2 - 6k + 1) + 6k^2 - 12k}{5n(n+1)} - \frac{6}{5} \sum_{i=1}^{k} \frac{1}{n_i}$$

ii) la distribution est tabulée pour des faibles valeurs de k et des n_i (p. 383).

iii) Si $n_i \geqslant 5\ \forall\, i = 1,\ldots,k$ on utilise l'approximation $KW \cong \chi^2_{(k-1)}$

Règle de décision. Pour un niveau de signification α on détermine le nombre c_α tel que $P_0(KW \geqslant c_\alpha) \cong \alpha$ et on rejette H_0 au profit de $\overline{H}_0$ lorsque la valeur prise par KW est supérieure à c_α. En effet sous l'hypothèse alternative $\overline{H}_0$, la statistique KW tend à prendre de grandes valeurs.

Exemple

Reprenant l'exemple précédent, on range par ordre croissant de valeurs croissantes les réalisations de chaque échantillon :

A_1 « promotion P_1 »	2,1	3,5	4,0	
A_2 « promotion P_2 »	1,8	3,6	4,5	
A_3 « promotion P_3 »	2,2	2,5	3,1	3,8

Dans cet exemple où il y a $k = 3$ sous populations, $n_1 = 3$ est la taille du premier échantillon issu de la première sous population, $n_2 = 3$ est la taille du deuxième échantillon issu de la seconde sous population, $n_3 = 4$ est la taille du troisième échantillon et $n = 10$ est la taille agrégée des différents échantillons $(3 + 3 + 4)$.

On attribue à chacune des 10 valeurs numériques proposées, son rang et on obtient le classement suivant :

Échantillon extrait de P_1	2	6	9		$R_{1\bullet} = 17$	$\overline{R}_{1\bullet} = 17/3$
Échantillon extrait de P_2	1	7	10		$R_{2\bullet} = 18$	$\overline{R}_{2\bullet} = 18/3$
Échantillon extrait de P_3	3	4	5	8	$R_{3\bullet} = 20$	$\overline{R}_{3\bullet} = 20/4$

La somme des rangs : $R_{1\bullet} + R_{2\bullet} + R_{3\bullet} = n(n+1)/2 = 55$.

La *v.a.* $KW = \dfrac{12}{10(10+1)} \quad \displaystyle\sum_{i=1}^{3} n_i \left(\overline{R}_{i\bullet} - \frac{10+1}{2}\right)^2$ prend la valeur numérique

$$kw^* = \frac{12}{10(10+1)}[3 \times (5{,}66 - 5{,}5)^2 + 3 \times (6 - 5{,}5)^2 + 4 \times (5 - 5{,}5)^2] = 0{,}2\,.$$

Règle de décision. Pour $\alpha = 0{,}05$ on cherche c_α tel que $P_0(KW \geqslant c_\alpha) \cong 0{,}05$ et l'on obtient $c_\alpha = 5{,}73$. La valeur prise par *KW* étant inférieure à 5,73 on ne peut rejeter H_0.

Section 2

ANALYSE DE VARIANCES À DEUX FACTEURS

Une population ***P*** sur laquelle on mesure un caractère est partagée en k sous populations $P_1, \ldots, P_k$. Voulant savoir si un facteur A qui a n modalités $(A_1, A_2, \ldots, A_n)$ influe sur la distribution de ce caractère, on extrait au hasard et avec remise de chaque sous population P_i, n éléments $\{e_{i1}, \ldots, e_{in}\}$. e_{i1} est soumis au facteur d'environnement A_1, e_{i2} au facteur $A_2, \ldots, e_{in}$ au facteur A_n. On a donc, à raison d'un élément par sous population, k éléments soumis à la modalité A_1 du facteur A, k éléments soumis à la modalité A_2 du facteur A, etc.

Ce type d'analyse peut être synthétisé par le tableau croisé suivant où x_{ij} désigne la réalisation de l'expérimentation faite sur l'élément tiré de la sous population i et soumis au facteur j, élément noté e_{ij} auquel correspond la *v.a* X_{ij} :

© Dunod. La photocopie non autorisée est un délit.

Tableau 15.1

Facteur : modalités A_j → **Échantillon de la s/population P_i ↓**	A_1	A_2	•	A_n	Moyennes sur l'échantillon $\overline{x}_{i\bullet}$
Échantillon de P_1	x_{11}	x_{12}	•	x_{1n}	$\overline{x}_{1\bullet}$
Échantillon de P_2	x_{21}	x_{22}	•	x_{2n}	$\overline{x}_{2\bullet}$
•	•	•	•	•	•
Échantillon de P_k	x_{k1}	x_{k2}	•	x_{kn}	$\overline{x}_{k\bullet}$
Moyennes des valeurs soumises à la modalité A_j $\overline{x}_{\bullet j}$	$\overline{x}_{\bullet 1}$	$\overline{x}_{\bullet 2}$	•	$\overline{x}_{\bullet n}$	

où $\overline{x}_{i\bullet} = \dfrac{1}{n}\sum_{j=1}^{n} x_{ij}$ et $\overline{x}_{\bullet j} = \dfrac{1}{k}\sum_{j=1}^{k} x_{ij}$

Généralisation. On peut envisager de soumettre au facteur A_i non pas un seul élément de P_j mais s éléments auxquels sont associées les *v.a.* $X_{ij1}, X_{ij2},\ldots,X_{ijs}$. Le nombre s étant constant $\forall\, i$ et $\forall\, j$ on a $s \times k \times n$ expérimentations. Il suffit alors de substituer à X_{ij} défini précédemment, la moyenne aléatoire $\overline{X}_{ij} = (X_{ij1} + \ldots + X_{ijs})/s$.

1 Test par les rangs de Friedman

X_{ij} désignant la *v.a.* définie sur $P_i \times A_j$ qui prend la valeur numérique x_{ij}, on considère les fonctions de répartition $F_{ij}(t) = P(X_{ij} \leqslant t)$. Sous l'hypothèse de base H_0 « les facteurs $A_1,\ldots,A_n$ n'ont pas d'influence stochastiquement différente » on doit avoir pour chaque $i = 1,2,\ldots,k$: $F_{i1}(t) = \ldots = F_{in}(t)$ noté $F_i(t)$.

La statistique $F_{k,n}$ utilisée. À l'intérieur de chaque échantillon on ordonne les valeurs obtenues et on détermine leur rang dans l'échantillon. Ainsi r_{ij} désigne le rang de x_{ij} dans l'échantillon issu de P_i, c'est-à-dire le rang de x_{ij} parmi les n termes $x_{i1},\ldots,x_{in}$.

Tableau 15.2

Facteur : modalités A_j → **Échantillon de la s/population P_i ↓**	A_1	A_2	•	A_n	Total
P_1	r_{11}	r_{12}	•	r_{1n}	$n(n+1)/2$
P_2	r_{21}	r_{22}	•	r_{2n}	$n(n+1)/2$
•	•	•	•	•	•
P_k	r_{k1}	r_{k2}	•	r_{kn}	$n(n+1)/2$
Total	$r_{\bullet 1}$	$r_{\bullet 2}$	•	$r_{\bullet n}$	$kn(n+1)/2$
$\overline{R}_{\bullet j}$ rang moyen de A_j	$\overline{r}_{\bullet 1} = r_{\bullet 1}/k$	$\overline{r}_{\bullet 2} = r_{\bullet 2}/k$		$\overline{r}_{\bullet n} = r_{\bullet n}/k$	

Notons R_{ij} la *v.a.* qui prend la valeur r_{ij}. La somme des rangs des observations soumises à la modalité A_j du facteur A est $R_{\bullet j} = \sum_{i=1}^{k} R_{ij}$, elle prend après expérimentation la valeur $r_{\bullet j}$. La statistique $F_{k,n}$ de Friedman est, à une constante multiplicative près, la somme des carrés des écarts entre les rangs moyens $\overline{R}_{\bullet j}$ et la moyenne des rangs égale à $(n+1)/2$: $F_{k,n} = \dfrac{12\,k}{n^2+n} \sum_{j=1}^{n} \left(\overline{R}_{\bullet j} - \dfrac{n+1}{2} \right)^2$.

Distribution de $F_{k,n}$ sous l'hypothèse nulle. Sous l'hypothèse de base H_0,

i) les valeurs critiques de la distribution de $F_{k,n}$ sont tabulées (*cf.* p. 386).

ii) pour k suffisamment grand, on peut utiliser l'approximation : $F_{k,n} \cong \chi^2_{n-1}$

Règle de décision. Pour un risque d'erreur α, on détermine le nombre c_α tel que $P_0(F_{k,n} \geqslant c_\alpha) \cong \alpha$ et on rejette H_0 au profit de $\overline{H}_0$ lorsque la valeur f^* prise par $F_{k,n}$ est supérieure à c_α. En effet, sous l'hypothèse H_0 « les facteurs $A_1, \ldots, A_n$ n'ont pas d'influence stochastiquement différente » la statistique $F_{k,n}$ a tendance à prendre de faibles valeurs puisqu'on a alors équiprobabilité des rangs moyens $\overline{R}_{\bullet j}$.

Exemple

Une entreprise a essayé trois types de rémunération de ses vendeurs afin de connaître l'impact du type de rémunération (A_1, A_2, A_3) sur le chiffre d'affaires qu'ils réalisent. La population des vendeurs est scindée en trois sous-populations selon le critère de l'ancienneté du vendeur et on prélève dans chaque sous population P_i un échantillon de 3 individus. Au premier individu de l'échantillon, on attribue le mode de rémunération A_1, au deuxième le mode de rémunération A_2, au troisième le mode de rémunération A_3. Sont relevés en fin de trimestre les chiffres d'affaires réalisés (les zones géographiques des vendeurs étant similaires) :

Facteur : modalités $A_j \rightarrow$ **Échantillon de la s/population $P_i \downarrow$**	A_1	A_2	A_3
Échantillon de P_1	$x_{11} = 150$	$x_{12} = 155$	$x_{13} = 160$
Échantillon de P_2	$x_{21} = 157$	$x_{22} = 156$	$x_{23} = 165$
Échantillon de P_3	$x_{31} = 164$	$x_{32} = 162$	$x_{33} = 163$

Pour tester l'hypothèse selon laquelle le type de rémunération n'influe pas sur la performance du vendeur on dresse le tableau des rangs des valeurs pour chaque sous population :

© Dunod. La photocopie non autorisée est un délit.

Facteur : modalités $A_j \rightarrow$ Échantillon de la s/population $P_i \downarrow$	A_1	A_2	A_3	Total
Échantillon de P_1	$r_{11} = 1$	$r_{12} = 2$	$r_{13} = 3$	6
Échantillon de P_2	$r_{21} = 2$	$r_{22} = 1$	$r_{23} = 3$	6
Échantillon de P_k	$r_{31} = 3$	$r_{32} = 1$	$r_{33} = 2$	6
Moyennes des rangs pour les modalités des facteurs	$\overline{r}_{\bullet 1} = 2$	$\overline{r}_{\bullet 2} = 4/3$	$\overline{r}_{\bullet 3} = 8/3$	6

La taille de chaque échantillon est $n = 3$ et le nombre de sous population $k = 3$, $F_{k,n}$ (soit $F_{3,3}$) prend pour valeur $f^* = \dfrac{12 \times 3}{3^2 + 3}[(\overline{r}_{\bullet 1} - 2)^2 + (\overline{r}_{\bullet 2} - 2)^2 + (\overline{r}_{\bullet 3} - 2)^2] = 8/3$.

Règle de décision. Si on utilisait l'approximation asymptotique $F_{k,n} \cong \chi^2_{n-1}$ quoique k ait une valeur trop faible, on chercherait c_α tel que $0{,}05 = P_0(F_{3,3} \geqslant c_\alpha) \cong P(\chi^2_2 \geqslant c_\alpha)$. On lirait $c_\alpha = 5{,}99$ et on conclurait qu'on ne peut rejeter H_0. Utilisant la table de $F_{3,3}$ on trouve $c_\alpha = 6$.

2 Test de Fisher

Modèle et formulation de l'hypothèse. Le modèle étudié ici est du type

$$X_{ij} = m_i + \alpha_j + \varepsilon_{ij}, \forall\, i = 1,\ldots,k \text{ et } j = 1,\ldots,n$$

où m_i et α_j sont des constantes et les ε_{ij} sont des *v.a.* indépendantes qui suivent une même loi normale centrée.

L'hypothèse H_0 « les facteurs $A_1,\ldots,A_n$ n'ont pas d'influence stochastiquement différente » devient alors « $\alpha_1 = \alpha_2 = \ldots = \alpha_n$ ».

La statistique $\phi_{k,n}$ utilisée. Notons $\overline{X}_{i\bullet} = \dfrac{1}{n}\sum_{j=1}^{n} X_{ij}$ la moyenne aléatoire sur l'échantillon i issu de la sous-population P_i, $\overline{X}_{\bullet j} = \dfrac{1}{k}\sum_{i=1}^{k} X_{ij}$ la moyenne aléatoire des valeurs soumises à la modalité A_j du facteur A, $\overline{X}_{\bullet\bullet} = \dfrac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}$ la moyenne aléatoire de l'ensemble des valeurs. Sous l'hypothèse H_0 on sait que la *v.a*

$$\phi_{kn} = \frac{k\sum_{i=1}^{k}(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2/(n-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X}_{i\bullet} - \overline{X}_{\bullet j} + \overline{X}_{\bullet\bullet})^2/(k-1)(n-1)} = F^{(n-1)}_{(k-1)(n-1)}$$

Règle de décision. Pour un niveau de signification α on détermine le nombre c_α tel que $P(F^{(n-1)}_{(k-1)(n-1)} \geqslant c_\alpha) \cong \alpha$ et on rejette H_0 au profit de $\overline{H}_0$ lorsque la valeur prise par $\phi_{k,n}$ est supérieure à c_α. Sous l'hypothèse H_0 « les facteurs $A_1,\ldots,A_n$ n'ont pas

d'influence stochastiquement différente » la statistique $\phi_{k,n}$ a tendance à prendre de faibles valeurs. On rejette l'hypothèse H_0 si la valeur prise par $\phi_{k,n}$ est trop grande.

Section 3

TEST DE CONCORDANCE DE KENDALL

Exemple introductif

Afin de pourvoir un emploi on demande séparément à trois examinateurs de classer les six candidats A, B, C, D, E et F de 1 à 6 : 6 pour le meilleur candidat, 5 pour le suivant par ordre de mérite, …, 1 pour le moins performant. Les résultats sont les suivants :

Candidats

Juges	A	B	C	D	E	F
juge n° 1	1	3	5	6	4	2
juge n° 2	1	2	4	5	3	6
juge n° 3	2	1	6	3	5	4
Total $R_{i\bullet}$ des rangs	4	6	15	14	12	12

Plus généralement considérons n candidats classés de 1 à n par chacun des k juges. Le rang attribué au candidat i par le juge n°j étant noté R_{ij} on considère le total $R_{i\bullet}$ des rangs attribués au candidat i par l'ensemble des juges et conséquemment son rang moyen $\overline{R}_{i\bullet} = R_{i\bullet}/k$.

Statistique utilisée

Le coefficient de concordance $W_{k,n}$ de Kendall prend en compte, pour chaque candidat i, le carré de l'écart entre la moyenne des rangs obtenus par ledit candidat et la moyenne de tous les rangs attribués $((n+1)/2)$

$$W_{k,n} = \frac{12}{(n^3-n)} \sum_{i=1}^{n} \left(\overline{R}_{i\bullet} - \frac{n+1}{2} \right)^2 = \frac{12 \sum_{i=1}^{n} R_{i\bullet}^2}{k^2(n^3-n)} - 3\,\frac{n+1}{n-1}$$

Les valeurs prises par le coefficient de concordance $W_{k,n}$ sont comprises entre 0 et 1. Dans le cas d'une concordance parfaite $(R_{i1} = R_{i2} = \ldots = R_{ik}$ $\forall i = 1,2,\ldots,n)$ $W_{k,n} = 1$. La valeur prise par $W_{k,n}$ doit être considérée comme un degré de concordance.

© Dunod. La photocopie non autorisée est un délit.

Distribution de $W_{k,n}$ sous l'hypothèse nulle. Sous l'hypothèse H_0 « toutes les distributions de classement sont équiprobables (avec une probabilité égale à $1/(n!)^k$ », certaines valeurs critiques de la distribution de $W_{k,n}$ sont données page 382.

Règle de décision. Pour tester l'hypothèse H_0 contre $\overline{H}_0$ avec un niveau de signification α on cherche w_α tel que $P_0(W_{k,n} \geqslant w_\alpha) \cong \alpha$. On rejette H_0 lorsque la valeur prise par $W_{k,n}$ est supérieure ou égale à w_α.

Exemple (suite)

Dans l'exemple introductif on obtient $W_{3;6} = 0{,}6317$. Or $P_0(W_{3;6} \geqslant 0{,}6317) \cong 5\ \%$. On rejettera donc l'hypothèse selon laquelle les classements réalisés par les trois juges ne sont pas liés et conclure qu'il y a un bon consensus dans les classements.

REPÈRE : Cas des ex-aequo

Dans chaque classement présentant des ex-aequo on attribue à chacun de ceux-ci le rang moyen du groupe d'ex-aequo auquel il appartient et qui n'est pas nécessairement un entier. Lorsque le classement n°j a h_j groupes d'ex-aequo, on attribue à ce classement la somme $T_j = \sum_{i=1}^{h_j} (t_i^3 - t_i)$ où t_i désigne le nombre d'éléments du i-ème de ces groupes. S'il n'a pas d'ex-aequo on a évidemment $T_j = 0$ puisque la répartition des n entiers du classement en classes de nombres égaux donne $h_j = n$ et $t_i = 1\ \forall i$. On considère alors la variable

$$W^*_{k,n} = \frac{12}{(n^3 - n) - \dfrac{1}{k}\displaystyle\sum_{j=1}^{k} T_j} \sum_{i=1}^{n} \left(\overline{R}_{i\bullet} - \frac{n+1}{2} \right)^2 .$$

Sous l'hypothèse H_0 et lorsque les conditions suivantes sont réalisées : « $k \geqslant 7$ et $q = (\Sigma_j T_j)/k(n^3 - n)$ a une valeur faible (c'est-à-dire $q < 0{,}1$) », on utilise les approximations précédemment définies pour le cas où les classements ne présentent pas d'ex-aequo. Lorsque le quotient q n'a pas une valeur négligeable on calcule les k nombres $\theta_i = T_i/(n^3 - n)$ puis on utilise l'approximation $(k-1)\dfrac{W^*_{k,n}}{1 - W^*_{k,n}} \approx F^{\nu_1}_{\nu_2}$ (variable de Fisher)

où $\nu_1 = \dfrac{(k-1)(n-1)}{2k} \times \dfrac{\left(\displaystyle\sum_{i=1}^{k} (1 - \theta_i) \right)^2}{\displaystyle\sum_{1 \leqslant i < j \leqslant k} (1 - \theta_i)(1 - \theta_j)} - \dfrac{2}{k}$ et $\nu_2 = (k-1)\nu_1$.

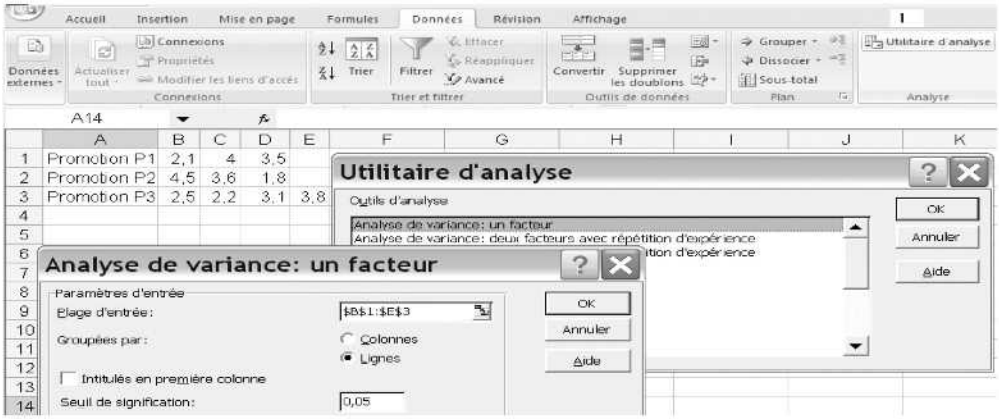
Section 4

TRAITEMENTS SOUS EXCEL ET SPSS

Reprenons l'exemple introductif où le dirigeant de magasin à succursales multiples souhaite connaître les impacts de différents types de promotions P_1, P_2, P_3 envisagées sur le chiffre d'affaires. Il assigne à 10 magasins ces campagnes de promotion selon la répartition suivante : 3 pour P_1, 3 pour P_2 et 4 pour P_3.

1 Analyse de variance à un facteur avec Excel

Le relevé du taux de croissance du chiffre d'affaires de chacun des 10 magasins pour la période des promotions est présenté dans la feuille excel ci-dessous. Partant de ces données, est réalisée l'analyse de variance à un facteur (test de Fisher). Considérant que les modalités du facteur « promotion « peuvent influer sur les moyennes de distribution on part de la décomposition « $X_{ij} = m_i + \varepsilon_{ij}\ \forall\ j = 1,2,\ldots,n_i$ où l'on suppose que $\varepsilon_{ij} \sim> \mathcal{N}(0,\sigma_0^2)$ et on teste l'hypothèse H_0 « $m_1 = m_2 = m_3$ » contre $\overline{H}_0$.



Procédure.

1. On entre les données (en cellule E_3 figure 3,8), on se place dans une cellule quelconque non utilisée, on clique sur (Données) puis (Utilitaire d'analyse).

2. Dans le **menu utilitaire d'analyse**, on sélectionne (Analyse de variance à un facteur).

3. Dans le **menu analyse de variance à un facteur**, on sélectionne les cellules correspondant aux observations ($B$1 :$E$3), le facteur A_i figurant en ligne on sélectionne (ligne) et on obtient le menu suivant. Le seuil de signification 0,05 est le risque d'erreur de première espèce.

© Dunod. La photocopie non autorisée est un délit.

– Résultats

RAPPORT DÉTAILLÉ				
Groupes	**Nombre d'échantillons**	**Somme**	**Moyenne**	**Variance**
Ligne 1	3	9,6	3,2	0,97
Ligne 2	3	9,9	3,3	1,89
Ligne 3	4	11,6	2,9	0,5

ANALYSE DE VARIANCE						
Source des variations	**Somme des carrés**	**Degré de liberté**	**Moyenne des carrés**	***F***	**Probabilité**	**Valeur critique pour *F***
Entre Groupes	0,309	2	0,15	0,15	0,86	4,737
À l'intérieur des groupes	7,22	7	1,03			
Total	7,529	9				

La somme des carrés expliqués par le facteur promotion ou interclasses $S_E = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2$ prend pour valeur 0,309. La somme des carrés résiduels ou intra-classes $S_R = \sum_i \sum_j (X_{ij} - \overline{X}_i)^2$ prend pour valeur 7,22. La statistique F de l'analyse de variance $F = \dfrac{S_E/(k-1)}{S_R/(n-k)}$ suit, sous l'hypothèse H_0, une loi de Fisher-Snédecor $F(2; 7)$. Elle prend pour valeur 0,15 et le risque de rejet à tort de H_0 est ici de $0,86 = P_0(F \geqslant 0,15) = P(F_7^2 \geqslant 0,15)$. On ne peut donc rejeter l'hypothèse d'absence d'influence du type de promotion.

2 Analyse de variance à un facteur avec SPSS (test de Fisher)

Procédure.

1. On entre les données x_{ij} en colonne dans la variable $\boldsymbol{x}$ et la modalité A_j du facteur A correspondant à l'observation dans une colonne adjacente (1 pour A_1, 2 pour A_2, 3 pour A_3.)

2. On clique sur Analyse, Comparer les moyennes et on sélectionne ANOVA à 1 facteur.

3. Dans le menu ANOVA dans lequel on sélectionne la variable dépendante x et la variable a représentant les facteurs appelée critère.

Les résultats obtenus sont identiques à ceux obtenus avec Excel (pour interprétation *cf*. ci-dessus).

3 Test de Kruskal-Wallis avec SPSS

Procédure.

1. Identique à précédemment.

2. On entre les données, on clique sur Analyse puis sur Tests non paramétriques.

3. Obtenant le menu test pour plusieurs échantillons indépendants on sélectionne H de Kruskall-Wallis (le test de Jonckheere-Terpstra est disponible), on sélectionne ***X* la variable à tester** et **dans critère de groupement** on fait figurer la variable a représentant les différentes modalités 1, 2 et 3 du facteur associées aux observations x_{ij}.

4. Cliquer sur Définir niveaux et sélectionner les modalités 1, 2, 3 du facteur A en mettant 1 dans minimum et 3 dans maximum.

Résultats Test de Kruskal Wallis

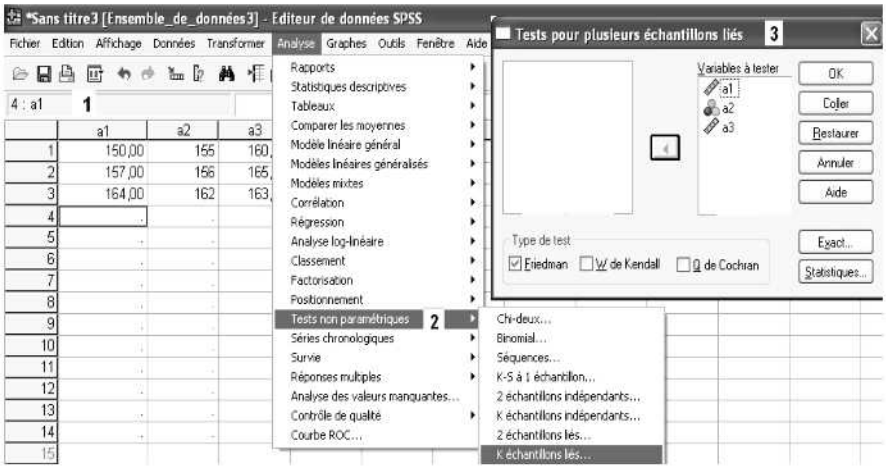
	A	***n***	**Rang moyen**		***X***
X	1	3	5,67	Khi-deux	0,20
	2	3	6,00	*ddl*	2,00
	3	4	5,00	Signification asymptotique	0,90
	Total	10			

Dans cet exemple, le nombre de sous populations $k = 3$, la taille de l'échantillon extrait de la sous population 1 est $n_1 = 3$, la taille de l'échantillon extrait de la sous population 2 est $n_2 = 3$, $n_3 = 4$ et $n = 10$. Le rang moyen occupé par les observations issues de la sous population 1 dans l'ensemble des observations est $\overline{r}_1 = r_{1\bullet}/3 = 5{,}67$ etc. La statistique KW prend pour valeur 0,2. Le risque de rejeter à tort H_0 étant égal à $P_0(KW \geqslant 0{,}2) \cong P(\chi^2_{(3-1)} > 0{,}2) = 0{,}9$, on décide de ne pas rejeter H_0. En fait l'approximation employée est ici peu compatible avec la taille des échantillons.

© Dunod. La photocopie non autorisée est un délit.

4 Test de Friedman avec SPSS

On reprend l'exemple de la section 2.



Procédure.

1. On entre les données correspondant à chaque modalité du facteur A dans une colonne d'où on obtient trois variables colonnes a_1,a_2 et a_3 , les observations en ligne sont les k échantillons, soit ici $k = 3$.

2. On clique sur (Analyse) puis sur (Tests non paramétriques) puis on sélectionne (K-échantillons liés).

3. Dans le menu **test pour plusieurs échantillons liés**, on sélectionne les variables à tester a_1,a_2,a_3 et le type de test : (Friedman).

Résultats du Test de Friedman

	Rang moyen	*N*	3
A_1	2,00	Khi-deux	2,67
A_2	1,33	ddl	2,00
A_3	2,67	Signification asymptotique	0,26

La statistique de Friedman $F_{k,n} = \dfrac{12}{kn(n+1)} \sum_{j=1}^{n} R_{j\bullet}^2 - 3k(n+1)$ prend pour valeur 2,67. La moyenne des rangs des observations soumises au facteur A_1 est 2. Sous l'hypothèse de base H_0, le logiciel utilise l'approximation asymptotique $F_{k,n} \cong \chi^2_{3-1}$ et calcule la probabilité maximum α_c de rejeter à tort H_0 : $\alpha_c = P_0(F_{k,n} \geqslant 2{,}67) \cong 0{,}26$. On ne peut donc rejeter H_0.

Remarque : cette procédure permet à l'étape 3 d'obtenir le test de concordance de Kendall.

Exercices (*cf.* corrections page 363)

Exercice

Un chercheur souhaite savoir si l'implication des vendeurs varie selon le critère choisi par l'entreprise pour calculer le montant de la part variable de rémunération. Le calcul de cette part variable se fait principalement de trois façons : en fonction de la réalisation des objectifs assignés au cadre (sous population G_1), en fonction de l'accroissement du chiffre d'affaires (sous population G_2), en fonction de la marge réalisée (sous population G_3). Pouvant interroger 5 *vendeurs* qui appartiennent au groupe G_1, 4 à G_2 et 4 à G_3, il obtient à l'aide du questionnaire *O.C.Q de Porter et alii* (*organizational commitment questionnaire*) un score X permettant d'apprécier son degré d'implication organisationnelle : plus le score est élevé et plus l'implication organisationnelle du cadre est grande. Dans le tableau ci-après figure, pour chaque élément de l'échantillon, le score d'implication x_{ij} obtenu par le vendeur n_j° du groupe *i.*

Observation *j* / **Groupe *i***	**1**	**2**	**3**	**4**	**5**
G_1	60	65	85	80	75
G_2	52	66	82	77	
G_3	36	58	92	84	

1. Utiliser l'analyse de variance de Fisher afin de tester l'hypothèse selon laquelle la distribution du score d'implication est la même quelque soit le mode de calcul de la part variable des rémunérations.

2. Utiliser le test de Jonckheere-Terpstra puis la statistique de Kruskal-Wallis afin de tester l'hypothèse selon laquelle la distribution du score d'implication est la même quelque soit le mode de calcul de la part variable des rémunérations (prendre un niveau de signification de 5 %).

© Dunod. La photocopie non autorisée est un délit.

16 TESTS SUR LA RÉGRESSION LINÉAIRE

Un bon modèle prévisionnel est un élément essentiel d'une bonne gestion. Ainsi, connaissant les ventes $Y(t_i)$ au cours des périodes précédentes où $i = 1,2,\ldots,n$, il peut être utile en raison d'aléas d'avoir une estimation par intervalle de confiance de la quantité $Y(t_{n+1})$ qui sera vendue au cours de la période à venir $(n+1)$. Sont présentés des modèles linéaires simples et multiples où sont introduits des aléas. Le lecteur doit préalablement se reporter au chapitre 2 avant de lire ce chapitre.

Section 1 ■ Régression d'une *v.a. Y* sur une variable certaine x

Section 2 ■ Régression d'une *v.a. Y* sur une variable aléatoire X

Section 3 ■ Test de Durbin-Watson d'autocorrélation des erreurs

Section 4 ■ Régression d'une *v.a. Y* sur k variables certaines x_i

Section 5 ■ Multicolinéarité

Section 6 ■ Traitements sous Excel et SPSS

Section 1

RÉGRESSION D'UNE *V.A. Y* SUR UNE VARIABLE CERTAINE *x*

1 Le modèle avec résidus aléatoires

On s'intéresse au cas où une variable aléatoire Y dépend de façon affine d'une variable certaine $\boldsymbol{x}$:

$$Y = ax + b + \varepsilon$$

où a et b sont des constantes réelles de valeurs inconnues et où ε désigne l'aléa ou erreur aléatoire.

Ce type de modèle s'applique notamment au cas d'expérimentations où la variable x est contrôlée par l'expérimentateur.

Partant de n observations (x_i, y_i), considérons le modèle où Y_i la variable aléatoire endogène prend la valeur y_i égale $ax_i + b$ aux aléas près notés ε_i :

$$Y_i = ax_i + b + \varepsilon_i$$

où x_i est la i-ème observation de la variable exogène certaine $\boldsymbol{x}$, a et b sont des nombres certains de valeurs inconnues.

Les variables erreurs ε_i sont des variables aléatoires supposées être indépendantes et suivre une même loi normale centrée $\mathcal{N}(0; \sigma_0^2) : \forall\, i = 1,2,\ldots,n\ E(\varepsilon_i) = 0$ et $V(\varepsilon_i) = \sigma_0^2$ (hypothèse d'homoscédasticité).

Sous ces hypothèses, on peut vérifier que Y_i suit la loi normale $\mathcal{N}(ax_i + b; \sigma_0^2)$:

Les différentes *v.a.* $Y_1 = ax_1 + b + \varepsilon_1, Y_2 = ax_2 + b + \varepsilon_2, \ldots, Y_n = ax_n + b + \varepsilon_n$ qui correspondent respectivement aux valeurs prédéterminées $x_1, x_2, \ldots, x_n$, sont indépendantes puisque les *v.a* $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ sont supposées indépendantes.

© Dunod. La photocopie non autorisée est un délit.

Exemple

Les consommations annuelles d'électricité en 10^{10} Kwh au cours des 7 précédentes années ont été les suivantes (t désigne le numéro d'ordre de l'intervalle de temps).

Années	2000	2001	2002	2003	2004	2005	2006
Quantité q_t	7,38	8,05	8,68	9,34	9,51	9,62	11,14
t	1	2	3	4	5	6	7

La représentation graphique des points $M_t = (t, q_t)$ permet de constater que ces 7 points sont sensiblement alignés. La droite d'ajustement a pour équation « $q'_t = 0{,}545 \times t + 6{,}92$ » car $\overline{t} = 4,\ V(t) = 4,\ \overline{q} = 9{,}10,\ \mathrm{Cov}(t, q_t) = 2{,}179$. Les valeurs ajustées q'_t et les résidus $e_t = q'_t - q_t$ figurent ci-après :

$q'_t = 0{,}545 \times t + 6{,}92$	7,469	8,014	8,558	9,103	9,648	10,192	10,737	
$e_t = q'_t - q_t$	– 0,090	0,040	0,120	0,240	– 0,140	– 0,70	0,400	$\sum e_t = 0$
e_t^2	0,008	0,002	0,014	0,058	0,020	0,325	0,160	$\sum e_t^2 = 0{,}586$

$V(q_t) = 1{,}27$, $r^2 = 0{,}9337 = [\mathrm{Cov}(t, q_t)]^2 / [V(t) \times V(qt)]$, $V_R(q_t) = \dfrac{1}{7}\sum_{t=1}^{7} e_t^2 = 0{,}084$.

La valeur de r^2 étant proche de 1, on peut estimer que l'ajustement linéaire est de bonne qualité. La prévision de consommation pour l'année 2007 doit donc être une valeur proche de $q'_8 = 11{,}28 = 0{,}545 \times 8 + 6{,}92$.

L'estimation de l'écart entre cette valeur centrale de 11,28 et la valeur q_8 qui sera effectivement consommée, nécessite l'introduction des variables aléas. La quantité consommée Q_t au cours de l'intervalle de temps t est définie par la relation

$$Q_t = at + b + \varepsilon_t$$

où a et b sont des constantes de valeurs inconnues qui définissent le trend et ε_t est une variable aléatoire dont la valeur dépend de caractères conjoncturels tels que les conditions climatiques, les variations du taux de croissance de la production par rapport à son propre trend, etc. Cette variable aléatoire est nécessairement centrée : $E(\varepsilon_t) = 0$. Les coefficients a_0 et b_0 déduits de la régression linéaire par la méthode des moindres carrés sont respectivement des estimations des valeurs numériques de a et de b. La valeur ε_t^* prise par la variable aléatoire ε_t est appelée résidu.

2 Tests et intervalles de confiance des coefficients

2.1 Estimateurs de a et b ainsi que de la variance σ_0^2 des variables aléas ε_i

Par la technique des moindres carrés ordinaires, on détermine les estimateurs $\hat{\mathbf{a}}$ et $\hat{\mathbf{b}}$ de a et b qui minimisent la somme des carrés des erreurs $\sum_{i=1}^{n} \hat{e}_i^2$.

• $\hat{\mathbf{a}} = \dfrac{\mathrm{Cov}(x, Y)}{V(x)} = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ (où $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ et $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$) est un estimateur convergent sans biais de a, il prend la valeur a_0 déduite de la régression linéaire par la méthode des moindres carrés ordinaires (*cf.* repère p. 53). Sa distribution est normale : $\hat{\mathbf{a}} \sim> \mathcal{N}\left(a; \dfrac{\sigma_0^2}{nV(x)}\right)$.

• $\hat{\mathbf{b}} = \overline{Y} - \hat{\mathbf{a}}\overline{x}$ est un estimateur convergent sans biais de b ; il prend la valeur b_0 déduite de la régression linéaire par la méthode des moindres carrés ordinaires, sa

distribution est normale : $\hat{\mathbf{b}} \sim> \mathcal{N}\left(b; \frac{\sigma_0^2 \sum_{i=1}^n x_i^2}{n^2 V(x)}\right)$. Après expérimentation, $\hat{\mathbf{b}}$ prend la valeur b_0.

- La *v.a.* $\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$ (où $\hat{e}_i = Y_i - \hat{\mathbf{a}}x_i - \hat{\mathbf{b}}$), qui prend pour valeur $V_R(y)$ la variance résiduelle déduite de la régression linéaire, est un estimateur biaisé de σ_0^2. Elle est indépendante de $\hat{\mathbf{a}}$ et $\hat{\mathbf{b}}$. La statistique $\frac{n}{\sigma_0^2} \times \hat{\sigma}_\varepsilon^2$ suit la loi du Khi-deux à $(n-2)$ degrès de liberté et est utilisée pour obtenir des intervalles de confiance de σ_0^2. La *v.a.* $\theta_\varepsilon^2 = \frac{n}{n-2}\hat{\sigma}_\varepsilon^2$ est quant à elle un estimateur convergent sans biais de la variance σ_0^2 soit $E(\theta_\varepsilon^2) = \sigma_0^2$.

Le coefficient de détermination ou coefficient de corrélation au carré R^2

$$R^2 = \left[\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})\right]^2 \Big/ \left[\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2 \times \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2\right].$$

qui prend pour valeur r^2 mesure la qualité de l'ajustement.

2.2 Tests et intervalles de confiance des coefficients

L'écart-type σ_0 étant généralement inconnu, on utilise les propriétés *i)* et *ii)* pour définir des intervalles de confiance des valeurs de a, b ou réaliser des tests d'hypothèse concernant leurs valeurs respectives.

Propriétés

i) $(\hat{\mathbf{a}} - a) \times \sqrt{(n-2)V(x)}/\hat{\sigma}_\varepsilon = \boldsymbol{t}_{n-2}$ (variable de Student)

ii) $(\hat{\mathbf{b}} - b) \times \sqrt{(n-2)\frac{V(x)}{V(x) + \overline{x}^2}} \Big/ \hat{\sigma}_\varepsilon = \boldsymbol{t}_{n-2}$

iii) $R^2/(1-R^2)$ cette variable prend pour valeur $V_E(y)/V_R(y) =$ variance expliquée/variance résiduelle.

iv) sous l'hypothèse « $a = 0$ » on a $\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \boldsymbol{t}_{n-2}$

Les propriétés *i)* et *iv)* permettent également de tester H_0 « $a = 0$ » contre $\overline{H}_0$ « $a \neq 0$ ».

© Dunod. La photocopie non autorisée est un délit.

Exemple. Tests et intervalles de confiance des valeurs *a*, *b* et test sur la qualité d'ajustement.

Analysons l'exemple introductif relatif aux consommations annuelles d'électricité.

Pour tester $\mathbf{H_0}$ « $a = 0$ » contre $\mathbf{H_1}$ « $a > 0$ » avec un niveau de signification $\alpha = 0{,}05$ » on utilise la propriété « $(\hat{\mathbf{a}} - a) \times \sqrt{(n-2)V(t)}/\hat{\sigma}_\varepsilon = \boldsymbol{t}_{n-2}$ où $n = 7$ » et donc sous l'hypothèse H_0 la variable statistique $T_0 = (\hat{\mathbf{a}} - 0) \times \sqrt{(n-2)V(t)}/\hat{\sigma}_\varepsilon = \boldsymbol{t}_5$.

Pour déterminer de domaine de rejet de H_0 on se place sous l'hypothèse alternative : sous H_1, $\hat{\mathbf{a}} \xrightarrow{\text{pr}} a > 0$ lorsque $n \to \infty$, donc T_0 tend à prendre de grandes valeurs positives. Par suite le domaine de rejet de H_0 est du type $[c_\alpha, \infty[$ où c_α est tel que $0{,}05 = P_0(T_0 \geqslant c_\alpha) = P(\boldsymbol{t}_5 \geqslant c_\alpha)$, soit $c_\alpha = 2{,}015$. On prend la décision D_1 de rejeter H_0 au profit de H_1 lorsque la valeur numérique t_0^* prise par T_0 est supérieure à c_α. On a (*cf.* début du chapitre) $a_0 = 0{,}545$, $V(t) = 4$, $V_R(y) = (\sum e_t^2)/n = (0{,}586)/7$, donc $t_0^* = 0{,}545 \times \sqrt{(7-2) \times 4}/\sqrt{0{,}586/7} = 8{,}42 > 2{,}015$ et on rejette H_0 au profit de H_1.

Pour construire un intervalle de confiance de *b* de niveau 0,95, on utilise la propriété :

$$(\hat{\mathbf{b}} - b) \times \sqrt{(n-2)\frac{V(t)}{V(t)+\bar{t}^2}}/\hat{\sigma}_\varepsilon = \boldsymbol{t}_{n-2} \text{ avec } n = 7. \text{ On a } P(-2{,}57 \leqslant \boldsymbol{t}_5 \leqslant 2{,}57) = 0{,}95,$$

$$P(-2{,}57 \leqslant (\hat{\mathbf{b}} - b) \times \sqrt{(5)\frac{V(t)}{V(t)+\bar{t}^2}}/\hat{\sigma}_\varepsilon \leqslant 2{,}57)$$

$$= P(-2{,}57 \times \hat{\sigma}_\varepsilon/\sqrt{(5)\frac{V(t)}{V(t)+\bar{t}^2}} + \hat{\mathbf{b}} \leqslant b \leqslant 2{,}57 \times \hat{\sigma}_\varepsilon/\sqrt{(5)\frac{V(t)}{V(t)+\bar{t}^2}} + \hat{\mathbf{b}}) = 0{,}95.$$

Avec un niveau de confiance de 95 %, on considère que l'inégalité

$$-2{,}57 \times 0{,}289/\sqrt{(5)\frac{4}{4+4^2}} + 6{,}92 \leqslant b \leqslant +2{,}57 \times 0{,}289/\sqrt{(5)\frac{4}{4+4^2}} + 6{,}92 \text{ est vraie,}$$

soit $6{,}17 \leqslant b \leqslant 7{,}67$.

3 Modèle prévisionnel et erreur de prévision

L'erreur de prévision constatée ou écart entre y'_{n+1} la valeur estimée par régression linéaire pour la période future $n+1$ (ou pour une $n+1$-ème observation qui sera associée à x_{n+1}) et la valeur effectivement constatée y_{n+1} est notée z_{n+1} :

$$z_{n+1} = y_{n+1} - y'_{n+1} \text{ où}$$

- y_{n+1} désigne la valeur que prendra la *v.a.* $Y_{n+1} = ax_{n+1} + b + \varepsilon_{n+1}$ à la « date à venir » x_{n+1},
- $y'_{n+1} = a_0 x_{n+1} + b_0$ est la valeur attendue pour une valeur prédéterminée x_{n+1} par application des coefficients a_0 et b_0 calculés à partir des seules n premières observations $x_1, x_2, \ldots, x_n$ et correspondant à la réalisation de $Y'_{n+1} = \hat{\mathbf{a}}x_{n+1} + \hat{\mathbf{b}}$.

L'erreur de prévision z_{n+1} est la valeur prise par la *v.a.* Z_{n+1} nommée écart prévisionnel aléatoire : $Z_{n+1} = Y_{n+1} - Y'_{n+1} = Y_{n+1} - (\hat{\mathbf{a}}x_{n+1} + \hat{\mathbf{b}})$.

Afin de déterminer un intervalle de confiance des erreurs de prévision z_{n+1} et de la valeur y_{n+1} que prendra Y_{n+1} on utilise la statistique $\dfrac{Z_{n+1}\sqrt{n-2}}{\hat{\sigma}_\varepsilon\sqrt{n+1+\delta}} = \boldsymbol{t}_{n-2}$ où $\delta = (x_{n+1} - \overline{x})^2 / V(x)$ et $\overline{x} = (x_1 + \ldots + x_n)/n$.

Exemple

Dans l'exemple introductif (le 1 de la section 1) on souhaite prévoir pour l'année 2007 ($t = 8$) la consommation d'électricité : $q'_t = 0{,}545 \times t + 6{,}92 = 11{,}28$. Un intervalle de confiance de la prévision de niveau 0,95 est obtenu en utilisant la propriété

$$\text{« } \frac{Z_{n+1}\sqrt{n-2}}{\hat{\sigma}_\varepsilon\sqrt{n+1+\delta}} = \boldsymbol{t}_{n-2} \text{ où } Z_{n+1} = Y_{n+1} - \hat{\mathbf{a}}t_{n+1} - \hat{\mathbf{b}} \text{ et } n = 7 \text{ ».}$$

Or $P(-2{,}57 \leqslant \boldsymbol{t}_5 \leqslant 2{,}57) = 0{,}95$ donc $P\left(-2.57 \leqslant \dfrac{Z_{7+1}\sqrt{7-2}}{\hat{\sigma}_\varepsilon\sqrt{7+1+\delta}} \leqslant 2{,}57\right) = 0{,}95$.

La *v.a.* $\dfrac{Z_8\sqrt{5}}{\hat{\sigma}_\varepsilon\sqrt{8+\delta}}$ prend la valeur $(y_8 - 0{,}545 \times 8 - 6{,}92) \times \sqrt{5}/[(0{,}289) \times \sqrt{1}2]$ (car $\delta = (t_{n+1} - \overline{t})^2 / V(t) = (8-4)^2/4 = 4$), donc avec un niveau de confiance de 95 % on a l'inégalité « $-2{,}57 \leqslant (y_8 - 0{,}545 \times 8 - 6{,}92) \times \sqrt{5}/(0{,}289) \times \sqrt{1}2 \leqslant 2{,}57$ » c'est-à-dire $10{,}13 \leqslant y_8 \leqslant 12{,}43$.

Section 2 RÉGRESSION D'UNE *V.A. Y* SUR UNE VARIABLE ALÉATOIRE *X*

À chaque valeur x prise par X on peut en général associer le nombre $E(Y/X = x)$ et on détermine ainsi la fonction de régression $\Psi(x) = E(Y/X = x)$ qui est définie sur Δ_X (*Cf.* chap. 13 p. 228)

Ainsi est déterminée la *v.a.* $\Psi(X) = E(Y/X)$ qui, lorsque $X = x$, prend la valeur $\Psi(x)$ et l'on a $E[(Y - \Psi(X))^2] \leqslant E[(Y - f(X))^2]$ quelque soit la fonction réelle $f(x)$ définie sur Δ_X.

Rappel. $E[(Y - f(X))^2] = \sum_{i=1}^{m}\sum_{j=1}^{n}((y_j - f(x_i))^2 p_{ij}$ lorsque (X,Y) suit une loi discrète (pour une loi continue *cf.* annexe p. 367)

Lorsque le phénomène aléatoire représenté par X aide à prévoir le phénomène aléatoire Y, on considère le modèle $Y = E(Y/X) + \varepsilon$ où ε est un résidu aléatoire,

© Dunod. La photocopie non autorisée est un délit.

souvent considéré comme la partie de Y non expliquée par X. Ce résidu a les propriétés suivantes : $E(\varepsilon) = 0$; $V(\varepsilon) = V(Y) - V[\Psi(X)]$; $\text{Cov}(X,\varepsilon) = 0$.

La qualité de l'approximation de Y par $\Psi(X)$ est mesurée par le rapport de corrélation $\eta^2_{Y/X} = V[\Psi(X)]/V(Y)$ prenant pour valeur $\dfrac{V_E(y)}{V(y)}$ = Variance expliquée/ variance totale.

Lorsque $E(Y/X)$ est une fonction affine de X du type $E(Y/X) = aX + b$, on a le modèle $Y = aX + b + \varepsilon$ où $\text{Cov}(X,\varepsilon) = 0$. Ce modèle peut être réécrit

$$\frac{Y - E(Y)}{\sigma(Y)} = r\frac{X - E(X)}{\sigma(X)} + \sqrt{1 - r^2}\,\varepsilon$$

où $r = r(X,Y)$ est le coefficient de corrélation linéaire, $E(\varepsilon) = 0$, $V(\varepsilon) = 1$ et $\text{Cov}(X,\varepsilon) = 0$.

Si on admet que ε suit une loi normale centrée $\mathcal{N}(0; \sigma_0^2)$ alors la loi de probabilité conditionnelle de Y « sachant que X prend la valeur x » est la loi normale $\mathcal{N}(ax + b; \sigma_0^2)$.

Par suite, si on dispose de n couples (x_i, y_i) correspondant aux valeurs prises par un couple (X,Y) on a $E(Y/X = x_i) = ax_i + b$, $V(Y/X = x_i) = \sigma_0^2$ et conséquemment $Y/X = x_i$ suit la loi normale $\mathcal{N}(ax_i + b; \sigma_0^2)$. Les estimateurs de a et b et les tests sur les coefficients sont les mêmes que ceux présentés dans le § 1 section 1.

Section 3

TEST D'AUTOCORRÉLATION DES ERREURS

Le modèle précédent suppose l'indépendance des variables aléas. Or, elles le sont rarement lorsque la variable explicative est le temps. Le test de Durbin-Watson teste la nullité du coefficient de corrélation ρ entre deux aléas consécutifs ε_{i-1} et ε_i. Considérant le modèle $\varepsilon_i = \rho \times \varepsilon_{i-1} + \eta_i$ où la constante ρ vérifie $|\rho| < 1$ et où les variables normales η_i sont centrées, indépendantes, de même écart-type σ_η, on teste à l'aide de la statistique DW de Durbin-Watson l'hypothèse H_0 « $\rho = 0$ ou absence d'autocorrélation d'ordre 1 » contre $\overline{H}_0$ « $\rho \neq 0$ ».

Procédure du test. La statistique $DW = \sum_{i=2}^{n}(\hat{e}_i - \hat{e}_{i-1})^2 / \sum_{i=1}^{n}\hat{e}_i^2$ est un estimateur de $2 \times (1 - \rho)$. Elle varie entre 0 et 4, prend une valeur proche de 2 lorsque $\rho = 0$. (rappelons que $\hat{e}_i = Y_i - \hat{\mathbf{a}}x_i - \hat{\mathbf{b}}$ prend la valeur $e_i = y'_i - y_i$ qui est l'écart entre la valeur estimée $y'_i = a_0x_i + b_0$ et la valeur effective y_i).

La règle de décision. Schématisée dans le tableau ci-dessous, elle est fondée sur la comparaison de la valeur numérique $d^* = \sum_{i=2}^{n}(e_i - e_{i-1})^2 / \sum_{i=1}^{n}e_i^2$ prise par DW

avec les valeurs critiques d_1 et d_2 obtenues par lecture de table (p. 389) en fonction de la taille de l'échantillon n, du nombre k de variables explicatives et du risque d'erreur.

Valeur d^*	$0 < d^* \leqslant d_1$	$d_1 < d^* \leqslant d_2$	$d_2 < d^* \leqslant 4 - d_2$	$4 - d_2 < d^* \leqslant 4 - d_1$	$4 - d_1 < d^* \leqslant 4$
Décision	$E(\varepsilon_i \varepsilon_{i+1}) > 0$	doute	indépendance des ε_i	doute	$E(\varepsilon_i \varepsilon_{i+1}) < 0$

Exemple

Reprenant l'exemple introductif on peut calculer la valeur d^* prise par la statistique *DW* de Durbin-Watson

t	1	2	3	4	5	6	7	
$e_i = q_t' - q_t$	– 0,090	0,040	0,120	0,240	– 0,140	–0,70	0,400	$\sum e_t = 0$
e_t^2	0,008	0,002	0,014	0,058	0,020	0,325	0,160	$\sum e_t^2 = 0{,}586$
$(e_t - e_{t-1})^2$		0,017	0,006	0,014	0,144	0,185	0,941	$\sum_t (e_t - e_{t-1})^2 = 1{,}308$

$d^* = 1{,}308/0{,}586 = 2{,}232$. Le nombre de variables explicatives est $k = 1$, le nombre d'observations est $n = 7$ et le niveau de signification retenu $\alpha = 0{,}05$. Par lecture de table on déduit que $d_1 = 0{,}700$ et $d_2 = 1{,}356$ et donc $4 - d_1 = 3{,}3$ et $4 - d_2 = 3{,}644$. $d_2 < d^* \leqslant 4 - d_2$ on décide donc d'accepter l'hypothèse d'indépendance.

Remarque. Ce test suppose que $(\varepsilon_1, \ldots, \varepsilon_n)$ suit une loi normale centrée de dimension n.

Section **4**

RÉGRESSION D'UNE *V.A. Y* SUR *K* VARIABLES CERTAINES X_I

1 Régression linéaire sur *k* variables explicatives

Considérons la série statistique composée de n observations où y est la variable endogène ou expliquée et $x_1, x_2, \ldots, x_k$ les variables exogènes ou explicatives :

Observations		1	2	•	n
Valeurs des variables explicatives	x_1	x_{11}	x_{12}	•	x_{1n}
	x_2	x_{21}	x_{22}	•	x_{2n}
	...	...	...	•	...
	x_k	x_{k1}	x_{k2}	•	x_{kn}
Valeurs de la variable expliquée	y	y_1	y_2	•	y_n

© Dunod. La photocopie non autorisée est un délit.

On a un nuage de n points $M_1 = (x_{11},x_{21},\ldots,x_{k1},y_1)$, $M_2 = (x_{12},x_{22},\ldots,x_{k2},y_2)$, $\ldots$, $M_n = (x_{1n},x_{2n},\ldots,x_{kn},y_n)$. Si ce nuage est situé à proximité d'un plan d'équation $y = a_1x_1 + a_2x_2 + \ldots + a_kx_k + a_{k+1}$ on peut ajuster la série en substituant à y_i la valeur $y'_i = a_1x_{1,i} + a_2x_{2,i} + \ldots + a_kx_{k,i} + a_{k+1}\ \forall\, i = 1,\ldots,n$.

Les coefficients $(a_1^0,a_2^0,\ldots,a_k^0,a_{k+1}^0)$ retenus selon la méthode des moindres carrés, sont les réels qui minimisent la fonction

$$e(a_1,a_2,\ldots,a_k,a_{k+1}) = \sum_{i=1}^{n}[y_i - (a_1x_{1i} + a_2x_{2i} + \ldots + a_kx_{ki} + a_{k+1})]^2$$

Selon la méthode des moindres carrés la valeur estimée par l'ajustement linéaire est $y'_i = a_1^0x_{1i} + \ldots + a_k^0x_k + a_{k+1}^0$.

Écriture matricielle. On cherche le plan d'équation $y = a_1x_1 + a_2x_2 + \ldots + a_kx_k + a_{k+1}$ qui minimise la fonction $e(a_1,a_2,\ldots,a_k,a_{k+1})$.

La matrice ligne de la variable expliquée étant notée $\boldsymbol{Y} = (y_1,y_2,\ldots,y_n)$, la matrice des variables explicatives étant notée $X = \begin{bmatrix} x_{11} & x_{12} & \cdot & \cdot & x_{1n} \\ x_{21} & x_{22} & \cdot & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{k1} & x_{k2} & \cdot & \cdot & x_{kn} \\ 1 & 1 & \cdot & \cdot & 1 \end{bmatrix}$, alors :

– la matrice ligne des coefficients obtenus par la méthode des moindres carrés est $\mathbf{a^0} = (a_1^0,a_2^0,\ldots,a_k^0,a_{k+1}^0)$ avec $\mathbf{a^0} = \boldsymbol{Y} \cdot {}^tX \cdot (X \cdot {}^tX)^{-1}$.

– la matrice ligne des valeurs attendues $\boldsymbol{Y}' = (y'_1,y'_2,\ldots,y'_n)$ est obtenue en faisant le produit $\boldsymbol{Y}' = \mathbf{a^0} \cdot X$,

– la matrice ligne des résidus $\boldsymbol{e} = (e_1,e_2,\ldots,e_n) = \boldsymbol{Y}' - \boldsymbol{Y}$.

2 Le modèle de régression linéaire multiple avec aléas

Le modèle concernant la régression linéaire d'une *v.a.* Y sur un k-uple de variables certaines $(x_1,x_2,\ldots,x_k)$ se caractérise par l'ajustement :

$$Y_i = a_1x_{1i} + a_2x_{2i} + \ldots + a_kx_{ki} + a_{k+1} + \varepsilon_i \quad \forall\, i = 1,2,\ldots,n$$

où les constantes $a_1,a_2,\ldots,a_{k+1}$ ont des valeurs inconnues, les variables aléas $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n$ sont indépendantes et suivent une même loi normale centrée $\mathcal{N}(0;\sigma_0^2)$, la valeur de la constante σ_0 étant inconnue.

Le modèle s'exprime de façon matricielle par la relation $Y = \boldsymbol{a} \cdot X + \varepsilon$ où

– la matrice ligne de la variable expliquée $Y = (Y_1,Y_2,\ldots,Y_n)$ est un n-uple aléatoire ;

– la matrice ligne des coefficients du modèle est $\boldsymbol{a} = (a_1,a_2,\ldots,a_{k+1})$,

– la matrice des variables explicatives est $X = \begin{bmatrix} x_{11} & x_{12} & \cdot & \cdot & x_{1n} \\ x_{21} & x_{22} & \cdot & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{k1} & x_{k2} & \cdot & \cdot & x_{kn} \\ 1 & 1 & \cdot & \cdot & 1 \end{bmatrix}$;

– la matrice ligne des résidus aléatoires est $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$.

3 Tests et intervalles de confiance des coefficients

On se propose de tester le caractère significatif de la régression correspondant au test de l'hypothèse H « $Y_i = a_1 x_{1i} + a_2 x_{2i} + \ldots + a_k x_{ki} + a_{k+1} + \varepsilon_i \quad \forall\, i$ », de trouver un intervalle de confiance de a_i et de tester l'hypothèse $a_i = 0$, enfin de donner un intervalle de confiance pour la prévision Y_{n+1}.

3.1 Estimateurs des coefficients (a_1, a_2, …, a_{k+1}) et de σ_0^2

– L'estimateur de $(a_1, a_2, \ldots, a_{k+1})$ est la matrice ligne ou $(k+1)$-uple aléatoire $\hat{\mathbf{a}} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k, \hat{a}_{k+1})$ défini par $\hat{\mathbf{a}} = Y \cdot {}^t X \cdot (X \cdot {}^t X)^{-1}$. Il prend pour valeur $\mathbf{a^0} = (a_1^0, a_2^0, \ldots, a_{k+1}^0)$ déduit de l'ajustement des n valeurs prises $(y_1, y_2, \ldots, y_n)$ sur les variables $(x_1, x_2, \ldots, x_k)$ selon méthode des moindres carrés ordinaires.

– L'estimateur sans biais de σ_0^2 est la *v.a.* $\theta_n^2 = \dfrac{1}{n-k-1} \times \sum_{i=1}^{n} \hat{e}_i^2$ où $\hat{e}_i = Y_i - Y_i'$ est l'écart aléatoire entre la valeur aléatoire Y_i et la valeur estimée par le modèle $Y_i' = \hat{a}_1 x_{1i} + \hat{a}_2 x_{2i} + \ldots + \hat{a}_k x_{ki} + \hat{a}_{k+1}$

– Le coefficient de détermination est noté $\hat{R}^2 = \dfrac{\sum_{i=1}^{n} (Y_i' - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} = 1 - \dfrac{\sum_{i=1}^{n} \hat{e}_i^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$,

3.2 Propriétés

Désignant par $\Omega = X \cdot {}^t X = (\omega_{ij})$ la matrice réelle symétrique dont le terme général ω_{ij} appartient à la i-ème ligne et à la j-ème colonne et par $\Omega^{-1} = (\omega'_{ij})$ sa matrice inverse, on a les propriétés suivantes :

i) la statistique $(\hat{a}_i - a_i)/(\theta_n \times \sqrt{\omega'_{ii}})$ suit la loi de Student à $(n-k-1)$ degrés de liberté : $\dfrac{(\hat{a}_i - a_i)}{\theta_n \sqrt{\omega'_{ii}}}) = \boldsymbol{t}_{n-k-1} \quad \forall\, i = 1, 2, \ldots, k+1$.

© Dunod. La photocopie non autorisée est un délit.

ii) la statistique $\dfrac{1}{k+1}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}(\hat{a}_i - a_i)(\hat{a}_j - a_j)/\theta_n^2 = F_{n-k-1}^{k+1}$ (variable de Fisher)

iii) sous l'hypothèse « $a_1 = a_2 = \ldots a_{k+1} = 0$ », la statistique $F = \dfrac{\hat{R}^2/k}{(1-\hat{R}^2)/(n-k-1)}$ qui prend pour valeur numérique $f^* = \dfrac{V_E(y)/k}{V_R(y)/(n-k-1)}$ suit une loi de Fisher-Snédécor $F(k, n-k-1)$.

La propriété *i)* est utilisée pour trouver un intervalle de confiance de la valeur a_i ou réaliser un test de signification de chaque coefficient. La propriété *iii)* est utilisée pour réaliser un test global d'ajustement permettant de rejeter le modèle si Y' est trop éloignée de Y car elle permet de tester l'hypothèse H_0 « $a_1 = a_2 = \ldots a_{k+1} = 0$ ».

Exemple

Une entreprise s'intéresse au lien entre les ventes (Y) et les dépenses publicitaires de télévision (x_1) et de radio (x_2). À cette fin, elle fait varier au cours de 13 trimestres ces dépenses publicitaires et observe leur effet sur les ventes.

Obs.	1	2	3	4	5	6	7	8	9	10	11	12	13
VENTE y_i	221	236	224	269	240	263	261	278	281	320	321	330	309
PUBT $x_{1,i}$	20	30	24	36	26	36	35	44	42	54	51	60	55
PUBR $x_{2,i}$	18	14	11	18	15	16	17	13	20	24	28	19	7

Le directeur marketing souhaite estimer le modèle : $Y_i = a_1 x_{1,i} + a_2 x_{2,i} + a_3 + \varepsilon_i$ $\forall\, i = 1,2,\ldots,13$ où $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{13}$ sont supposées indépendantes et suivre une même loi normale centrée $\mathcal{N}(0; \sigma_0^2)$. Le modèle peut s'exprimer sous la forme matricielle $Y = \boldsymbol{a} \cdot X + \varepsilon$ où $Y = (221,236,\ldots,309)$, $\boldsymbol{a} = (a_1, a_2, a_3)$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{13})$,

$$X = \begin{bmatrix} 20 & 30 & \ldots & 55 \\ 18 & 14 & \ldots & 7 \\ 1 & 1 & \ldots & 1 \end{bmatrix} \text{ et}$$

$$(X \cdot {}^tX)^{-1} = (\omega'_{ij}) = \begin{pmatrix} 0{,}000551 & -0{,}000365 & -0{,}015580 \\ -0{,}000365 & 0{,}003092 & -0{,}0379035 \\ -0{,}015580 & -0{,}0379035 & 1{,}3331972 \end{pmatrix}$$

Les estimations des coefficients (a_1, a_2, a_3) sont définies par $\mathbf{a^0} = (a_1^0, a_2^0, a_3^0) = Y \cdot {}^tX \cdot (X \cdot {}^tX)^{-1} = (2{,}75,\ 1{,}14,\ 145{,}44)$. On en déduit les n valeurs estimées $y'_i = 2{,}75x_{1,i} + \ 1{,}14x_{2,i} + 145{,}44$ et les résidus correspondant $e_i = y_i - y'_i$:

y'_i	221,01	243,95	224,02	265,02	234,09	262,73	261,12	281,31	283,8	321,38	317,7	332,17	304,71
e_i	– 0,01	– 7,95	– 0,02	3,98	5,91	0,27	– 0,12	– 3,31	– 2,80	– 1,38	3,30	– 2,17	4,29

La somme des carrés des résidus $\sum_{i=1}^{13} e_i^2 = (-0{,}01)^2 + \ldots + 4{,}29^2 = 168{,}79$ et $\theta_n^2 = \dfrac{1}{13-2-1} \times \sum_{i=1}^{13} \hat{e}_i^2$ prend la valeur 16,88.

– Pour tester par exemple, l'hypothèse H_0 « $a_1 = 0$ » contre $\overline{H}_0$ « $a_1 \neq 0$ » on utilise la propriété selon laquelle $\dfrac{(\hat{a}_i - a_i)}{\theta_n \sqrt{\omega'_{ii}}} = \boldsymbol{t}_{n-k-1}$, avec n égal à 13 et $k = 2$.

Sous H_0 « $a_1 = 0$ » on a $T_0 = \hat{a}_1/(\theta_n \sqrt{\omega'_{11}}) = \boldsymbol{t}_{10}$.

Compte tenu de $a_1^0 = 2{,}75$ et $\omega'_{11} = 0{,}000551$ on constate que T_0 prend après expérimentation la valeur $t_0^* = 2{,}75/(\sqrt{16{,}88} \times \sqrt{0{,}000551}) = 28{,}51$.

Le domaine de rejet de H_0 est du type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$ où $c_{\alpha/2}$ est tel que $1 - \alpha/2 = P_0(T_0 \leqslant c_{\alpha/2}) = P(\boldsymbol{t}_{10} \leqslant c_{\alpha/2})$. Pour $\alpha = 0{,}01$ on lit $c_{\alpha/2} = 3{,}17$. $t_0^* = 28{,}51$ appartient au domaine de rejet de H_0.

4 Évaluation prévisionnelle

Si on donne à chacune des variables certaines $x_1, x_2, \ldots, x_k$ de nouvelles valeurs numériques notées respectivement $x_{1,n+1}, x_{2,n+1}, \ldots, x_{k,n+1}$ alors

$$Y_{n+1} = a_1 x_{1,n+1} + a_2 x_{2,n+1} + \ldots + a_{k+1} + \varepsilon_{n+1}$$

prendra une valeur numérique y_{n+1}.

Cette valeur y_{n+1} diffèrera généralement de la valeur prévue $y'_{n+1} = a_1^0 x_{1,n+1} + a_2^0 x_{2,n+1} + \ldots + a_k^0 x_{k,n+1} + a_{k+1}^0$ calculée à partir des valeurs $x_{1,n+1}, x_{2,n+1}, \ldots, x_{k,n+1}$ et des coefficients $(a_1^0, a_2^0, \ldots, a_k^0, a_{k+1}^0)$ déterminés sur la base des n observations antérieures. y'_{n+1} est la réalisation de $Y'_{n+1} = \hat{a}_1 x_{1,n+1} + \hat{a}_2 x_{2,n+1} + \ldots + \hat{a}_k x_{k,n+1} + \hat{a}_{k+1}$.

Si la nouvelle variable aléas ε_{n+1} suit la loi $\mathcal{N}(0; \sigma_0^2)$ et est indépendante de $\varepsilon_1, \ldots, \varepsilon_n$ on a la propriété suivante : $\dfrac{(Y_{n+1} - Y'_{n+1})}{\theta_n \sqrt{1 + \sum_{i=1}^{k+1} \sum_{j=1}^{k+1} \omega'_{ij} x_{in+1} x_{jn+1}}} = \boldsymbol{t}_{n-k-1}$.

Cette statistique permet de construire des intervalles de confiance pour les valeurs prévisionnelles. Le lecteur se réfèrera au § 2 pour détecter une éventuelle autocorrélation des erreurs, la lecture de la table de Durbin-Watson nécessitant la prise en compte du nombre k de variables exogènes.

© Dunod. La photocopie non autorisée est un délit.

Section 5

MULTICOLINÉARITÉ

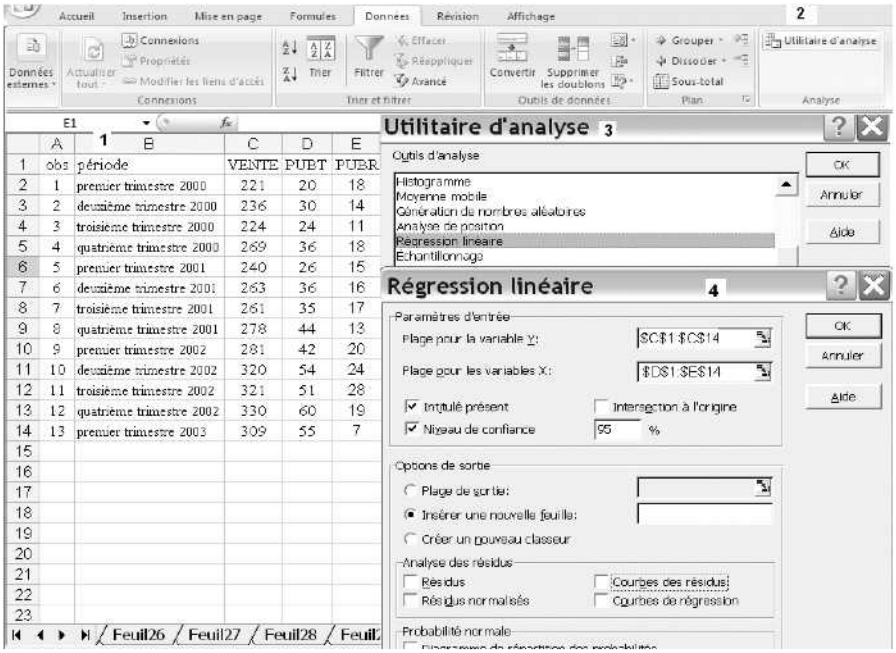
Les estimations de la méthode des moindres carrés sont fiables si les variables explicatives ne sont pas liées. Pour diagnostiquer cette multicolinéarité, Klein compare le coefficient de détermination R^2 calculé sur le modèle à k variables aux coefficients de corrélation entre variables explicatives. Si $R^2 < r^2_{x_i x_j}$ ($r^2_{x_i x_j}$ est le carré du coefficient de corrélation simple entre x_i et x_j) il y a présomption de multicolinéarité.

Section 6

TRAITEMENT SOUS EXCEL ET SPSS

1 Traitements statistiques avec Excel

Reprenons l'exemple précédent. Y désignant le montant des ventes VENTE, x_1 les dépenses publicitaires à la télévision PUBT et x_2 les dépenses publicitaires à la radio (PUBR), on souhaite estimer le modèle : $Y_i = a_1 x_{1i} + a_2 x_{2i} + a_3 + \varepsilon_i \forall i = 1,2,\ldots,13$ avec les variables aléas $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{13}$ supposées indépendantes et suivre une même loi normale centrée $\mathcal{N}(0; \sigma_0^2)$.



Procédure.

1. On entre les données, on se place dans une cellule quelconque non utilisée.

2. On clique sur (Données) puis sur (Utilitaire d'analyse).

3. Dans le menu **utilitaire d'analyse** on déroule et on clique sur (Regression linéaire) puis (OK).

4. Dans le menu **régression linéaire** on désigne dans plage pour la variable *Y* les cellules correspondant à l'intitulé et aux données de la variable expliquée VENTE soit (C1:C14) (de la ligne 1 de la colonne C à la ligne 14 de la colonne C) et dans plage pour les variables *X* les cellules correspondant à l'intitulé et aux données des variables explicatives PUBT et PUBR soit (D1:E14). On sélectionne (le niveau de confiance) et (résidus).

Résultats obtenus

RAPPORT DÉTAILLÉ		ANALYSE DE VARIANCE					
Statistiques de la régression			Degré de liberté	Somme des carrés	Moyenne des carrés	*F*	Valeur critique de *F*
Coefficient de détermination multiple	0,995	Régression	2,00	16 799,97	8 399,99	497,65	0,00
Coefficient de détermination R^2	0,990	Résidus	10,00	168,79	16,88		
Erreur-type	4,108	Total	12,00	16 968,77			
Observations	13,00						

Qualité de l'ajustement. Le coefficient de détermination prend la valeur $R^2 = V_E(y)/V_R(y) = 0{,}990$, $16\,799{,}97 = n \times V_E(y)$ où $V_E(y)$ est la variance expliquée par le modèle et n est égal à 13, $168{,}79 = n \times V_R(y)$ où $V_R(y)$ est la variance résiduelle ; $n \times V(y) = 16\,968{,}77$.

La statistique F prend pour valeur $f^* = \dfrac{16\,799{,}97/2}{168{,}79/10} = 8\,399{,}99/16{,}88 = 497{,}65$.

Testant l'hypothèse H_0 « tous les coefficient sont nuls » contre $\overline{H}_0$ on utilise la statistique F qui, sous H_0 suit la loi $\mathcal{F}(k; n-k-1)$ avec $k = 2$ et $n = 13$. Le niveau de signification observé du test $\alpha_c = 2 \times P_0(F > 497{,}65) = 2 \times P(F^2_{10} > 497{,}65) \cong 0{,}00$. La probabilité α_c de rejeter à tort H_0 étant négligeable, le modèle proposé est donc significatif.

Les coefficients du modèle.

	Coefficients	Erreur-type	Statistique *t*	Probabilité	Limite inférieure pour 95 %	Limite supérieure pour 95 %
PUBT	2,75	0,10	28,51	0,00	2,54	2,97
PUBR	1,14	0,23	5,00	0,00	0,63	1,65
Constante	145,44	4,74	30,66	0,00	134,87	156,01

© Dunod. La photocopie non autorisée est un délit.

Les estimateurs $(\hat{a}_1, \hat{a}_2, \hat{a}_3)$ des coefficients du modèle prennent la valeur (2,75, 1,14, 145,44). Pour $i = 1$ et 2 la variable $T_i = (\hat{a}_i - a_i)/(\theta_n \times \sqrt{\omega'_{ii}}) = \boldsymbol{t}_{n-k-1}$ avec ici $n = 13$ et $k = 2$.

Est testée pour chaque a_i, l'hypothèse H_0 « $a_i = 0$ » contre $\overline{H}_0$ « $a_i \neq 0$ ».

Sous H_0, $T_0 = \hat{a}_i/(\theta_n \sqrt{\omega'_{ii}}) = \boldsymbol{t}_{10}$. Le logiciel calcule la valeur t_0^* prise par T_0 et détermine le niveau de signification du test bilatéral $\alpha_c = 2 \times P_0(T_0 \geqslant |t_i^*|) = 2 \times P(\boldsymbol{t}_{10} \geqslant |t_0^*|)$.

Ainsi concernant le coefficient a_1, on observe que $\hat{a}_1$ prend pour valeur 2,75, que T_0 prend pour valeur $t_0^* = 2{,}75/0{,}10 = 28{,}51$ et donc $\alpha_c = 2 \times P(\boldsymbol{t}_{10} \geqslant 28{,}51) \cong 0{,}00$.

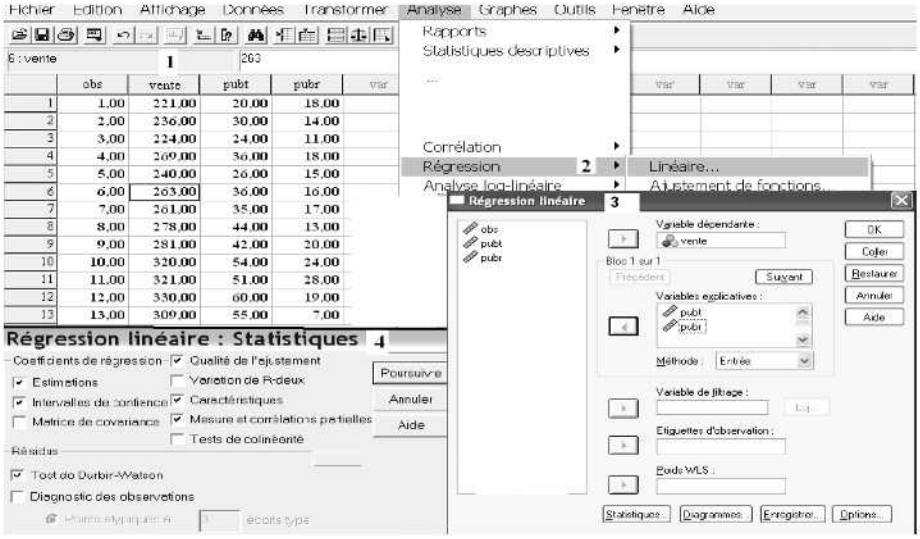
Le logiciel utilisant la variable T_0 détermine les intervalles de confiance avec un niveau de confiance fixé à 95 % de a_1 soit [2,54; 2,97], de a_2 soit [0,63; 1,65] et de a_3 soit [134,87; 156,01]. Il fournit également la liste des résidus e_i :

Analyse des résidus

Observation	1	2	3	4	5	6	7	8	9	10	11	12	13
Prévisions pour *Y*	221,0	243,9	224,0	265,0	234,1	262,7	261,1	281,3	283,8	321,4	317,7	332,2	221,0
Résidus	0,0	– 7,9	0,0	4,0	5,9	0,3	– 0,1	– 3,3	– 2,8	– 1,4	3,3	– 2,2	0,0

2 Traitements statistiques avec SPSS

On reprend l'exemple précédent.



Procédure.

1. Entrer les données.

2. Pointer sur (Analyse) puis (Regression) puis (Linéaire).

3. Dans le menu **régression linéaire** on sélectionne la variable *vente* que l'on met dans (variable dépendante) et les variables **pubt** et **pubr** en tant que (variables explicatives).

4. Cliquant sur le sous-menu (Statistiques) on sélectionne les options (estimations), (Intervalle de confiance), (Matrice de covariance) et (Test de Durbin-Watson)… On clique sur (Poursuivre) puis (OK).

Résultat : Récapitulatif du modèle

R	*R*-deux	*R*-deux ajusté	Erreur standard de l'estimation	Durbin-Watson
1,00	0,99	0,99	4,11	1,68
Valeurs prédites : (constantes), PUBR, PUBT Variable dépendante : VENTE				

Le coefficient de détermination $R^2 = 0{,}99$. Le R-deux ajusté $= 1 - (1 - R^2) \times [(n-1)/(n-k-1)]$ sert à apprécier la qualité de l'ajustement lorsque le nombre de degrès de liberté est faible. (En effet dans un modèle où le nombre d'observations est égal au nombre de variables explicatives on a $R^2 = 1$ alors même que le pouvoir explicatif du modèle est nul).

La statistique $DW = \sum_{i=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2 / \sum_{i=1}^{n} \hat{e}_i^2$ (où rappelons que $\hat{e}_i = Y'_i - Y_i$ est l'écart aléatoire entre la valeur estimée par le modèle $Y'_i = \hat{a}_1 x_{1,i} + \hat{a}_2 x_{2,i} + \hat{a}_3$ qui prend la valeur $y'_i = a_1^0 x_1 + a_2^0 x_2 + a_3^0$ et Y_i) prend la valeur 1,68. Cette valeur est comprise entre $d_2 = 1{,}038$ et $4 - d_2$, valeur lue sur la table pour un risque d'erreur de 5 % et un nombre de variables explicatives $k = 2$ (*cf.* p. 389). Aussi accepte-t-on l'hypothèse d'indépendance des ε_i.

ANOVA					
Modèle	**Somme des carrés**	**ddl**	**Carré moyen**	**F**	**Signification**
Régression	16 799,97	2,00	8 399,99	497,65	0,00
Résidu	168,79	10,00	16,88		
Total	16 968,77	12,00			

Pour l'interprétation des résultats se reporter au § 6.1, le tableau des coefficients de la régression est identique à celui d'Excel.

	Modèle	
		PUBT
Corrélations	PUBR	– 0,28

Les coefficients de corrélation entre variables explicatives étant très faibles on peut exclure tout problème de multicolinéarité.

© Dunod. La photocopie non autorisée est un délit.

Exercices (*cf.* corrections page 364)

Exercice

Un analyste financier souhaite estimer la relation entre l'évolution du produit net bancaire Y de la banque et l'évolution du temps t, les relevés portant sur 16 trimestres consécutifs depuis le début de l'année N.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Y	100	120	131	200	135	130	141	210	200	135	151	220	210	145	161	165

1. Estimer le modèle $Y_i = ax + b + \varepsilon_i \quad \forall\, i = 1,2,\ldots,16$ avec a et b des nombres certains de valeurs inconnues, les variables aléatoires $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_{16}$ étant supposées indépendantes et suivre une même loi normale centrée $\mathcal{N}(0;\sigma_0^2)$: $E(\varepsilon_i) = 0$; $V(\varepsilon_i) = \sigma_0^2\ \forall\, i = 1,2,\ldots,16$.

2. Tester à l'aide de la statistique de Durbin-Watson l'hypothèse d'autocorrélation des résidus.

3. Tester l'hypothèse H_0 « $a = 0$ » contre $\overline{H}_0$ « $a \neq 0$ » avec un niveau de signification $\alpha = 0{,}10$.

4. Donner un intervalle de confiance de a avec un niveau de confiance de 95 %.

5. Sachant qu'il peut exister un mouvement saisonnier, tester $Y_t = a_1 D_{1,t} + a_2 D_{2,t} + a_3 D_{3,t} + a_5 t + a_6 + \varepsilon_i\ \forall\, i = 1,2,\ldots,16$ où $D_{1,t}$ est une variable indicatrice qui prend la valeur 1 s'il s'agit du premier trimestre de l'année considérée et 0 dans le cas contraire, $D_{2,t}$ est une variable indicatrice qui prend la valeur 1 s'il s'agit du deuxième trimestre de l'année considérée et 0 dans le cas contraire, etc. Compte tenu de $\sum_{i=1}^{4} D_{i,t} = 1\ \forall\, t$, la variable indicatrice du quatrième trimestre n'est pas à inclure dans le modèle car elle est égale à $1 - (D_{1,j} + D_{2,j} + D_{3,j})$.

17 MODÈLES LOG-LINÉAIRE ET LOGIT

Les modèles Log-linéaires permettent d'analyser des tableaux de contingence à deux, trois ou *n* dimensions. Ils permettent notamment d'étudier les interactions entre deux ou plusieurs variables nominales. La régression logistique a pour objet de prévoir le comportement d'une variable dichotomique ou binaire à l'aide d'une ou de plusieurs variables (de nature quelconque) en prenant en compte l'effet propre de chaque variable mais également l'effet de leur interaction.

Section 1 ■ Modèle log-linéaire
Section 2 ■ Modèle de régression logistique de type logit
Section 3 ■ Traitement sous SPSS

© Dunod. La photocopie non autorisée est un délit.

Section 1

MODÈLE LOG-LINÉAIRE

1 Modèle log-linéaire pour un tableau 2×2

Exemple introductif

L'objet de l'étude concerne « le profil » des fumeurs de tabac. Pour cela on a créé un questionnaire concernant ***A*** « la quantité consommée par rapport à seuil donné » et ***B*** « le sexe ». Les résultats ont été les suivants :

	B_1 « hommes »	B_2 « femmes »	Total $n_{i\bullet}$
A_1 « consommer moins de 15 paquets par mois »	$n_{11} = 240$	$n_{12} = 222$	$n_{1\bullet} = 462$
A_2 « consommer 15 paquets ou plus par mois »	$n_{21} = 185$	$n_{22} = 115$	$n_{2\bullet} = 300$
Total $n_{\bullet j}$	$n_{\bullet 1} = 425$	$n_{\bullet 2} = 337$	$n_{\bullet\bullet} = 762$

Chacune des n personnes interrogées est positionnée dans l'une des $2 \times 2 = 4$ classes $A_i \times B_j$ où $i = 1,2$ et $j = 1,2$. Notons p_{ij} la probabilité associée à la classe $A_i \times B_j$, c'est-à-dire la fréquence relative étendue à toute la population $\boldsymbol{P}$ des adultes de la région concernée par l'étude. À partir des données statistiques de l'échantillon on souhaite savoir s'il existe un lien entre le sexe de l'individu et le fait d'avoir une consommation excessive de tabac. Sous l'hypothèse d'indépendance, on sait que $p_{ij} = p_{i\bullet} \times p_{j\bullet}\ \forall i, j$. Dans le cas contraire, si par exemple $p_{21} > p_{2\bullet} \times p_{\bullet 1}$, on peut considérer qu'il y a un lien concomitant entre le fait d'être un homme et de consommer de façon excessive du tabac.

1.1 Les paramètres du modèle log-linéaire

Les facteurs $\boldsymbol{A}$ et $\boldsymbol{B}$ étant respectivement partagés en $I = 2$ et $J = 2$ classes, on note p_{ij} la probabilité associée à la classe $A_i \times B_j$. Le modèle logistique consiste à décomposer les probabilités p_{ij} suivant un modèle multiplicatif ou de façon équivalente, de décomposer suivant un modèle additif les $\pi_{ij} = \ln(p_{ij})$. Dans le tableau ci-dessous figurent $\pi_{ij} = \ln(p_{ij})$, $\pi_{i\bullet} = (\sum_j \pi_{ij})/J$, $\pi_{\bullet j} = (\sum_i \pi_{ij})/I$ et $\pi_{\bullet\bullet} = (\sum_{ij} \pi_{ij})/I \times J$.

Tableau 17.1

	B_1	B_2	$\pi_{i\bullet}$
A_1	$\pi_{11} = \ln(p_{11})$	$\pi_{12} = \ln(p_{12})$	$\pi_{1\bullet} = (\pi_{11} + \pi_{12})/2$
A_2	$\pi_{21} = \ln(p_{21})$	$\pi_{22} = \ln(p_{22})$	$\pi_{2\bullet} = (\pi_{21} + \pi_{22})/2$
$\pi_{\bullet j}$	$\pi_{\bullet 1} = (\pi_{11} + \pi_{21})/2$	$\pi_{\bullet 2} = (\pi_{12} + \pi_{22})/2$	$\pi_{\bullet\bullet} = (\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22})/4$

Les quatre termes $\pi_{ij} = \ln(p_{ij})$ associés aux probabilités peuvent être décomposés de façon unique sous la forme $\pi_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$ si on impose les conditions $\sum_i \lambda_i^A = 0, \sum_j \lambda_j^B = 0, \sum_i \lambda_{ij}^{AB} = 0, \sum_j \lambda_{ij}^{AB} = 0$ où μ est l'effet global, λ_i^A l'effet dû au facteur A, λ_j^B l'effet dû au facteur B et λ_{ij}^{AB} l'effet dû à l'interaction entre les facteurs A et B. Autrement dit, on a les décompositions suivantes

Tableau 17.2

	B_1	B_2
A_1	$\pi_{11} = \mu + \lambda_1^A + \lambda_1^B + \lambda_{11}^{AB}$	$\pi_{12} = \mu + \lambda_1^A + \lambda_2^B + \lambda_{12}^{AB}$
A_2	$\pi_{21} = \mu + \lambda_2^A + \lambda_1^B + \lambda_{21}^{AB}$	$\pi_{22} = \mu + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB}$

En raison des contraintes, $\lambda_{12}^{AB} = -\lambda_{11}^{AB}$ et $\lambda_{22}^{AB} = -\lambda_{21}^{AB} = \lambda_{11}^{AB}$ et $\lambda_2^A = -\lambda_1^A$ et $\lambda_2^B = -\lambda_1^B$. Par résolution d'un système de neuf équations à neuf inconnues sont obtenues les valeurs des paramètres associés : $\mu = \pi_{\bullet\bullet}$, $\lambda_i^A = \pi_{i\bullet} - \pi_{\bullet\bullet}$, $\lambda_j^B = \pi_{\bullet j} - \pi_{\bullet\bullet}$, $\lambda_{ij}^{AB} = \pi_{ij} - \pi_{i\bullet} - \pi_{\bullet j} + \pi_{\bullet\bullet}$.

1.2 Les estimateurs des paramètres du modèle log-linéaire

Un modèle est dit hiérarchique si la présence dans le modèle du coefficient d'interaction entre variables implique la présence des coefficients de toutes interactions d'ordre inférieur. Ainsi, un modèle 2×2 où λ_{11}^{AB} est supposé non nul doit nécessairement intégrer les paramètres λ_1^A et λ_1^B. À partir des observations sur un échantillon de taille n on détermine, par la méthode du maximum de vraisemblance, les estimateurs respectifs $\hat{\lambda}_i^A$, $\hat{\lambda}_j^B$, $\hat{\lambda}_{ij}^{AB}$, $\hat{\mu}$ de λ_i^A, λ_j^B, λ_{ij}^{AB}, μ.

– **Le modèle saturé.** Si on admet l'existence d'interaction entre A_i et B_j, c'est-à-dire $\lambda_{ij}^{AB} \neq 0$, le modèle est dit ***saturé*** puisqu'il y a autant d'inconnues que d'équations. Les valeurs prises par ces estimateurs sont notées : λ_i^{A*}, λ_j^{B*}, λ_{ij}^{AB*} et μ^*. Désignant par n_{ij} l'effectif de la classe $A_i \times B_j$ et par $f_{ij} = n_{ij}/n$ sa fréquence relative, on obtient une décomposition multiplicative des fréquences relatives f_{ij} ou évidemment une décomposition additive en utilisant le logarithme $z_{ij} = \ln(f_{ij})$:

$$z_{ij} = \mu^* + \lambda_i^{A*} + \lambda_j^{B*} + \lambda_{ij}^{AB*}$$

où $\lambda_i^{A*} = z_{i\bullet} - z_{\bullet\bullet}$, $\lambda_j^{B*} = z_{\bullet j} - z_{\bullet\bullet}$, $\lambda_{ij}^{AB*} = z_{ij} - z_{i\bullet} - z_{\bullet j} + z_{\bullet\bullet}$, $\mu^* = z_{\bullet\bullet}$ avec $z_{i\bullet} = (\sum_j z_{ij})/J$, $z_{\bullet j} = (\sum_i z_{ij})/I$ et $z_{\bullet\bullet} = (\sum_{ij} z_{ij})/(I \times J)$.

Dans le cadre d'un modèle saturé, les probabilités attendues $p_{ij}^* = \mathrm{e}^{z_{ij}}$ sont évidemment égales aux fréquences relatives observées.

Exemple

Les données étant celles de l'exemple introductif, on calcule $z_{ij} = \ln(f_{ij}) = \ln(n_{ij}/n_{\bullet\bullet})$

	B_1	B_2	$z_{i\bullet}$
A_1	$z_{11} = \ln(240/762) = -1{,}155$	$z_{12} = \ln(222/762) = -1{,}233$	$z_{1\bullet} = -1{,}194$
A_2	$z_{21} = \ln(185/762) = -1{,}415$	$z_{22} = \ln(115/762) = -1{,}891$	$z_{2\bullet} = -1{,}653$
$z_{\bullet j}$	$z_{\bullet 1} = -1{,}285$	$z_{\bullet 2} = -1{,}562$	$z_{\bullet\bullet} = -1{,}424$

Dans le cas où l'on retient un modèle saturé, on a : $\lambda_i^{A*} = -1{,}285 - (-1{,}424) = 0{,}2295$;
$\lambda_1^{B*} = -1{,}285 - (-1{,}424) = 0{,}139$;
$\lambda_{11}^{AB*} = -1{,}155 - [-1{,}194 - 1{,}285] - 1{,}424 = -0{,}099$; $\mu^* = -1{,}423$.

Les estimateurs $\hat{\lambda}_i^A$, $\hat{\lambda}_j^B$, $\hat{\lambda}_{ij}^{AB}$ divisés par $\frac{1}{4} \times \sqrt{\sum_{ij} (n_{ij})^{-1}}$ suivent sensiblement la loi normale standard.

© Dunod. La photocopie non autorisée est un délit.

Remarque. Si une des valeurs λ_i^{A*}, λ_j^{B*}, λ_{ij}^{AB*} n'appartient pas à l'intervalle $(-2,2)$ il est raisonnable de penser que le modèle n'est pas saturé. On substitue au modèle initial un modèle où $\lambda_{11}^{AB} = 0$. Les logiciels donnent alors les valeurs prises par les nouveaux estimateurs déduits de la méthode du maximum de vraisemblance ainsi que le niveau de signification concernant ces valeurs.

– **Le modèle d'indépendance** est caractérisé par « $\lambda_{ij}^{AB} = 0$ » c'est-à-dire $\pi_{ij} = \mu + \lambda_i^A + \lambda_j^B \; \forall i, j$, ou de façon équivalente par $p_{ij} = p_{i\bullet} \times p_{\bullet j} \; \forall i, j$. À partir des observations de l'échantillon de taille n, on détermine, par la méthode du maximum de vraisemblance[1], les estimations $\mu^*, \lambda_1^{A*}, \lambda_1^{B*}$ des paramètres supposés non nuls :

$$\lambda_1^{A*} = (1/2)\ln(f_{1\bullet}/f_{2\bullet}), \lambda_1^{B*} = (1/2)\ln(f_{\bullet 1}/f_{\bullet 2}), \mu^* = (1/2)\ln(f_{1\bullet} \times f_{2\bullet} \times f_{3\bullet} \times f_{4\bullet})$$

où $f_{i\bullet} = \sum_j f_{ij}$ et $f_{\bullet j} = \sum_i f_{ij}$.

De ces valeurs sont déduites les fréquences attendues :

$$f_{11}^* = \exp[\mu^* + \lambda_1^{A*} + \lambda_1^{B*}], f_{12}^* = \exp[\mu^* + \lambda_1^{A*} + \lambda_2^{B*}] = \exp[\mu^* + \lambda_1^{A*} - \lambda_1^{B*}], \text{etc.}$$

Pour tester l'hypothèse d'indépendance H_0 « $\lambda_{ij}^{AB} = 0 \; \forall i, j$ », on utilise la propriété selon laquelle, sous l'hypothèse H_0 et si l'effectif de l'échantillon est suffisamment grand $(nf_{ij}^* > 3 \; \forall i, j)$, la *v.a.* Y^2 qui prend la valeur $2n \sum_{ij} f_{ij}[\ln(f_{ij}) - \ln(f_{ij}^*)]$ suit sensiblement la loi $\chi^2(1)$

Le nombre de degrés de liberté correspond au nombre de paramètres supposés nuls. Ce test est appelé test du rapport de vraisemblance.

Exemple

Les données sont celles de l'exemple introductif soit $f_{11} = 240/762 = 0{,}3150$, $f_{12} = 222/762 = 0{,}2913\ldots$

	B_1	B_2	Total $f_{i\bullet}$
A_1	$f_{11} = 0{,}315$	$f_{12} = 0{,}291$	$f_{1\bullet} = 0{,}606$
A_2	$f_{21} = 0{,}243$	$f_{22} = 0{,}151$	$f_{2\bullet} = 0{,}394$
Total $f_{\bullet j}$	$f_{\bullet 1} = 0{,}558$	$f_{\bullet 2} = 0{,}442$	$f_{\bullet\bullet} = 1$

1. Pour le modèle considéré « $\pi_{ij} = \mu + \lambda_i^A + \lambda_j^B$ » on a $p_{11} = \exp(\mu + \lambda_1^A + \lambda_1^B)$, $p_{12} = \exp(\mu + \lambda_1^A - \lambda_1^B)$, $p_{21} = \exp(\mu - \lambda_1^A + \lambda_1^B)$, $p_{22} = \exp(\mu - \lambda_1^A - \lambda_1^B)$. Selon la méthode du maximum de vraisemblance, on cherche le triplet de valeurs estimées $(\mu^*, \lambda_1^{A*}, \lambda_1^{B*})$ qui maximise $FV(\mu, \lambda_1^A, \lambda_1^B) = (n!/(n_{11}!n_{12}!n_{21}!n_{22}!) \times (p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}}))$ et donc $\mathcal{L}(\mu, \lambda_1^A, \lambda_1^B) = \ln(FV)$ sous la contrainte : $\sum_{ij} p_{ij} - 1 = 0$. $(\mu^*, \lambda_1^{A*}, \lambda_1^{B*}, \theta)$ est solution de $\frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial \lambda_1^A} = \frac{\partial L}{\partial \lambda_1^B} = \frac{\partial L}{\partial \theta} = 0$ où $L = \mathcal{L}(\mu, \lambda_1^A, \lambda_1^B) + \theta[\exp(\mu + \lambda_1^A \lambda_1^B) + \ldots + \exp(\mu - \lambda_1^A - \lambda_1^B) - 1]$.

Les valeurs estimées des inconnues principales μ, λ_1^A, λ_1^B par la méthode du maximum de vraisemblance sont : $\lambda_1^{A*} = (1/2)\ln(f_{1\bullet}/f_{2\bullet}) = (1/2)\ln(n_{1\bullet}/n_{2\bullet}) = 0{,}216$, $\lambda_1^{B*} = (1/2)\ln(f_{\bullet 1}/f_{\bullet 2}) = 0{,}116$ et $\mu^* = (1/2)\ln(f_{1\bullet} \times f_{2\bullet} \times f_{3\bullet} \times f_{4\bullet}) = -1{,}416$.

Sont déduites les fréquences relatives estimées : $f_{11}^* = \exp[\mu^* + \lambda_1^{A*} + \lambda_1^{B*}] = 0{,}338$; $f_{12}^* = \exp[\mu^* + \lambda_1^{A*} - \lambda_1^{B*}] = 0{,}268$; $f_{21}^* = \exp[\mu^* - \lambda_1^{A*} + \lambda_1^{B*}] = 0{,}22$; $f_{22}^* = 0{,}174$.

Les effectifs attendus ou estimés sous l'hypothèse d'absence d'interaction sont :

$n_{11}^* = 0{,}338 \times 762 = 257{,}7$, $n_{12}^* = 0{,}268 \times 762 = 204{,}3$, $n_{21}^* = 167{,}3$, $n_{22}^* = 132{,}7$.

Pour tester l'hypothèse « $\lambda_{11}^{AB} = 0$ », on détermine la valeur prise par Y^2 :

$$y^{*2} = 2 \times 762[0{,}3150 \times (\ln(0{,}3150) - \ln(0{,}338)) + \ldots + 0{,}151 \times (\ln(0{,}151) - \ln(0{,}174))] = 7.$$

Le risque de rejet à tort de H_0 correspond à $P_0(Y^2 \geqslant 7) \cong P(\chi_1^2 \geqslant 7) = 0{,}008$; ce niveau de risque étant infime, on rejette l'hypothèse de non interaction.

– **Le modèle hiérarchique avec équiprobabilité des catégories *B*** : « $\lambda_1^B = 0$ » (condition qui implique $\lambda_{11}^{AB} = 0$). Le modèle est du type : $\pi_{ij} = \mu + \lambda_i^A \; \forall i, j$.

Par la méthode du maximum de vraisemblance[1], on établit que : $\mu^* = (1/2)\ln(f_{1\bullet} \times f_{2\bullet}/4)$, $\lambda_1^{A*} = (1/2)\ln(f_{1\bullet}/f_{2\bullet})$ où $f_{i\bullet} = \sum_j f_{ij}$.

Les fréquences relatives estimées f_{ij}^* des classes $A_i \times B_j$ sont : $f_{11}^* = f_{12}^* = f_{1\bullet}/2$, $f_{21}^* = f_{22}^* = f_{2\bullet}/2$. En effet $f_{11}^* = \exp[\mu^* + \lambda_1^{A*}] = \exp[\ln(f_{1\bullet} - \ln(2)] = f_{1\bullet}/2$, etc.

1. **Justification.** Notons n_{ij} l'effectif observé de la classe $A_i \times B_j$. La répartition $(N_{11}, N_{12}, N_{21}, N_{22})$ d'un échantillon de taille n suit la loi 4-nomiale $\mathcal{M}_4(n; p_{11}, p_{12}, p_{21}, p_{22})$. La fonction de vraisemblance $FV(\mu, \lambda_1^A) = P(N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}) = K \prod_{ij} p_{ij}^{n_{ij}}$ où $K = n! / \prod_{ij} n_{ij}!$, où $p_{11} = \exp(\mu + \lambda_1^A)$, $p_{12} = \exp(\mu + \lambda_1^A)$, $p_{21} = \exp(\mu - \lambda_1^A)$, $p_{22} = \exp(\mu - \lambda_1^A)$. Le couple des valeurs estimées (μ^*, λ_1^{A*}) maximise $FV(\mu, \lambda_1^A)$ ou de façon équivalente $\mathcal{L}(\mu, \lambda_1^A) = \ln[FV(\mu, \lambda_1^A)]$ sous la contrainte $\sum_{ij} p_{ij} = 1$ c'est-à-dire $2\exp(\mu + \lambda_1^A) + 2\exp(\mu - \lambda_1^A) - 1 = 0$. En utilisant la méthode du multiplicateur de Lagrange et donc en considérant la fonction $\mathcal{L}(\mu, \lambda_1^A, \theta) = \mathcal{L}(\mu, \lambda_1^A) + \theta(2\exp(\mu + \lambda_1^A) + 2\exp(\mu - \lambda_1^A) - 1)$ on sait que $(\hat{\mu}, \hat{\lambda}_1^A, \theta)$ est solution du système: $\frac{\partial L}{\partial \mu} = 0$, $\frac{\partial L}{\partial \lambda_1^A} = 0$, $\frac{\partial L}{\partial \theta} = 0$. On en déduit immédiatement $\exp(\mu^* + \lambda_1^{A*}) = (f_{11} + f_{12})/2$, $\exp(\mu^* - \lambda_1^{A*}) = (f_{21} + f_{22})/2$, puis $\mu^* = 1/2\ln(f_{1\bullet} \times f_{2\bullet}/4)$, $\lambda_1^{A*} = (1/2)\ln(f_{1\bullet}/f_{2\bullet})$ et les fréquences estimées $f_{11}^* = f_{12}^* = \exp(\mu + \lambda_1^A) = f_{1\bullet}/2$, $f_{21}^* = f_{22}^* = \exp(\mu - \lambda_1^{*A}) = f_{2\bullet}/2$.

© Dunod. La photocopie non autorisée est un délit.

Propriété. Sous l'hypothèse H_0 « $\lambda_1^B = 0, \lambda_{11}^{AB} = 0$ » et lorsque $nf_{ij}^* > 3\ \forall i, j$, la *v.a.* $Y^2 = 2n \sum_{ij} f_{ij}[\ln(f_{ij}) - \ln(f_{ij}^*)]$ suit sensiblement la loi $\chi^2(2)$ (le nombre de degrés de liberté est 2 puisque 2 paramètres sont supposés nuls).

Remarque. Les modèles « $\lambda_1^A = 0$ » et « $\lambda_1^A = \lambda_1^B = 0$ » peuvent également être envisagés.

2 Modèle log-linéaire pour une table à $I \times J \times K$ classes

Exemple introductif

L'objet de l'étude concerne « le profil » des fumeurs de tabac. Pour cela on a crée un questionnaire concernant ***A*** « la quantité consommées par rapport à seuil donné », ***B*** « l'âge », ***C*** « le sexe . Les résultats sont les suivants

Hommes

	< 30 ans	entre 30 et 50 ans	$\geqslant 50$ ans	Total $n_{i\bullet 1}$
< 15 paquets par mois	$n_{111} = 83$	$n_{121} = 72$	$n_{131} = 85$	$n_{1\bullet 1} = 240$
$\geqslant 15$ paquets par mois	$n_{211} = 52$	$n_{221} = 68$	$n_{231} = 65$	$n_{2\bullet 1} = 185$
Total $n_{\bullet j 1}$	$n_{\bullet 11} = 135$	$n_{\bullet 21} = 140$	$n_{\bullet 31} = 150$	$n_{\bullet\bullet 1} = 425$

Femmes

	< 30 ans	entre 30 et 50 ans	$\geqslant 50$ ans	Total $n_{i\bullet 2}$
< 15 paquets par mois	$n_{112} = 77$	$n_{122} = 64$	$n_{132} = 81$	$n_{1\bullet 2} = 222$
$\geqslant 15$ paquets par mois	$n_{212} = 43$	$n_{222} = 48$	$n_{232} = 24$	$n_{2\bullet 2} = 115$
Total $n_{\bullet j 2}$	$n_{\bullet 12} = 120$	$n_{\bullet 22} = 112$	$n_{\bullet 32} = 105$	$n_{\bullet\bullet 2} = 337$

On est en présence d'un échantillon à 3 caractères ***A***, ***B*** et ***C*** qui se répartissent ainsi :

– A_1 « fumer moins de 15 paquets » , A_2 « fumer plus de 15 paquets » ;

– B_1 « être âgé de moins de 30 ans », B_2 « avoir entre 30 et 50 ans », B_3 « avoir plus de 50 ans » ;

– C_1 « être un homme », C_2 « être une femme » ;

Chacune des 762 personnes interrogées est positionnée dans l'une des $2 \times 3 \times 2 = 12$ classes $A_i \times B_j \times C_k$ où $i = 1,2$; $j = 1,2,3$; $k = 1,2$.

Plus généralement, ***A*** , ***B*** et ***C*** étant respectivement partagés en I, J et K classes, on notera n_{ijk} l'effectif (resp. f_{ijk} la fréquence relative) de la classe $A_i \times B_j \times C_k$.

– **Les paramètres du modèle.** Les probabilités p_{ijk} associées aux classes $A_i \times B_j \times C_k$, c'est-à-dire les fréquences relatives étendues à toute la population ***P*** vont être décomposées suivant un modèle multiplicatif. De façon équivalente, on peut décomposer sous la forme d'un modèle additif, les $I \times J \times K$ termes $\pi_{ijk} = \ln(p_{ijk})$ associés aux probabilités.

Dans le cas d'un ***modèle saturé*** où aucun paramètre n'est nul on a :

$$\pi_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

avec les conditions : $\sum_i \lambda_i^A = 0$, $\sum_j \lambda_j^B = 0$, $\sum_k \lambda_k^C = 0$, $\sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0$, $\sum_i \lambda_{ik}^{AC} = \sum_k \lambda_{ik}^{AC} = 0$, $\sum_j \lambda_{jk}^{BC} = \sum_k \lambda_{jk}^{BC} = 0$, $\sum_j \lambda_{ijk}^{ABC} = \sum_j \lambda_{ijk}^{ABC} = \sum_k \lambda_{ijk}^{ABC} = 0$.

– **Les estimateurs des paramètres du modèle saturé.** Partant du tableau des fréquences relatives f_{ijk} obtenues sur l'échantillon on réalise le même type de décomposition sur $z_{ijk} = \ln(f_{ijk})$.

Soit $z_{ij\bullet} = (\sum_k z_{ijk})/K$, $z_{i\bullet k} = (\sum_j z_{ijk})/J$, $z_{\bullet jk} = (\sum_i z_{ijk})/I$,
$z_{i\bullet\bullet} = (\sum_{jk} z_{ijk})/J \times K$, $z_{\bullet j\bullet} = (\sum_{ik} z_{ijk})/I \times K$, $z_{\bullet\bullet k} = (\sum_{ij} z_{ijk})/I \times J$ et
$z_{\bullet\bullet\bullet} = (\sum_{ijk} z_{ijk})/I \times J \times K$.

On a la décomposition additive :

$$z_{ijk} = \ln(f_{ijk}) = \mu^* + \lambda_i^{A*} + \lambda_j^{B*} + \lambda_k^{C*} + \lambda_{ij}^{AB*} + \lambda_{ik}^{AC*} + \lambda_{jk}^{BC*} + \lambda_{ijk}^{ABC*}$$

où $\lambda_i^{B*} = z_{i\bullet\bullet} - z_{\bullet\bullet\bullet}$, $\lambda_j^{B*} = z_{\bullet j\bullet} - z_{\bullet\bullet\bullet}$, $\lambda_k^{C*} = z_{\bullet\bullet k} - z_{\bullet\bullet\bullet}$,
$\lambda_{ij}^{AB*} = z_{ij\bullet} - z_{i\bullet\bullet} - z_{\bullet j\bullet} + z_{\bullet\bullet\bullet}$,
$\lambda_{ik}^{AC*} = z_{i\bullet k} - z_{i\bullet\bullet} - z_{\bullet\bullet k} + z_{\bullet\bullet\bullet}$, $\lambda_{jk}^{BC*} = z_{\bullet jk} - z_{\bullet j\bullet} - z_{\bullet\bullet k} + z_{\bullet\bullet\bullet}$,
$\lambda_{ijk}^{ABC*} = z_{ijk} - z_{ij\bullet} - z_{i\bullet k} - z_{\bullet jk} + z_{i\bullet\bullet} + z_{\bullet j\bullet} + z_{\bullet\bullet k} - z_{\bullet\bullet\bullet}$, $\mu^* = z_{\bullet\bullet\bullet}$.

Lorsque le modèle est saturé, les valeurs λ_i^{A*}, λ_j^{B*}, λ_k^{C*}, λ_{ij}^{AB*}, λ_{ik}^{AC*}, λ_{jk}^{BC*}, λ_{ijk}^{ABC*} et μ^* sont les valeurs estimées par la méthode du maximum de vraisemblance des paramètres correspondants.

Propriété. Chaque valeur estimée λ^* du paramètre λ est une combinaison linéaire des z_{ijk} : $\lambda^* = \sum_{ijk} a_{ijk} z_{ijk}$ où les a_{ijk} sont des constantes appropriées. Le quotient $\sqrt{n}\lambda^* / \sqrt{\sum_{ijk} (a_{ijk})^2 / f_{ijk}}$ est la valeur prise par une *v.a.* qui suit sensiblement la loi normale standard.

Un test d'un modèle non saturé. Si le modèle hiérarchique retenu est correct, la variable Y^2 qui prend la valeur $2n \sum_{ijk} f_{ijk}[\ln(f_{ijk}) - \ln(f_{ijk}^*)]$ suit sensiblement la loi $\chi^2(h)$ où h est le nombre de paramètres supposés nuls.

© Dunod. La photocopie non autorisée est un délit.

Section 2

LES MODÈLES LOGISTIQUES

Le résultat de la i-ème expérience e_i est de type binaire : succès ou échec, oui ou non, …, avec les probabilités respectives p_i, $q_i = 1 - p_i$. On associe à e_i, la *v.a.* Y_i qui prend la valeur 1 si succès ou 0 si échec : $P(Y_i = 1) = p_i$, $P(Y_i = 0) = 1 - p_i$.

On suppose que cette probabilité de succès p_i dépend de deux facteurs $\boldsymbol{A}$ et $\boldsymbol{B}$, ou des valeurs $x_{1,i}$ et $x_{2,i}$ prises par deux variables explicatives x_1 et x_2 : $p_i = \varphi(x_{1,i}, x_{2,i})$. Le modèle logistique permet d'estimer cette probabilité, autrement dit il estime la probabilité que $Y = 1$ sachant les valeurs prises par les variables explicatives : $P(Y = 1/(x_1, x_2) = (x_{1,i}, x_{2,i}))$.

Exemple

On considère un échantillon de 300 PME ayant opté ou non pour un dispositif fiscal (variable binaire Y) qui se répartissent selon x_1 leur effectif et l'âge x_2 du dirigeant.

Notons m_1 et m_2 les moyennes respectives des variables numériques exogènes x_1 et x_2 définies sur la population $\boldsymbol{P}$. Sous certaines conditions de régularité de la fonction φ et si les écarts $|x_{1,i} - m_1|$ et $|x_{2,i} - m_2|$ ne sont pas trop élevés on sait que $p_i \cong \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i}$ où α, β_1 et β_2 sont des constantes dont les valeurs sont inconnues. Cette approximation n'est plus valide lorsque $\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} < 0$ ou lorsque $\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} > 1$. Aussi afin de remédier à cet inconvénient majeur, on introduit une fonction $F : \mathbb{R} \mapsto]0,1[$ continue et strictement croissante et on considère la fonction $\Psi(x_{1,i}, x_{2,i}) = F^{-1}[\varphi(x_{1,i}, x_{2,i})]$.

Par la suite, on étudiera le modèle $p_i = F(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i})\ \forall\, i = 1,2,\ldots,n$ où les constantes α, β_1 et β_2 ont des valeurs inconnues.

Le modèle est dit *probit* lorsque $F(t) = \Phi(t)$.

Plus fréquemment, on utilise le modèle *logit* qui est associé à $F(t) = 1/(1 + \mathrm{e}^{-t})$ $\forall\, t \in \mathbb{R}$, fonction de répartition d'une loi logistique centrée. Dans le cas d'un modèle logit on a $\ln[p_i/(1 - p_i)] = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i}$, autrement dit

$$P(Y = 1/(x_1, x_2) = (x_{1,i}, x_{2,i})) = \exp[(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i})]/(1 + \exp[(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i})])$$

1 Modèles logit lorsque les variables exogènes sont continues

– Les paramètres du modèle

Soit un élément e_i d'une population statistique $\boldsymbol{P}$ dont les caractères $\boldsymbol{A}$ et $\boldsymbol{B}$ ont pour mesures respectives $x_1(e_i) = x_{1,i}$ et $x_2(e_i) = x_{2,i}$ et on s'intéresse à la probabilité p_i pour que e_i possède le caractère étudié C :

$$p_i = F(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i}) = (1 + \exp[-(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i})])^{-1},$$
$$q_i = 1 - p_i = (1 + \exp[(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i})])^{-1}.$$

On associe à e_i la variable aléatoire Y_i qui prend la valeur 1 si e_i possède le caractère C et la valeur 0 dans le cas contraire. Y_i suit la loi de Bernoulli de paramètre p_i. Les valeurs numériques α, β_1, β_2 (et donc p_i) sont inconnues.

– Les estimateurs par la méthode du maximum de vraisemblance

Disposant de n expériences e_i ($i = 1,2,\ldots,n$) réalisées sous les conditions exogènes $x_1(e_i) = x_{1,i}$, $x_2(e_i) = x_{2,i}$, on constate que $Y_i = 1$ si e_i possède le caractère C, 0 dans le cas contraire. Partant de ces n expériences, on définit les estimateurs $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$ des coefficients α, β_1, β_2 ainsi que les estimations numériques α^*, β_1^*, β_2^* de ces coefficients et p_i^* de p_i.

La fonction de vraisemblance $FV(\alpha,\beta_1,\beta_2) = P(Y_1 = y_i \text{ et } Y_2 = y_2 \text{ et } \ldots Y_n = y_n)$ où $y_i = 0$ ou 1. Donc

$$FV(\alpha,\beta_1,\beta_2) = p_1^{y_1}(1-p_1)^{(1-y_1)} \times p_2^{y_2}(1-p_2)^{(1-y_2)} \times \ldots p_n^{y_n}(1-p_n)^{(1-y_n)}$$

où $p_i = (1 + \exp[-(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i})])^{-1}$

FV maximum lorsque $\dfrac{\partial}{\partial\alpha}\ln[FV(\alpha,\beta_1,\beta_2)] = \dfrac{\partial}{\partial\beta_1}\ln FV = \dfrac{\partial}{\partial\beta_2}\ln FV = 0.$

La solution $(\alpha^*,\beta_1^*,\beta_2^*)$ de ce système de 3 équations à 3 inconnues est la valeur prise par le triplet d'estimateurs $(\hat{\alpha},\hat{\beta}_1,\hat{\beta}_2)$.

– Le test d'Hosmer-Lemeshow. Pour tester la validité du modèle, Hosmer et Lemeshow ont proposé un test du khi-deux de validité du modèle sur un tableau établi à partir des probabilités prédites $p_i^* = (1 + \exp[-(\alpha^* + \beta_1^* x_{1,i} + \beta_2^* x_{2,i})])^{-1}$. Les n nombres $p_1^*, p_2^*, \ldots, p_n^*$ étant classés par ordre croissant de valeurs puis répartis en k classes adjacentes C_h ($h = 1,\ldots,k$), on calcule l'effectif m_h de chaque classe C_h. Il convient de faire en sorte que les classes aient à peu près le même effectif, soit $m_h \cong n/k$ et $\sum_{h=1}^{k} m_h = n$.

Parmi les m_h expériences regroupées dans la classe C_h, il en existe r_h qui possèdent réellement le caractère C. La statistique HL_k d'Hosmer-Lemeshow est la statistique qui prend la valeur $hl_k^* = \sum_{h=1}^{k} \dfrac{(r_h - m_h\overline{p}_h)^2}{m_h\overline{p}_h(1-\overline{p}_h)}$ où $\overline{p}_h = (\sum_{j\in J_h} p_j^*)/m_h$ est la moyenne des probabilités prédites appartenant à la classe C_h.

Sous l'hypothèse H_0 de validité du modèle logit, c'est-à-dire « $p_i^* \cong p_i \ \forall\, i$ » et si $k \geqslant 10, m_h \geqslant 30$, la statistique HL_k suit sensiblement la loi $\chi^2(k)$. Le niveau de signification observé ou risque de rejet à tort de l'hypothèse H_0 est $\theta \cong P(\chi_k^2 \geqslant hl_k^*)$.

© Dunod. La photocopie non autorisée est un délit.

2 Modèle logit avec variables exogènes dichotomiques

2.1 Description du modèle avec deux variables exogènes

Les variables exogènes x_1 et x_2 sont dichotomiques et ne peuvent prendre que les valeurs 0 ou 1 ($x_1 = 0$ ou 1 et $x_2 = 0$ ou 1). Les résultats des n expérimentations peuvent être présentés à l'aide d'un tableau du type suivant où figurent les effectifs n_{ij} et n'_{ij}

Valeurs de $(x_1, x_2) \rightarrow$	(0,0)	(1,0)	(0,1)	(1,1)	Total
$Y = 1$	n_{00}	n_{10}	n_{01}	n_{11}	ν
$Y = 0$	n'_{00}	n'_{10}	n'_{01}	n'_{11}	ν'
Total	ν_{00}	ν_{10}	ν_{01}	ν_{11}	n

Lire : sur un total de ν_{00} expérimentations réalisées sous des conditions $x_1 = 0, x_2 = 0$ on a constaté que Y prenait n_{00} fois la valeur 1 et n'_{00} fois la valeur 0 ; etc.

Notons $p_{ij} = P(Y = 1|x_1 = i, x_2 = j)$ où i et j valent 0 ou 1. Ainsi $p_{00} = P(Y = 1|x_1 = 0, x_2 = 0)$, $p_{10} = P(Y = 1|x_1 = 1, x_2 = 0)$, etc.

On a donc pour la h-ème expérience $p(x_{1,h}, x_{2,h}) = p_{ij}$ où $i = x_{1,h}$ et $j = x_{2,h}$.

Le modèle logistique consiste à linéariser $\ln[p(x_{1,h}, x_{2,h})/(1 - p(x_{1,h}, x_{2,h}))] = \alpha + \beta_1 x_{1,h} + \beta_2 x_{2,h} + \gamma x_{1,h} x_{2,h}$ ou de façon équivalente $\ln[p_{ij}/(1 - p_{ij})] = \alpha + \beta_1 i + \beta_2 j + \gamma ij$.

– Les estimateurs dans le cas d'un modèle saturé

Dans le cas du modèle saturé, les meilleurs estimateurs des probabilités p^*_{ij} par la méthode du maximum de vraisemblance sont les proportions observées p^*_{ij} c'est-à-dire $p^*_{00} = n_{00}/\nu_{00}$, $p^*_{10} = n_{10}/\nu_{10}$, $p^*_{01} = n_{01}/\nu_{01}$, $p^*_{11} = n_{11}/\nu_{11}$, d'où l'on déduit les estimations α^*, β^*_1, β^*_2, γ^* de α, β_1, β_2, γ :

$\alpha^* = \ln(n_{00}/n'_{00})$, $\beta^*_1 = \ln(n_{10} \times n'_{00}/n_{00} \times n'_{10})$, $\beta^*_2 = \ln(n_{01} \times n'_{00}/n_{00} \times n'_{01})$, $\gamma^* = \ln(n_{11} \times n'_{00} \times n'_{10} \times n'_{01}/n'_{11} \times n_{00} \times n_{10} \times n_{01})$ [1]

Exemple

> On considère un échantillon de 700 PME ayant opté ou non pour un dispositif fiscal (variable Y) que l'on répartit selon x_1 leur effectif ($x_1 \leqslant 50$ ou $x_1 > 50$) et l'âge x_2 du dirigeant ($x_2 \leqslant 40$ ans ou $x_2 > 40$ ans). Y est une variable dichotomique qui prend la valeur 1 lorsque la PME a opté pour le dispositif et la valeur 0 dans le cas contraire. $x_1 = 0$ si l'effectif est inférieur à 50 et $x_1 = 1$ dans le cas contraire, $x_2 = 0$ si le dirigeant a moins de 40 ans et $x_2 = 1$ dans le cas contraire.

---

1. **Justification.** $FV(\alpha, \beta_1, \beta_2, \gamma) = \prod_{h=1}^{n} P(Y_h = y_h) = \prod_{h=1}^{n} p_{00}^{n_{00}} p_{10}^{n_{10}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}} (1 - p_{00})^{n'_{00}}$
$(1 - p_{10})^{n'_{10}} (1 - p_{01})^{n'_{01}} (1 - p_{11})^{n'_{11}}$ où $p_{ij} = (1 + \exp[-(\alpha + \beta_1 i + \beta_2 j + \gamma ij)])^{-1}$,
(ainsi, $p_{00} = (1 + \exp[-(\alpha)])^{-1}$, $p_{10} = (1 + \exp[-(\alpha + \beta_1)])^{-1}$, etc.

Valeurs de $(x_1, x_2) \rightarrow$	(0,0)	(1,0)	(0,1)	(1,1)	Total
$Y = 1$	$n_{00} = 5$	$n_{10} = 5$	$n_{01} = 5$	$n_{11} = 10$	$\nu = 25$
$Y = 0$	$n'_{00} = 348$	$n'_{10} = 232$	$n'_{01} = 48$	$n'_{11} = 47$	$\nu' = 675$
Total	$\nu_{00} = 353$	$\nu_{10} = 237$	$\nu_{01} = 53$	$\nu_{11} = 57$	$n = 700$

$p^*_{ij} = n_{ij}/\nu_{ij}$; $\alpha^* = \ln(5/348) = -4{,}2428$; $\beta^*_1 = \ln(5 \times 348/5 \times 232) = 0{,}405$; $\beta^*_2 = \ln(5 \times 348/5 \times 48) = 1{,}98$; $\gamma^* = \ln(n_{00} \times n'_{10} \times n'_{01} \times n_{11}/n'_{00} \times n_{10} \times n_{01} \times n'_{11}) = \ln(5 \times 232 \times 48 \times 10/348 \times 5 \times 5 \times 47) = 0{,}3087$.

Propriétés. Les estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\gamma}$ de β_1, β_2 et γ, élevés au carré et divisés par l'estimateur de leur variance respectif, suivent sensiblement la loi du khi-deux à 1 degré de liberté. La valeur prise par chacun des trois quotients est fournie par logiciel. θ désignant la plus grande de ces 3 valeurs, on évalue $P(\chi^2_1 \geqslant \theta)$ et on rejette l'hypothèse du modèle saturé (β_1, β_2 et γ non nuls) lorsque cette probabilité est trop faible, par exemple inférieure à 5 %.

Remarque. Le modèle saturé s'avère sans intérêt puisqu'il se borne à transcrire de façon bijective les données observées. Pour qu'un modèle soit utile, il doit être facilement interprétable. L'élaboration d'un tel modèle doit résulter d'un regroupement minimal de termes permettant de saisir les fondements sous-jacents du problème analysé, les regroupements éventuels de variables étant réalisés après une analyse en composantes principales. On doit ensuite réduire au minimum le nombre de paramètres non nuls du modèle.

– Étude du modèle sans interaction entre les variables exogènes

Dans le cas d'un modèle où γ le coefficient d'interaction entre les facteurs mesurés par x_1 et x_2 est supposé nul, la recherche des valeurs des estimateurs par la méthode du maximum de vraisemblance consiste à déterminer le triplet de valeurs $(\alpha^*, \beta^*_1, \beta^*_2)$ qui maximisent la fonction $L(\alpha, \beta_1, \beta_2) = \ln[FV]$ en déterminant la solution du système de 3 équations à 3 inconnues : $\dfrac{\partial L}{\partial \alpha} = 0$, $\dfrac{\partial L}{\partial \beta_1} = 0$, $\dfrac{\partial L}{\partial \beta_2} = 0$ où $p_{ij} = (1 + \exp[-(\alpha + \beta_1 i + \beta_2 j)])^{-1}$.

Pour un modèle de type logit « $p(x_{1,h}, x_{2,h}) = \alpha + \beta_1 x_{1,h}$ » on détermine le couple (α^*, β^*_1) solution de $\dfrac{\partial L}{\partial \alpha} = 0$, $\dfrac{\partial L}{\partial \beta_1} = 0$.

La méthode du maximum de vraisemblance consiste à rechercher le quadruplet de valeurs $(\alpha^*, \beta^*_1, \beta^*_2, \gamma^*)$ qui maximise la fonction de vraisemblance FV ou de façon équivalente maximise $L(\alpha, \beta_1, \beta_2, \gamma) = \ln(FV)$. Le quadruplet $(\alpha^*, \beta^*_1, \beta^*_2, \gamma^*)$ est solution du système : $\dfrac{\partial L}{\partial \alpha} = \dfrac{\partial L}{\partial \beta_1} = \dfrac{\partial L}{\partial \beta_2} = \dfrac{\partial L}{\partial \gamma} = 0$

© Dunod. La photocopie non autorisée est un délit.

2.2 Test du rapport de vraisemblance

Pour décrire le modèle avec le minimum de paramètres non nuls on considère le modèle saturé **[0]** $\ln[p_{ij}/(1-p_{ij})] = \alpha + \beta_1 \times i + \beta_2 \times j + \gamma \times ij$ où le nombre de paramètres non nuls h est égal à 4 puis successivement les modèles :

[1] $\alpha \neq 0, \beta_1 \neq 0, \beta_2 \neq 0, \gamma = 0$ où $h = 3$;

[2] $\alpha \neq 0, \beta_1 \neq 0, \beta_2 = 0, \gamma = 0$; ou $\alpha \neq 0, \beta_1 = 0, \beta_2 \neq 0, \gamma = 0$ où $h = 2$;

[3] $\alpha \neq 0, \beta_1 = 0, \beta_2 = 0, \gamma = 0$ où $h = 1$.

Pour chacun des modèles envisagés ci-dessus on a une estimation des valeurs des paramètres non nuls. Par exemple pour le modèle **[1]** on obtient un triplet $(\alpha^*, \beta_1^*, \beta_2^*)$ qui maximise $L(\alpha, \beta_1, \beta_2) = \ln[FV]$. La fonction de vraisemblance étant un produit de probabilité, $\ln[FV]$ est inférieur à 0. La valeur numérique $L_h^* = L_3^* = -2L(\alpha^*, \beta_1^*, \beta_2^*)$ est la valeur prise par une *v.a.* appelée ***statistique du log de vraisemblance***.

On définit ainsi successivement les statistiques L_4^* (associé au modèle saturé), L_3^*, L_2^*, $L_2'^*$, L_1^* où $L_h^* < L_k^*$ si le modèle donnant L_h^* a plus de paramètres non nuls que le modèle donnant L_k^*. Les tests fondés sur les rapports de vraisemblance permettent de comparer la qualité de l'ajustement réalisé par deux modèles distincts et permettent ainsi de déterminer quels sont les paramètres à intégrer pour améliorer significativement l'ajustement. Ces tests s'appuient sur les propriétés suivantes :

Propriétés.

i) si $\nu_{ij} > 4 \; \forall (i,j)$, la différence $D_h = L_k^* - L_4^* = 2\ln\left[\dfrac{FV \text{ (modèle saturé)}}{FV \text{ (modèle étudié)}}\right]$ suit sensiblement la loi khi-deux à $[2^r - (\textit{nombre de paramètres non nuls})]$, où r désigne le nombre de variables exogènes (dans l'exemple précédent $r = 2$ car l'étude concerne 2 variables exogènes dichotomiques x_1 et x_2),

ii) plus généralement, $\Delta_{h,k} = L_k^* - L_h^* \cong \chi^2_{h-k}$ où $(h-k)$ désigne la différence entre le nombre de paramètres non nuls du modèle [$\boldsymbol{h}$] et celui du modèle [$\boldsymbol{k}$].

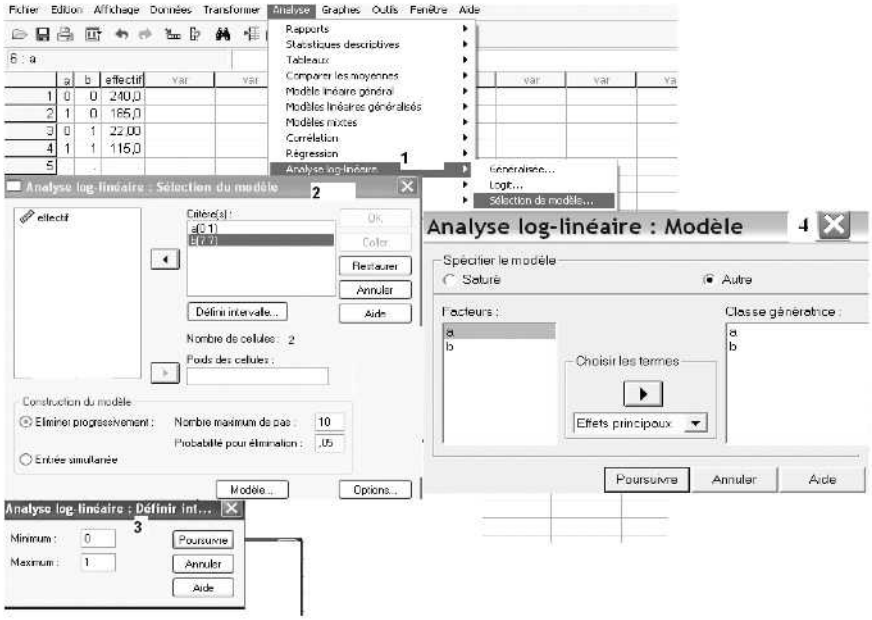
Section **3**

PROCÉDURES DE TRAITEMENT SOUS SPSS

1 Modèle log-linéaire

Reprenons l'exemple introductif (p. 291).

Partant de ce tableau on élabore sous SPSS la feuille de travail. Aux facteurs A et B sont associées les variables a et b. La variable *effectif* sert à pondérer les différentes classes par les effectifs.



Procédure.

Pour la pondération

Pour pondérer les observations il faut pointer sur (Données) puis (Pondérer les observations) et dans le menu pondération il faut retenir pour variable de pondération la variable *effectif*.

Pour obtenir le modèle saturé

1. Une fois qu'on a entré les données, on pointe sur (Analyse) puis (Analyse Log-linéaire) et (Sélection de modèle).

2. Dans le menu **analyse log-linéaire**, on sélectionne les facteurs *a* et *b* en tant que (critères).

3. On précise les modalités des facteurs (valeur 0 et 1) en cliquant sur (intervalle) et en indiquant que (minimum) = 0 et (maximum) = 1.

4. Cliquant sur (Modèle) on sélectionne le modèle sans interaction en cliquant sur (autre). On retient les paramètres *a* et *b* sans interaction (l'interaction correspond à a^*b).

Résultats

HIERARCHICAL LOG LINEAR				
Factor	**code**	**OBS count**	**EXP count**	**Residual**
A	0			
B	1	240	257,7	– 17,7
B	0	222	204,3	17,7
A	1			
B	0	185	167,3	17,7
B	1	115	132,7	17,7

© Dunod. La photocopie non autorisée est un délit.

Goodness-of-fit test statistics

Likelihood ratio chi square = (7,000) DF = 1 P = (0,008) Pearson chi square = 6,96479 DF = 1 P = 0,008

Commentaires. Les effectifs attendus sous l'hypothèse d'absence d'interaction (EXP count) sont $n^*_{11} = 257{,}7$, $n^*_{12} = 204{,}3$, $n^*_{21} = 167{,}3$, $n^*_{22} = 132{,}7$. Ils différent évidemment des effectifs observés (OBS count) puisque le modèle n'est pas saturé.

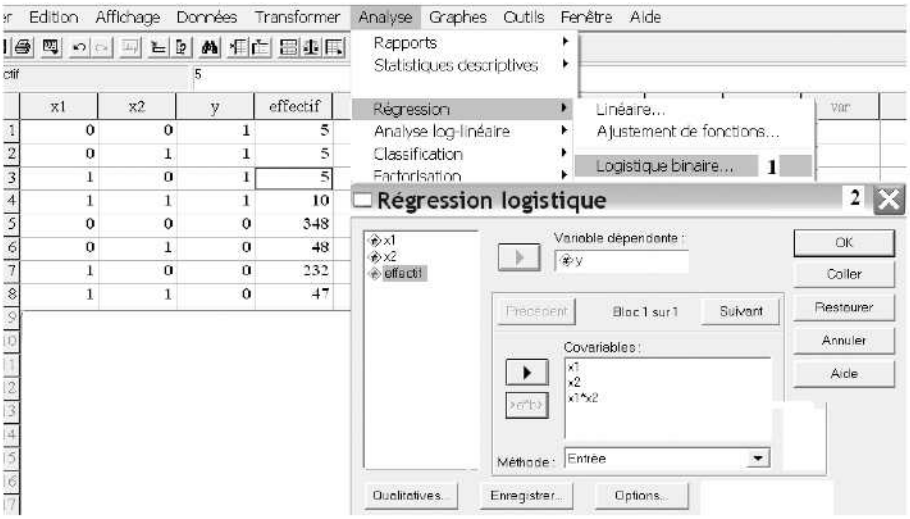
Afin de tester l'hypothèse d'indépendance ou d'absence d'interaction entre la consommation excessive A et le sexe B , on teste « $\lambda^{AB}_{11} = \lambda^{AB}_{12} = \lambda^{AB}_{21} = \lambda^{AB}_{22} = 0$ » puis on détermine la valeur prise par Y^2 appelée ici *Likelihood ratio chi square* qui sous H_0 suit une loi du Khi-deux à un degré de liberté (DF = 1 soit un paramètre nul) : $y^{*2} = 2 \times [240 \times (\ln(240) - \ln(257{,}7)) + \ldots + 115 \times (\ln(115) - \ln(132{,}7))] = (7)$. Le risque de rejet à tort de l'hypothèse H_0 de non interaction est égal à $P_0(Y^2 \geqslant (7)) \cong P(\chi^2_1 \geqslant 7) = (0{,}008)$ et peut être considéré comme négligeable.

2 Modèle Logit

Reprenons le second exemple du cours. On considère un échantillon de 700 PME ayant opté ou non pour un dispositif fiscal (variable Y) qui se répartissent selon x_1 leur effectif (inférieur ou égal à 50 ou supérieur à 50) et l'âge x_2 du dirigeant (inférieur ou égal à 40 ans ou supérieur à 40 ans) (*cf.* p. 300).

2.1 Modèle saturé

Partant de ce tableau, on élabore sous SPSS la feuille suivante en entrant les variables y, x_1 et x_2 en colonne. La variable *effectif* sert à pondérer les différentes classes par les effectifs (5, 5, 5, 10, 348, …) (*cf.* p. 303 pour voir la procédure servant à pondérer les variables).



Procédure.

1. Une fois rentrées les données, on pointe sur (Analyse), (Régression) puis (Logistique binaire).

2. Voulant un modèle saturé, on sélectionne dans le menu régression logistique ci-dessous, y en tant que (variable dépendante) et les variables x_1, x_2 et leur interaction $x_1^* x_2$ (pour avoir l'interaction sélectionner x_1, x_2 et cliquer sur >a*b> en tant que facteurs nommés ici (covariables).

Résultats

Variables dans l'équation		B	E.S.
Étape 1	x_1 by x_2	0,309	0,866
	x_1	0,405	0,638
	x_2	1,981	0,651
	Constante	– 4,243	0,450

Commentaire. Le logiciel permet d'obtenir les coefficients du modèle saturé : $\text{logit}\, p(x_{1,h},x_{2,h}) = \ln[p(x_{1,h},x_{2,h})/(1 - p(x_{1,h},x_{2,h}))] = \alpha + \beta_1 x_{1,h} + \beta_2 x_{2,h} + \gamma x_{1,h} x_{2,h}$. On lit les coefficients $\alpha^* = -4{,}2428$, $\beta_1^* = 0{,}405$, $\beta_2^* = 1{,}98$, $\gamma^* = 0{,}309$. Les autres éléments non figurés ici n'ont pas à être commentés puisque le modèle est saturé. (x_1 by $x_2 =$ interaction)

2.2 Modèle avec le minimum de paramètres non nuls

Pour décrire le modèle avec le minimum de paramètres non nuls, on utilise pour critère le niveau de signification associé au test de Wald (rapport de vraisemblance).

Procédure.

1. *cf.* ci-desssus. **2.** On sélectionne Descendante Wald pour méthode.

3. On clique sur option et on sélectionne les test de Hosmer-Lemeshow.

Résultats

Tableau 17.3 – *Bloc 0 : bloc de départ*

Variables dans l'équation							
		B	E.S.	Wald	ddl	Signif.	Exp(B)
Étape 0	Constante	– 3,296	0,204	261,867	1	0,000	0,037

Tableau 17.4 – *Bloc 1 : Statistique-2log-vraisemblance*

Étape	
1	187,043
2	187,170
3	189,008

Tableau 17.5 – *Tests de spécification du modèle*

		Khi-deux	ddl	Signif.
Étape 1	Étape	28,66	3	2,6348E–06
	Bloc	28,66	3	2,6348E–06
	Modèle	28,66	3	2,6348E–06
Étape 2	Étape	– 0,13	1	0,72072809
	Bloc	28,54	2	6,3602E–07
	Modèle	28,54	2	6,3602E–07
Étape 3	Étape	– 1,84	1	0,17525655
	Bloc	26,70	1	2,3778E–07
	Modèle	26,70	1	2,3778E–07

© Dunod. La photocopie non autorisée est un délit.

Une valeur du khi-deux négative indique que la valeur du khi-deux a diminué depuis l'étape précédente.

Tableau 17.6 – *Variables dans l'équation*

		B	E.S.	Wald	ddl	Signif.	Exp(B)	IC pour Exp(B) 95,0 %	
								Inférieur	Supérieur
Étape 1	x_1	0,405	0,638	0,404	1	0,525	1,500	0,429	5,239
	x_2	1,981	0,651	9,262	1	0,002	7,250	2,024	25,966
	x_1 by x_2	0,309	0,866	0,127	1	0,721	1,362	0,250	7,429
	Constante	– 4,243	0,450	88,731	1	4,5226E–21	0,014		
Étape 2	x_1	0,575	0,428	1,803	1	0,17933213	1,777	0,768	4,110
	x_2	2,157	0,425	25,706	1	3,9766E–07	8,646	3,756	19,906
	Constante	–4,331	0,394	120,931	1	3,9574E–28	0,013		
Étape 3	x_2	2,215	0,423	27,412	1	1,6439E–07	9,158	3,997	20,982
	Constante	–4,060	0,319	162,078	1	3,9782E–37	0,017		

Variable(s) entrées à l'étape 1: x_1, x_2, $x_1^* x_2$.

Commentaire

Tableau 17.3. Le modèle de référence est $\ln[p_{ij}/(1-p_{ij})] = \alpha$ puisque dans le menu option on a retenu le modèle avec une constante.

Tableau 17.4. Il présente la valeur prise par la variable aléatoire $L_h^* = -2 \times \ln[FV]$ où FV est la valeur maximale de la fonction de vraisemblance pour le modèle étudié comportant h paramètres non nuls.

Tableau 17.5. Le logiciel calcule pour chaque étape (1, 2 et 3) la statistique $\Delta_{h,k} = L_k^* - L_h^* \cong \chi^2_{h-k}$ où $(h-k)$ désigne la différence entre le nombre de paramètres non nuls du modèle $[h]$ et celui du modèle de l'étape 0, modèle de référence avec la constante $\alpha \neq 0$ soit $k = 1$ nombre de paramètres non nuls.

Dans le tableau 17.6 figurent les variables restant dans le modèle à chaque étape et dans **le tableau 17.7** figurent réciproquement les variables exclues.

Commentaires par étape

Étape 0. Dans l'étape nommée Bloc 0, on teste $\ln[p_{ij}/(1-p_{ij})] = \alpha + \beta_1 \times i + \beta_2 \times j + \gamma \times ij$ avec les conditions [1] $\alpha \neq 0$, $\beta_1 = 0$, $\beta_2 = 0$, $\gamma = 0$ soit un nombre h de paramètres non nuls égal à 1 et donc $\ln[p_{ij}/(1-p_{ij})] = \alpha$. On trouve une estimation de α égale – 3,296 par la méthode du maximum de vraisemblance. La statistique de Wald prend pour valeur le carré du rapport entre la valeur estimée du paramètre et son écart-type soit $(-3{,}296/0{,}204)^2 = 261{,}867$ et suit une loi du Khi deux à 1 degré de liberté (h nombre de paramètre non nul est égal à 1).

Dans les étapes nommées Bloc 1, le logiciel part du modèle saturé et utilisant pour critère la *statistique du log de vraisemblance* il fait sortir, à chaque étape, les variables qui n'améliorent pas significativement la qualité de l'ajustement. Il commence évidemment par celle qui contribue le moins à la qualité de l'ajustement.

Étape 1. Le logiciel part tout d'abord du modèle saturé (*cf.* tableau 17.6 étape 1) $\ln[p_{ij}/(1-p_{ij})] = \alpha + \beta_1 \times i + \beta_2 \times j + \gamma \times ij$ où $i = x_{1,h}$ et $j = x_{2,h}$. et où le

nombre h de paramètres non nuls est 4 puisque $\alpha \neq 0$, $\beta_1 \neq 0$, $\beta_2 \neq 0$ et $\gamma \neq 0$. Il calcule la valeur de la fonction de vraisemblance FV correspondante (*cf.* tableau 17.4 étape 1), plus exactement la statistique $L_4^* = -2 \times \ln(FV)$ qui prend pour valeur 187,043. Dans le tableau 17.5 on compare la qualité du modèle de l'étape 1 à $h = 4$ paramètres non nuls au modèle de l'étape 0 à un paramètre non nul incluant seulement la constante $\alpha \neq 0$ [nombre de paramètre non nul $k = 1$]. Ainsi la statistique du Log de vraisemblance $\Delta_{h,k} = L_k^* - L_h^* = L_1^* - L_4^* \cong \chi^2_{h-k}$ $= \chi^2_3$ prend la valeur 28,66 et, testant l'hypothèse H_0 « le modèle de l'étape 1 n'est pas meilleur que le modèle de l'étape 0 » c'est-à-dire que « $\beta_1 = \beta_2 = \gamma = 0$ », on constate que le niveau de signification est égal à $P(\chi^2_3 \geqslant 23{,}66) = 2{,}63486 \times 10^{-6}$ et on rejette H_0. On détermine dans le tableau 16.6 si les coefficients sont significativement différents de 0. Les estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\gamma}$ et $\hat{\alpha}$ prennent pour valeur respective 0,405, 1,981, 0,309, – 4,24. Élevés au carré puis divisés par leur variance ils suivent sensiblement la loi du khi-deux à 1 degré de liberté. Testant l'hypothèse H_0 « $\beta_1 = 0$ » contre $\overline{H}_0$, le risque de rejet à tort de H_0 est $\cong P(\chi^2_1 \geqslant (0{,}405/0{,}638)^2) = P(\chi^2_1 \geqslant 0{,}404) = 0{,}525$ et on conclut au non rejet de H_0.

Étape 2. Le logiciel sort la variable la moins significative, soit ici la variable d'interaction ou paramètre γ et considère le modèle $\ln[p_{ij}/(1 - p_{ij})] = \alpha + \beta_1 i + \beta_2 j$ où il y a trois paramètres non nuls ($h = 3$) soit $\alpha \neq 0$, $\beta_1 \neq 0$, $\beta_2 \neq 0$ et $\gamma = 0$. Dans le tableau 17.4, la statistique $L^*_{3'} = -2 \times \ln(FV) = 187{,}17$. Dans le tableau 17.5, utilisant la statistique $\Delta_{h,k} = L_h^* - L_k^* \cong \chi^2_{h-k}$, on calcule la différence de la statistique du Log de vraisemblance entre le modèle à l'étape 2 (à 3 variables non nulles) et le modèle à l'étape 1 (à 4 variables non nulles). Ainsi $\Delta_{4,3} = L_3^* - L_4^* =$ $187{,}17 - 187{,}043 \cong 0{,}13$. Cette différence n'est pas significativement différente de 0. L'inclusion ou non de l'effet d'interaction (du paramètre γ) n' a pas d'effet significatif sur la qualité de l'ajustement (ce résultat peut être également constaté dans le tableau 17.7 dans la rubrique statistiques globales). Remarquons que la comparaison du modèle de l'étape 2 au modèle de l'étape 0 donne la statistique $\Delta_{3,1} = L_1^* - L_3^*$ qui prend la valeur 28,54. Aussi testant l'hypothèse H_0 que « $\beta_1 = \beta_2 = 0$ » on constate que le niveau de signification est égal à $P(\chi^2_2 \geqslant 28{,}54) = 6{,}3602 \times 10^{-7}$ et on rejette donc H_0. Figurent dans le tableau 17.7, comme pour l'étape 1, les tests de signification des coefficients. Testant l'hypothèse H_0 que « $\beta_2 = 0$ » on constate que le niveau de signification est égal à $P(\chi^2_1 \geqslant 25{,}706) = 3{,}9766 \times 10^{-7}$ on rejette donc l'hypothèse H_0 de non influence de x_2.

Dans l'étape 3 on exclut la variable x_1 et on teste $\ln[p_{ij}/(1 - p_{ij})] = \alpha + \beta_2 j$ un modèle avec deux paramètres non nuls $h = 2$. Dans le tableau 17.4 la statistique $L^*_{2'} = -2 \times \ln(FV) = 189{,}08$. On calcule les valeurs prises par $\Delta_{3,2} = L_2^* - L_3^*$ prend la valeur $189{,}08 - 187{,}17 = 1{,}84$ et $P(\chi^2_{3-2} > 1{,}84) = 0{,}175$, aussi le modèle incluant le paramètre β_1 n'est pas significativement meilleur. L'inclusion de

© Dunod. La photocopie non autorisée est un délit.

l'effet de x_1 n'a pas d'effet significatif sur la qualité de l'ajustement. On observe dans le tableau 17.6 qu'il reste la constante et la variable x_2, les coefficients estimés sont $\beta_2^* = (2{,}215)$ et $\alpha^* = (-4{,}06)$. Testant l'hypothèse H_0 « $\beta_2 = 0$ » *contre* $\overline{H}_0$, le risque de rejet à tort de H_0 est $\cong P_0(\chi_1^2 \geqslant 27{,}412) = 1{,}6439 \times 10^{-7}$. Seule l'entrée de la variable x_2 améliore significativement l'ajustement, autrement dit seule l'entrée du facteur âge du dirigeant améliore significativement la qualité de l'ajustement.

Selon le test associé à la statistique du *log de vraisemblance* la prise en compte de x_1 et de l'interaction $x_1^* x_2$ avec les estimations de leur paramètre β_1 et γ_{12} n'améliore pas le niveau de signification associé à la *statistique du log de vraisemblance*.

Le modèle estimé a donc pour expression : $\ln[p(x_{1,h},x_{2,h})/(1 - p(x_{1,h},x_{2,h}))] = -4{,}06 + 2{,}215x_{2,h}$ soit

$$p(x_{1,h},x_{2,h}) = F(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i}) = (1 + \exp[-(-4{,}06 + 2{,}215x_{2,h})])^{-1}$$

La probabilité d'adoption du dispositif fiscal s'accroît avec l'âge du dirigeant.

Résultat du test de Hosmer-Lemeshow

Test de Hosmer-Lemeshow			
Étape	**Khi-deux**	**ddl**	**Signif.**
1	1,88031E–09	2	1
2	0,127	2	0,938

Tableau de contingence pour le test de Hosmer-Lemeshow

		Y = ,00		Y = 1,00		Total
		Observé	Théorique	Observé	Théorique	
Étape 1	1	348	348,000	5	5,000	353
	2	232	232,000	5	5,000	237
	3	48	48,000	5	5,000	53
	4	47	47,000	10	10,000	57
Étape 2	1	348	348,415	5	4,585	353
	2	232	231,585	5	5,415	237
	3	48	47,585	5	5,415	53
	4	47	47,415	10	9,585	57

Commentaires. Pour tester la validité du modèle, Hosmer et Lemeshow ont proposé un test du khi-deux sur un tableau établi à partir des probabilités prédites $p_i^* = (1 + \exp[-(\alpha^* + \beta_1^* x_{1,i} + \beta_2^* x_{2,i})])^{-1}$ à l'aide des modèles des différentes étapes. Pour chaque modèle, sous l'hypothèse H_0 de validité du modèle logit c'est-à-dire « $p_i^* \cong p_i \; \forall \, i$ » (probabilités prédites égales aux probabilité réelles) ces probabilités sont réparties entre deux classes ($k = 2$) et la statistique HL_k suit ici sensiblement la loi $\chi^2(2)$. La valeur du niveau de signification observé ou risque de rejet à tort de l'hypothèse H_0 est ainsi à l'étape 2 égale à $\theta \cong P(\chi_2^2 \geqslant 0{,}127) = 0{,}938$ aussi peut-on accepter H_0.

Exercices (*cf.* corrections page 365)

Exercice 1

L'objet de l'étude concerne le lien pouvant exister entre les rémunérations annuelles des cadres et leur profil de formation dans un secteur donné. Chacune des 400 personnes interrogées est positionnée dans l'une des $2 \times 2 = 4$ classes $A_i \times B_j$ où $i = 1,2$ et $j = 1,2$

	B_1 « gagner – de 40 K-euros »	B_1 « gagner + de 40 K-euros »	Total $n_{i\bullet}$
A_1 « avoir une formation diplômante inférieure à une troisième cycle »	$n_{11} = 50$	$n_{12} = 100$	$n_{1\bullet} = 150$
A_2 « avoir un diplôme de troisième cycle ou d'école d'ingénieur »	$n_{21} = 200$	$n_{22} = 50$	$n_{2\bullet} = 250$
Total $n_{\bullet j}$	$n_{\bullet 1} = 250$	$n_{\bullet 2} = 150$	$n_{\bullet\bullet} = 400$

Utilisant l'analyse log-linéaire testez l'hypothèse d'absence d'interaction entre A le niveau de formation et B le niveau de rémunération.

Exercice 2

Soit une population $\boldsymbol{P} = \{e_1, e_2, \ldots, e_N\}$. À chaque élément e_i est associé la mesure numérique $x(e_i) = x_i$ d'un caractère A $(0 \leqslant x(e_i) \leqslant 0,9) \quad \forall\, i$. Cet élément e_i possède ou ne possède pas le caractère étudié C avec les probabilités respectives $p_i = \varphi(x_i)$ et $q_i = 1 - p_i$ et l'on considère la *v.a.* $Y_i = 1$ si e_i possède le caractère C, $Y_i = 0$ dans le cas contraire. On dispose d'un échantillon de taille 8 où les valeurs x_i ont été ordonnées de façon croissante :

Valeurs x_i	0,24	0,31	0,35	0,42	0,52	0,63	0,68	0,83
Valeurs y_i	0	1	0	0	1	0	1	1

À partir de ces données on émet l'hypothèse H_0 selon laquelle $p_i = ax_i$ où a est une constante de valeur inconnue.

Montrer que $a^* = 1,0892$ est la valeur estimée de a par la méthode du maximum de vraisemblance.

© Dunod. La photocopie non autorisée est un délit.

18 ACP & AFC

L'analyse en composante principale (ACP) consiste à résumer l'information contenue dans un tableau de deux entrées *individus* × *variables* (généralement individus en lignes, variables en colonne) en remplaçant les variables initiales par un plus petit nombre de nouvelles variables. L'analyse factorielle des correspondances (AFC) permet de visualiser les relations pouvant exister entre les modalités de deux caractères (par exemple la couleur des yeux et des cheveux).

Section 1 ■ Analyse en composante principale

Section 2 ■ Analyse factorielle des correspondances

Section 1

ANALYSE EN COMPOSANTES PRINCIPALES

Partant d'un tableau à deux entrées *individus* × *variables* appelé matrice des données, l'analyse en composantes principales permet de visualiser les corrélations entre les différentes variables associées aux caractères étudiés. Elle sert également à repérer des groupes d'individus ayant un comportement semblable vis-à-vis des caractères étudiés.

1 La Matrice des données

Exemple introductif

Une enquête a été réalisée auprès de 10 banques allemandes $e_1, e_2, \ldots, e_{10}$ afin de dégager les relations pouvant exister entre différents indicateurs de rentabilité (« ROE = *Bénéfice*/(*Fonds propres*) », « ROA = *Bénéfice*/(*Total de l'actif*) »), de productivité (mesuré par le coefficient d'exploitation CExp) et de taille (Effectif). Le tableau présenté ci-dessous récapitule les données concernant les 10 banques et les quatre indicateurs retenus. On souhaite visualiser les corrélations et déterminer s'il existe des groupes de banques ayant des caractéristiques analogues.

Nom des banques	Obs	x_1 Effectifs	x_2 CExp	x_3 ROE	x_4 ROA	x'_1	x'_2	x'_3	x'_4
DEUTSCHE BANK AG	A	76 141	0,81	3,22	0,11	1,86	0,74	– 1,72	– 1,30
DRESDNER BANK AG	B	46 000	0,77	11,31	0,3	0,60	0,35	1,21	1,30
WEST.LAND. GIR..	C	10 116	0,89	5,58	0,15	– 0,89	1,51	– 0,86	– 0,75
COMMERZBANK AG	D	29 615	0,66	10,26	0,3	– 0,08	– 0,70	0,83	1,30
BAYERIS.VEREINSB.	E	22 001	0,7	8,49	0,21	– 0,40	– 0,32	0,19	0,07
BAY. LAND.GIR.	F	5 192	0,57	8,89	0,16	– 1,10	– 1,57	0,34	– 0,61
Moyenne $\overline{x}_j$		31 510,833	0,733	7,958	0,205	0	0	0	0
Écart– type σ_j		23 966,083	0,104	2,763	0,073	1	1	1	1

Le point moyen ou centre de gravité du nuage de points est le point g = (31 510,833, 0,733, 7,958, 0,205). À chaque valeur x_{ij} on associe sa valeur centrée réduite $x'_{ij} = (x_{ij} - \overline{x}_j)/\sigma_j$ où $\overline{x}_i$ et σ_i sont la valeur moyenne et l'écart-type du caractère i.
$x'_{11} = (x_{11} - \overline{x}_1)/\sigma_1 = (76\,141 - 31\,510{,}833)/23\,966{,}083 = 1{,}86$, $x'_{21} = 0{,}6, \ldots,$

Partant d'un tableau à h variables et n individus, on associe à chaque élément e_i d'une population $\boldsymbol{P}$ le h-uple de variables réelles $x(e_i) = (x_{i1}, x_{i,2}, \ldots, x_{ih})$. La matrice de données associée M est à n lignes et h colonnes :

$$M = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1h} \\ x_{21} & x_{22} & \ldots & x_{2h} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nh} \end{bmatrix}$$

Dans la i-ème ligne figurent les valeurs numériques des h caractères (variables) de l'individu e_i

À chaque élément e_i on attribue un poids p_i tel que $p_i > 0$ et $\sum_{i=1}^{p} p_i = 1$, généralement $p_i = 1/n$[1]. Afin d'obtenir une analyse centrée et objectivement indépen-

1. Si par exemple l'enquête concerne n magasins d'une chaîne, p_i peut représenter la part relative du chiffre d'affaires $K(e_i)$ réalisé par le magasin e_i : $p_i = K(e_i)/\sum_{j=1}^{n} K(e_j)$.

© Dunod. La photocopie non autorisée est un délit.

dante de l'unité de mesure choisie pour chaque caractère étudié C_j, on substitue aux observations x_{ij} leur valeur centrée réduite : $x'_{ij} = (x_{ij} - \overline{x}_j)/\sigma_j$ où $\overline{x}_j = \sum_{i=1}^{n} p_i x_{ij}$ et $\sigma_j^2 = \sum_{i=1}^{n} p_i (x_{ij} - \overline{x}_j)^2$ sont la moyenne et la variance de la variable j.

La moyenne de ces observations centrées réduites est nulle, $\sum_{i=1}^{n} p_i x'_{ij} = 0$ et $\sum_{i=1}^{n} p_i (x'_{ij})^2 = 1 \ \forall\, j = 1,2,\ldots,h$ et donc leur centre de gravité $g' = (0,0,\ldots,0)$.
Par la suite on suppose que pour chaque caractère, les données sont préalablement centrées réduites et on utilise la matrice $M' = (x'_{ij})$ où $1 \leqslant i \leqslant n$ et $1 \leqslant j \leqslant h$.

2 Nuage de points dans l'espace affine R^h des observations

Chaque individu e_i est représenté par le point P_i de coordonnées $(x'_{i1}, x'_{i2}, \ldots, x'_{ih})$ dans l'espace affine R^h des observations. Les n points P_i affectés des poids p_i constituent un nuage N de points dont le centre gravité est au point $O = (0,0,\ldots,0)$.

Soient deux points quelconques $P = (x'_1,\ldots,x'_h)$ et $Q = (y'_1,\ldots,y'_h)$ de l'espace affine $\mathbb{R}^h$ des variables, la longueur du segment PQ c'est-à-dire la distance du point P au point Q sera notée $|PQ| : |PQ|^2 = \sum_{j=1}^{h} (y'_j - x'_j)^2$ (Pythagore)

– Imaginons que les n points du nuage se situent à proximité d'une droite passant par le centre de gravité O puisque les variables sont centrées. La meilleure droite retenue Δ sera celle pour laquelle la somme pondérée des carrés des distances de chaque point P_i à cette droite sera minimum.

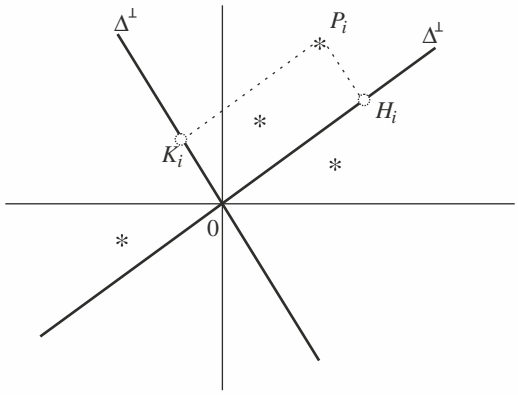


Figure 18.1

H_i est la projection orthogonale de P_i sur Δ, sa coordonnée sur cet axe est Ψ_i

K_i est la projection de P_i sur l'hyperplan $\Delta^{\perp}$ orthogonal à la droite Δ.

On a $|OP_i|^2 = |OH_i|^2 + |H_iP_i|^2$ et donc pour le nuage de points $\sum_{i=1}^{n} p_i|OP_i|^2 = \sum_{i=1}^{n} p_i|OH_i|^2 + \sum_{i=1}^{n} p_i|P_iH_i|^2$ où

- $\sum_{i=1}^{n} p_i|OP_i|^2 = \sum_{j=1}^{h}\sum_{i=1}^{n} p_i(x'_{ij})^2 = h \times 1 = h$ est la variance du nuage de points, elle est appelé inertie totale du nuage par rapport au point O et notée $I(N,O)$,
- $\sum_{i=1}^{n} p_i|OH_i|^2 = \sum_{i=1}^{n} p_i\Psi_i^2$ est la variance expliquée par l'ajustement du nuage par la droite Δ, elle est appelée également inertie expliquée par l'axe Δ.

La qualité de l'ajustement sera d'autant meilleure que le quotient $\left(\sum_{i=1}^{n} p_i|OH_i|^2\right) / \left(\sum_{i=1}^{n} p_i|OP_i|^2\right)$ la part d'inertie expliquée par la droite Δ est élevé.

Exemple

Dans notre exemple, la contribution du point A à l'inertie du nuage est $I(A,O) = \sum_{j=1}^{4}((1/n) \times (x'_{ij})^2) = [1{,}86^2 + 0{,}74^2 + (-1{,}72)^2 + (-1{,}30)^2]/4 = 2{,}16$.

Généralisation. Envisageons maintenant le cas où le nuage de points N se situe à proximité d'un plan Π passant nécessairement par le point $O = (0,0,\ldots,0)$. Le meilleur plan permettant de visualiser le nuage est celui pour lequel la somme pondérée des carrés des distances des points P_i à ce plan Π est minimum.

3 Matrice d'inertie et détermination des axes principaux

La matrice d'inertie du nuage est $W = {}^tM' \cdot D \cdot M'$ où $D = (d_{ij})$ est la matrice diagonale de dimension n qui caractérise la répartition des poids ($d_{ii} = p_i$), soit

$$W = \begin{bmatrix} x'_{11} & x'_{21} & \ldots & x'_{n1} \\ x'_{12} & x'_{22} & \ldots & x'_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ x'_{1h} & x'_{2h} & \ldots & x'_{nh} \end{bmatrix} \begin{bmatrix} p_1 & 0 & \ldots & 0 \\ 0 & p_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & p_n \end{bmatrix} \begin{bmatrix} x'_{11} & x'_{12} & \ldots & x'_{1h} \\ x'_{21} & x'_{22} & \ldots & x'_{2h} \\ \ldots & \ldots & \ldots & \ldots \\ x'_{n1} & x'_{n2} & \ldots & x'_{nh} \end{bmatrix},$$

le terme général étant noté w_{ij}.

Cette matrice correspond à la matrice des corrélations entre variables si les poids p_i correspondent aux fréquences relatives (généralement $p_i = 1/n$).

© Dunod. La photocopie non autorisée est un délit.

– La trace de cette matrice est égale à l'inertie du nuage de points :

$$I(N,O) = \sum_{j=1}^{h} \sum_{i=1}^{n} p_i (x'_{ij})^2 = \text{Trace}(W)\,.$$

– On calcule valeurs propres λ_ℓ de cette matrice réelle symétrique W (c'est-à-dire les λ_ℓ vérifiant $\det(W - \lambda_\ell I_h) = 0$. Ces h valeurs propres positives peuvent être considérées comme distinctes et indexées selon leurs valeurs décroissantes : $\lambda_1 > \lambda_2 > \ldots > \lambda_h > 0$. Associons à chaque valeur propre λ_ℓ le vecteur propre unitaire $\overrightarrow{U}_\ell$ défini par : $\overrightarrow{U}_\ell = \begin{pmatrix} u_{\ell 1} \\ u_{\ell 2} \\ \cdots \\ u_{\ell h} \end{pmatrix}$ tel que $W \cdot \overrightarrow{U}_\ell = \lambda_\ell \overrightarrow{U}_\ell$ avec ${}^t\overrightarrow{U}_\ell \cdot \overrightarrow{U}_\ell = \sum_{j=1}^{h} u_{\ell j}^2 = 1$.

On obtient h vecteurs propres $\overrightarrow{U}_1, \overrightarrow{U}_2, \ldots, \overrightarrow{U}_h$ qui forment une nouvelle base orthonormée de $\mathbb{R}^h$. Ces vecteurs propres sont en fait les vecteurs directeurs des différents axes factoriels. Ainsi, la droite Δ_1 passant par O de vecteur directeur $\overrightarrow{U}_1$ est appelée premier axe factoriel, la droite Δ_2 passant par O et ayant $\overrightarrow{U}_2$ comme vecteur directeur est appelée second axe factoriel…

Exemple

Dans l'exemple précédent la matrice d'inertie $W = {}^t M' \cdot D \cdot M'$ est en fait la matrice des coefficients de corrélation et $D_{(6,6)}$ est la matrice diagonale de terme $d_{ii} = 1/6$. On a :

$$M'_{(6,4)} = \begin{pmatrix} 1{,}86 & 0{,}74 & -1{,}72 & -1{,}3 \\ 0{,}6 & 0{,}35 & 1{,}21 & 1{,}3 \\ -0{,}89 & 1{,}51 & -0{,}86 & -0{,}75 \\ -0{,}08 & -0{,}7 & 0{,}83 & 1{,}3 \\ -0{,}4 & -0{,}32 & 0{,}19 & 0{,}07 \\ -1{,}1 & -1{,}57 & 0{,}34 & -0{,}61 \end{pmatrix}$$

et $W = {}^t M' \cdot D \cdot M' = \begin{pmatrix} 1 & 0{,}36 & -0{,}37 & -0{,}07 \\ 0{,}36 & 1 & -0{,}55 & -0{,}27 \\ -0{,}37 & -0{,}55 & 1 & 0{,}89 \\ -0{,}07 & -0{,}27 & 0{,}89 & 1 \end{pmatrix}$

Les 4 valeurs propres λ_ℓ de cette matrice des coefficients de corrélation, sont les racines du polynôme caractéristique $P(\lambda) = \det \cdot ({}^t M' \cdot D \cdot M' - \lambda I_4) = 0$. On obtient $\lambda_1 = 2{,}33$, $\lambda_2 = 1{,}04$, $\lambda_3 = 0{,}60$, $\lambda_4 = 0{,}03$. À chaque valeur propre λ_ℓ on associe le vecteur unitaire $\overrightarrow{U}_\ell$ propre de la matrice W :

$$\overrightarrow{U}_1 = \begin{pmatrix} -0{,}33 \\ -0{,}46 \\ 0{,}63 \\ 0{,}53 \end{pmatrix}, \overrightarrow{U}_2 = \begin{pmatrix} 0{,}73 \\ 0{,}37 \\ 0{,}19 \\ 0{,}54 \end{pmatrix}, \overrightarrow{U}_3 = \begin{pmatrix} -0{,}57 \\ 0{,}78 \\ 0{,}10 \\ 0{,}21 \end{pmatrix}, \overrightarrow{U}_4 = \begin{pmatrix} 0{,}18 \\ 0{,}20 \\ 0{,}79 \\ -0{,}66 \end{pmatrix},$$

$\overrightarrow{U}_1$ et $\overrightarrow{U}_2$ sont respectivement les vecteurs directeurs des droites Δ_1 et Δ_2 appelés respectivement premier et second axe factoriel.

– La part d'inertie conservée par le premier axe principal Δ_1 est égale à $(\lambda_1 / \sum_{\ell=1}^{h} \lambda_\ell)$. Le plan Π ayant pour vecteur directeur $\overrightarrow{U}_1$ et $\overrightarrow{U}_2$ est le meilleur plan de projection. La part d'inertie conservée par le nuage de points projetés sur ce plan Π est égale à $(\lambda_1 + \lambda_2) / \sum_{\ell=1}^{h} \lambda_\ell$. Généralement de façon empirique, on retient pour l'analyse tous les axes dont les valeurs propres sont supérieures à 1 et donc tous les plans correspondants.

4 Nuage des points individus projetés sur un plan Π

Les composantes principales sont les nouvelles coordonnées que l'on obtient par projection des points-individus P_i sur l'axe factoriel Δ_ℓ de vecteur directeur $\overrightarrow{U}_\ell$ où $\ell = 1,2,\ldots,h$ ($\ell = 1$ pour le premier axe, $\ell = 2$ pour le deuxième axe…). (*Cf.* fig. 1). Les nouvelles coordonnées du point P_i sont notées $\Psi_{i\ell}$ (ainsi fig.1 pour $\Delta = \Delta_1$ on a $\overline{OH}_i = \Psi_{i1}$) et les n nouvelles coordonnées des n points P_i peuvent être représentées par le vecteur $\Psi_\ell = M' \cdot \overrightarrow{U}_\ell$ appelé composante principale :

$$\Psi_\ell = \begin{pmatrix} \Psi_{1\ell} \\ \Psi_{2\ell} \\ \cdots \\ \Psi_{n\ell} \end{pmatrix} \text{ où } {}^1\ \overline{\Psi}_\ell = \sum_{i=1}^{n} p_i \Psi_{i\ell} = 0\,,$$

$$\text{Var}\,(\Psi_\ell) = \sum_{i=1}^{n} p_i (\Psi_{i\ell} - \overline{\Psi}_\ell)^2 = \sum_{i=1}^{n} p_i (\Psi_{i\ell})^2 = \lambda_\ell\,.$$

Les variables $\Psi_1, \Psi_2, \ldots, \Psi_h$ sont considérées comme non liées car $\text{Cov}\,(\Psi_\ell, \Psi_k) = 0\ \forall k \neq \ell$ [2].

---

1. Soit $p = (p_1,\ldots,p_n)$. On a $\overline{\Psi}_\ell = p \cdot \Psi_\ell = p \cdot M \cdot U_\ell = (0,\ldots,0).U_\ell$. et $\sum_{i=1}^{n} p_i (\Psi_{\lambda_{\ell,i}})^2$
$= {}^t\Psi_\ell \cdot D \cdot \Psi_\ell = {}^t(M \cdot U_\ell).D \cdot (M \cdot U_\ell) = {}^tU_\ell({}^tM \cdot D \cdot M) \cdot U_\ell = {}^tU_\ell \cdot (W U_\ell)$
$= {}^tU_\ell \cdot (\lambda_\ell \cdot U_\ell) = \lambda_\ell \cdot ({}^tU_\ell \cdot U_\ell) = \lambda_\ell$

2. $\text{Cov}\,(\Psi_\ell, \Psi_k) = \sum_i p_i \Psi_{i\ell} \Psi_{ik} = {}^tU_\ell \cdot ({}^tM \cdot D \cdot M \cdot U_k) = {}^tU_\ell \cdot (\lambda_k U_k) = \lambda_k ({}^tU_\ell \cdot U_k) = 0$

© Dunod. La photocopie non autorisée est un délit.

Autrement dit, chaque composante principale Ψ_ℓ du nuage des points est centrée et a pour variance λ_ℓ. La variance expliquée par l'axe factoriel Δ_ℓ est donc égale à la valeur propre λ_ℓ et la part d'inertie conservée par la droite Δ_ℓ est $(\lambda_\ell / \sum_{\ell=1}^{h} \lambda_\ell)$.

Ψ_1 et Ψ_2 sont appelées première et seconde composante principale.

La qualité de la représentation de chaque point-individu P_i sur l'axe factoriel Δ_ℓ de direction $\overrightarrow{U}_\ell$ est évaluée par le cosinus de l'angle $\theta_{i\ell}$, entre l'axe et la droite OP_i soit $\cos^2(\theta_{i,\ell}) = |OH_i|^2/|OP_i|^2 = (\Psi_{i\ell})^2 / \sum_{j=1}^{h} (x'_{i,j})^2$.

La contribution à l'axe du point-individu P_{i_0} est égale à $(p_{i_0} \Psi^2_{i_0\ell} / \lambda_\ell)$.

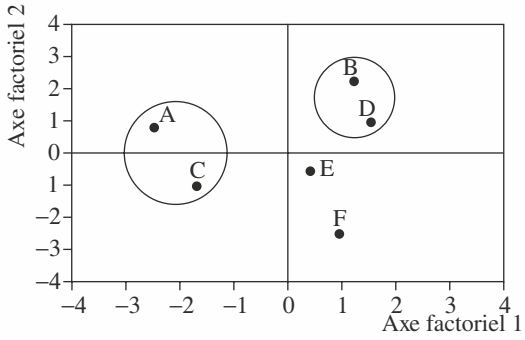
Exemple

Les nouvelles coordonnées des n individus dans la base (u_1, u_2) correspondent respectivement à $\Psi_{\lambda_1} = M' \cdot \overrightarrow{U}_1$ et $\Psi_{\lambda_2} = M' \cdot \overrightarrow{U}_2$. Ainsi

$$\Psi_{\lambda_1} = \begin{pmatrix} 1{,}86 & 0{,}74 & -1{,}72 & -1{,}3 \\ 0{,}6 & 0{,}35 & 1{,}21 & 1{,}3 \\ -0{,}89 & 1{,}51 & -0{,}86 & -0{,}75 \\ -0{,}08 & -0{,}7 & 0{,}83 & 1{,}3 \\ -0{,}4 & -0{,}32 & 0{,}19 & 0{,}07 \\ -1{,}1 & -1{,}57 & 0{,}34 & -0{,}61 \end{pmatrix} \begin{pmatrix} -0{,}33 \\ -0{,}46 \\ 0{,}63 \\ 0{,}53 \end{pmatrix} = \begin{pmatrix} -2{,}72 \\ 1{,}10 \\ -1{,}34 \\ 1{,}57 \\ 0{,}43 \\ 0{,}96 \end{pmatrix}$$

	$u_1 = \Psi_{\lambda_1}$	$u_2 = \Psi_{\lambda_2}$
A	– 2,72	0,91
B	1,10	2,26
C	– 1,34	– 1,00
D	1,57	0,81
E	0,43	– 0,50
F	0,96	– 2,48

Graphiquement plaçant les différents points-observations des six banques de A à F on obtient



– L'inertie du nuage de points $I(N,O) = \text{Trace}\,(W) = \sum_{\ell=1}^{4} \lambda_\ell = 4$. L'inertie expliquée par le premier axe est $\text{Var}(\Psi_{\lambda_1}) = (1/6)[(-2{,}72)^2 + (1{,}1)^2 + \ldots + 0{,}96^2] = 2{,}33$ valeur de la première valeur propre. L'inertie expliquée par le deuxième axe est $\text{Var}(\Psi_{\lambda_2}) = (1/6)[(0{,}91)^2 + \ldots + (-2{,}48)^2] = 1{,}04$ valeur de la deuxième valeur propre. Les contributions respectives à l'inertie des axes factoriels 1 et 2 sont respectivement 2,33/4 = 0,583 soit 58,3 % et 1,04/4 = 0,26 soit 26 %. La part d'inertie expliquée par le plan est donc de 84,3 %. La contribution du premier point-individu A à l'inertie expliquée par le premier axe est égale à $p_1 \times \Psi^2_{A,\lambda_1}/\lambda_1 = (1/6)((-2{,}72)^2/2{,}33) = 52{,}9$ %. La qualité de représentation du point A par l'axe 1 s'exprime par $\cos^2(\theta_{11}) = (\Psi_{A,1})^2/|OA|^2$ où θ_{11} est l'angle (OA, Δ_1), or $|OA|^2 = 1{,}86^2 + 0{,}74^2 + (-1{,}72)^2 + (-1{,}3)^2 = 8{,}64$ et $(\Psi_{A,1})^2 = (-2{,}72)^2$ donc $\cos^2(\theta_{11}) = (-2{,}72)^2/(8{,}64) = 0{,}856$. De même on obtient la qualité de représentation par l'axe 2 : $\cos^2(\theta_{12}) = (\Psi_{A,2})^2/|OA|^2 = (0{,}91)^2/8{,}64 = 0{,}096$ et l'on en déduit la qualité de représentation sur le plan en faisant la somme des deux cosinus carrés. Le premier axe oppose le premier groupe observations (B, D) au deuxième groupe (A, C). Le point E est très mal représenté.

5 Nuage des variables et composantes principales

Ces variables artificielles sont liées aux variables colonnes initiales centrées réduites X'_j par la relation : $\text{cor}\,(\Psi_\ell, X'_j) = \sqrt{\lambda_\ell} u_{\ell j}$ [1] où $u_{\ell j}$ est la j-ème composante du vecteur propre $\overrightarrow{U}_\ell$ de la matrice W dans la base canonique de R^h. Les valeurs de ces coefficients de corrélation permettent souvent de donner aux axes une interprétation concrète. La carte des variables représente dans un repère orthonormé les h points A_j de coordonnées $((\text{cor}\,(X'_j, \Psi_1), \text{cor}\,(X'_j, \Psi_2))$.

Exemple (suite) On a $\text{cor}\,(X'_1, \Psi_{\lambda_1}) = \lambda_1^{1/2} u_{11} = 2{,}33^{1/2} \times (-0{,}33) = -0{,}50$,
$\text{cor}\,(X'_2, \Psi_{\lambda_1}) = \lambda_1^{1/2} u_{21} = 2{,}33^{1/2} \times (-0{,}46) = -0{,}7$;

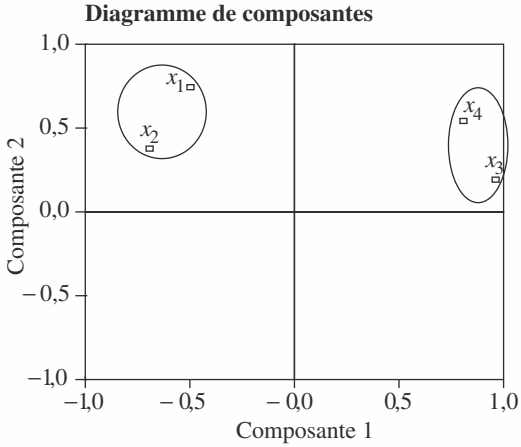
D'une façon plus générale sont présentées les corrélations entre les anciennes variables et les nouveaux axes, cela afin de déterminer quelles sont les variables qui contribuent le plus à la détermination de chacun des nouveaux axes.

Matrice des corrélations entre composantes et variables d'origine

	U_1	U_2	U_3	U_4
X'_1	– 0,50	0,74	– 0,45	0,03
X'_2	– 0,70	0,38	0,61	0,04
X'_3	0,97	0,19	0,08	0,14
X'_4	0,81	0,55	0,16	– 0,11

1. Les relations ${}^tM \cdot D \cdot M = \lambda_\ell U_\ell$ et $\Psi_\ell = M \cdot U_\ell$ impliquent $\text{Cov}\,(\Psi_\ell, X_j) = \sum_i x_{ij} \Psi_{i\ell} = \lambda_\ell u_{\ell j}$

© Dunod. La photocopie non autorisée est un délit.



Toutes les variables semblent être assez bien représentées puisque proches du cercle des corrélations. Dans l'exemple on s'aperçoit que l'axe 1 est tiré notamment par les variables x_3 (ROE return on equity) et x_2 (Cexp coefficient d'exploitation) qui s'opposent et sont donc corrélées négativement. L'axe 2 est tiré par x_1 (variable effectif). Deux groupes de variables très corrélés peuvent être visualisés (x_1,x_2) d'une part et (x_3,x_4) d'autre part, ces groupes de variables semblant être corrélés négativement.

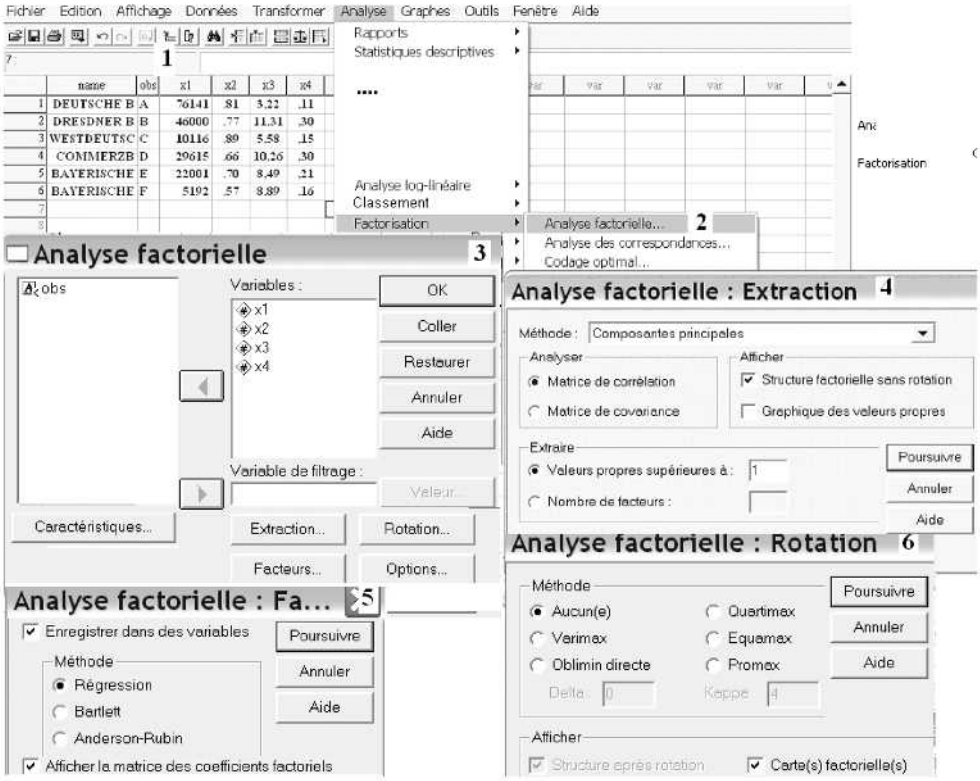
6 Représentation simultanée

Les deux nuages ne sont pas dans le même repère ce qui rend impossible la représentation simultanée des individus et des variables. Cependant, si l'on considère non plus des points-variables mais des directions de variable dans $\mathbb{R}^h$, on peut réaliser une représentation simultanée.

7 Procédure de traitement sous SPSS

7.1 Carte factorielle des variables

On reprend l'exemple précédent sur les banques et on souhaite déterminer les variables qui sont fortement corrélées entre-elles.



Procédure.

1. Une fois réalisée cette entrée de données,

2. On clique sur (Analyse) et l'on sélectionne dans le menu déroulant (Factorisation) puis (Analyse factorielle).

3. Dans le menu analyse factorielle on sélectionne x_1, x_2, x_3, x_4.

4. Cliquant sur (Extraction) on obtient le menu **Analyse factorielle : Extraction** dans lequel on sélectionne (valeurs propres supérieures à) 1.

5. Cliquant sur (Facteur) apparaît le menu **Analyse factorielle : Fa** dans lequel on sélectionne (Enregistrer dans les variables) et (Afficher la matrice des coefficients) l'enregistrement des composantes dans des variables permettra de représenter les points-individus.

6. Cliquant sur (Facteur) on obtient le menu **Analyse factorielle : rotation** dans lequel on sélectionne (Afficher cartes facorielles) ainsi apparaîtra la carte factorielle où figurent les corrélations entre variable.

© Dunod. La photocopie non autorisée est un délit.

Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	2,332	58,297	58,297	2,332	58,297	58,297
2	1,035	25,876	84,173	1,035	25,876	84,173
3	0,599	14,970	99,143			
4	0,034	0,857	100,000			

Dans ce tableau figurent les quatre axes factoriels associés aux quatre valeurs propres possibles λ obtenues en calculant $\det({}^tM \cdot D \cdot M - \lambda I) = 0$. L'inertie associée à chacun des axes ou part de la variance expliquée par les axes (la qualité de la représentation des nuages de point sur les axes) figure dans la troisième colonne. Elle est égale au rapport entre la valeur propre associée à la composante et la somme des valeurs propres. Dans cette analyse on constate que les deux premiers axes expliquent à eux deux plus de 84 % de l'inertie. Aussi l'ordinateur se contente-t-il de retenir pour l'analyse le plan constitué de l'axe 1 et 2. Dans la partie extraction du tableau, on constate qu'il n'y a que deux composantes extraites.

Il affiche également les corrélations entre les anciennes variables et les composantes 1 et 2, ce qui permet de montrer que l'axe factoriel 1 est essentiellement déterminé par X_3, X_4 et X_2 puisque les corrélations entre les variables d'origine et la composante 1 est 0,97, 0,81 et – 0,7, l'axe 2 est essentiellement déterminé par X_1.

Matrice des composantes

	Composante	
	1	2
X_1	– 0,50	0,74
X_2	– 0,70	0,38
X_3	0,97	0,19
X_4	0,81	0,55

Le logiciel fournit le graphe présenté p. 298 et interprété p. 298.

7.2 Carte factorielle des individus

Suite à la réalisation du **4** de la procédure, apparaissent dans la fenêtre des données les deux variables fac1_1 et fac2_1 correspondant aux deux axes factoriels et l'on peut alors réaliser la carte factorielle des points individus présentée p. 296.

	name	obs	x1	x2	x3	x4	fac1_	fac2_
1	DEUTSC	A	76141	,81	3,22	,11	-1,63	,55
2	DRESDN	B	46000	,77	11,31	,30	,66	1,35
3	WESTDE	C	10116	,89	5,58	,15	-,80	-,60
4	COMME	D	29615	,66	10,26	,30	,94	,48
5	BAYERIS	E	22001	,70	8,49	,21	,26	-,30
6	BAYERIS	F	5192	,57	8,89	,16	,57	-1,48

Procédure.

1. On sélectionne (Graphes), (Diagramme de dispersion).

2. Dans le menu Diagramme de dispersion on opte pour (Simple).

3. Dans le menu Diagramme de dispersion simple on retient **Fac_1** pour axe *X* et **Fac_2** pour axe *Y*.

Section 2

ANALYSE FACTORIELLE DES CORRESPONDANCES

1 Tableau de contingence

L'analyse factorielle des correspondances permet d'établir des correspondances entre deux caractères sur une population. Elle permet de visualiser sous forme de cartes graphiques des résultats statistiques figurant dans un tableau de contingence (tableau où l'on peut sommer les effectifs en ligne et en colonne).

Exemple

Disposant d'un échantillon de 13 085 entreprises, on souhaite déterminer le lien entre la taille de l'entreprise et la part du capital détenu par des personnes étrangères.

		Contrôle du capital			Total $n_{i\bullet}$
		Étranger majoritaire	Étranger minoritaire	National	
TAILLE	grande	50	5	200	255
	moyenne	352	24	1 600	1 976
	petite	856	76	9 922	10 854
Total $n_{\bullet j}$		1 258	105	11 722	13 085

lire: il y a 856 petites entreprises sous capital étranger majoritaire.

© Dunod. La photocopie non autorisée est un délit.

Le tableau des fréquences relatives et fréquences relatives cumulées en lignes et colonnes dans l'exemple étudié est présenté ci-dessous.

		Étranger majoritaire	Étranger minoritaire	National	Total
TAILLE	grande	f_{11} = 0,0038	f_{12} = 0,0004	f_{13} = 0,015	$f_{1\bullet}$ = 0,0195
	moyenne	f_{21} = 0,0269	f_{22} = 0,0018	f_{23} = 0,122	$f_{2\bullet}$ = 0,151
	petite	f_{31} = 0,0654	f_{32} = 0,0058	f_{33} = 0,7583	$f_{3\bullet}$ = 0,8295
Total $f_{\bullet j}$		$f_{\bullet 1}$ = 0,0961	$f_{\bullet 2}$ = 0,008	0,8958	*1,000*

lire: $f_{11} = 50/13085 = 0{,}004$, $f_{12} = 5/13085 = 0{,}004$

Plus précisément, on s'intéresse à deux variables X et Y définies sur une population P, les modalités possibles de X étant $a_1, a_2, \ldots, a_h$ et celles de Y étant $b_1, b_2, \ldots, b_k$. La population étudiée est donc répartie en $h \times k$ classes où l'effectif de la classe $a_i \times b_j$ sera noté n_{ij} et on construit le tableau des fréquences relatives $f_{ij} = n_{ij}/n$:

	b_1	…	b_j	…	b_k	Total marge en ligne
a_1	f_{11}	…	f_{1j}	…	$f_{1k} = n_{1k}/n$	$f_{1\bullet}$
…	…	…	…	…	…	…
a_i	f_{i1}	…	f_{ij}	…	$f_{ik} = n_{ik}/n$	$f_{i\bullet}$
…	…	…	…	…	…	…
a_h	f_{h1}	…	f_{hj}	…	$f_{hk} = n_{hk}/n$	$f_{h\bullet}$
Total en colonne	$f_{\bullet 1}$	…	$f_{\bullet j}$	…	$f_{\bullet k}$	1

De ce tableau on déduit :

– la matrice F d'ordre (h,k) et de terme général f_{ij},

– la matrice diagonale D_h d'ordre (h,h) dont les éléments diagonaux sont les marges en ligne $f_{i\bullet} = n_{i\bullet}/n$

– la matrice diagonale D_k d'ordre (k,k) dont les éléments diagonaux sont les marges en colonne $f_{\bullet j} = n_{\bullet j}/n$

Exemple

Ainsi la matrice des fréquences relatives F, la matrice diagonale des marges en ligne D_h et la matrice diagonale des marges en colonne D_k ont pour expression

$$F = \begin{pmatrix} 0{,}0038 & 0{,}0004 & 0{,}0153 \\ 0{,}0269 & 0{,}0018 & 0{,}1223 \\ 0{,}0654 & 0{,}0058 & 0{,}7583 \end{pmatrix} ; D_h = \begin{pmatrix} 0{,}0195 & 0 & 0 \\ 0 & 0{,}151 & 0 \\ 0 & 0 & 0{,}8295 \end{pmatrix} ;$$

$$D_k = \begin{pmatrix} 0{,}0961 & 0 & 0 \\ 0 & 0{,}008 & 0 \\ 0 & 0 & 0{,}8958 \end{pmatrix}.$$

2 Nuage des points-lignes dans R^k

Le tableau des profils-lignes a pour terme général $f_{ij}/f_{i\bullet}$. Il s'exprime de façon matricielle par $X = D_h^{-1} F$:

Y / Points-lignes	b_1	...	b_j	...	b_k	Total
Point L_1	$f_{11}/f_{1\bullet}$	...	$f_{1j}/f_{1\bullet}$	...	$f_{1k}/f_{1\bullet}$	1
...	...	...	...	...	...	...
Point L_i	$f_{i1}/f_{i\bullet}$	...	$f_{ij}/f_{i\bullet}$	...	$f_{ik}/f_{i\bullet}$	1
...	...	...	...	...	...	...
Point L_h	$f_{h1}/f_{h\bullet}$	...	$f_{hj}/f_{h\bullet}$	...	$f_{hk}/f_{h\bullet}$	1

Le nuage $N(L)$ constitué des h points lignes $L_1 = (f_{11}/f_{1\bullet},\ldots,f_{1k}/f_{1\bullet})$, $L_2 = (f_{21}/f_{2\bullet},\ldots,f_{2k}/f_{2\bullet})$, ..., $L_h = (f_{h1}/f_{h\bullet},\ldots,f_{hk}/f_{h\bullet})$ correspondant aux lignes du tableau est situé sur l'hyperplan d'équation $\sum_{i=1}^{k} x_i = 1$ puisque la somme des composantes de chacun des points vaut 1. À chaque profil ligne, on associe un poids $f_{i\bullet} = n_{i\bullet}/n$ et le centre de gravité de ce nuage est alors le point $G_L = (f_{\bullet 1},\ldots,f_{\bullet k})$.

La distance « du khi-deux » entre deux profils-lignes L_{i_0} et L_{i_1}

$||L_{i_0}L_{i_1}||^2 = \sum_{i=1}^{k}(f_{i_0 j}/f_{i_0\bullet} - f_{i_1 j}/f_{i_1\bullet})^2/f_{\bullet j}$ attribue à chaque écart $|f_{i_0 j}/f_{i_0\bullet} - f_{i_1 j}/f_{i_1\bullet}|$ le poids $1/\sqrt{f_{\bullet j}}$ qui permet d'augmenter l'importance des écarts dans les colonnes de faible poids, donnant ainsi à chaque colonne la même importance dans l'évaluation de la distance entre deux point-lignes.

© Dunod. La photocopie non autorisée est un délit.

Exemple

Le tableau des profils lignes correspond à la division des fréquences relatives f_{ij} par les fréquences relatives cumulées en ligne $f_{i\bullet} = n_{i\bullet}/n$, matriciellement $X = D_h^{-1}F = \begin{pmatrix} 0{,}1961 & 0{,}0196 & 0{,}7843 \\ 0{,}1781 & 0{,}0121 & 0{,}8097 \\ 0{,}0789 & 0{,}007 & 0{,}9141 \end{pmatrix}$

Profil ligne

	Étranger majoritaire	Étranger minoritaire	National	Total
Grande	0,1961	0,0196	0,7843	1
Moyenne	0,1781	0,0121	0,8097	1
Petite	0,0789	0,007	0,9141	1
Profil moyen	0,0961	0,008	0,8958	

Ainsi 19,61 % des grandes entreprises sont sous contrôle étranger majoritaire, 1,96 % sous contrôle étranger minoritaire, et 78,43 % sous contrôle d'une personne nationale. Dans l'exemple précédent, on a $G_L = (0{,}0961, 0{,}008, 0{,}8958)$ car $f_{\bullet 1} = 0{,}0961$, $f_{\bullet 2} = 0{,}008$, $f_{\bullet 3} = 0{,}8958$ et $\mathrm{d}^2(\ell_1, \ell_2) = (0{,}1961 - 0{,}1781)^2/0{,}0961 + (0{,}0196 - 0{,}0121)^2/0{,}008 + (0{,}7843 - 0{,}8097)^2/0{,}08958 = 0{,}011$.

L'inertie totale du nuage est $I(N, G_L) = \sum_{i=1}^{h}\sum_{j=1}^{k}(f_{ij} - f_{i\bullet} \times f_{\bullet j})^2/(f_{i\bullet} \times f_{\bullet j})$ expression du Khi-deux, calculé sur le tableau des fréquences relatives. En effet,

$$(\text{poids en } L_i) \times ||G_L L_i||^2 = f_{i\bullet} \sum_{j=1}^{k} \left(\frac{|f_{ij}/f_{i\bullet} - f_{\bullet j}|}{f_{\bullet j}^{0,5}} \right)^2$$

La *matrice d'inertie du nuage des points lignes pour la métrique* D_k^{-1} est $W = ({}^tF \cdot D_h^{-1} \cdot F) \cdot D_k^{-1}$ et l'inertie du nuage $I(N, G_L) = \text{Trace}\,(W)$.

Les k valeurs propres λ_i de la matrice réelle symétrique W sont indexées selon leurs valeurs décroissantes : $\lambda_0 = 1 > \lambda_1 > \ldots > \lambda_{k-1} > 0$. Associant à chaque valeur propre λ_ℓ le vecteur colonne propre $\overrightarrow{U}_\ell$ de la matrice W caractérisée par « $W \cdot \overrightarrow{U}_\ell = \lambda_\ell \overrightarrow{U}_\ell$ et ${}^t\overrightarrow{U}_\ell \cdot D_k^{-1} \cdot \overrightarrow{U}_\ell = 1$ » on obtient une base orthonormée de $\mathbb{R}^k$ pour la métrique caractérisée par D_k^{-1}, constituée des vecteurs $\overrightarrow{U}_0, \overrightarrow{U}_1, \overrightarrow{U}_2, \ldots, \overrightarrow{U}_{k-1}$. Ainsi à $\lambda_0 = 1$ correspond le vecteur propre $\overrightarrow{U}_0 = \vec{g}$ barycentre du nuage $N(L)$, il oriente l'axe factoriel qui joint l'origine O au centre de gravité G_L. À λ_1 correspond le vecteur $\overrightarrow{U}_1$ qui donne la direction du premier axe principal Δ_1 passant par O. À λ_2 correspond le vecteur $\overrightarrow{U}_2$ qui donne la direction du second axe principal Δ_2.

Si l'on procède à une analyse par rapport au centre de gravité en excluant la première valeur propre $\lambda_0 = 1$, on en déduit la part d'inertie expliquée par l'axe Δ_1 est $\lambda_1 / \left(\sum_{i=1}^{k-1} \lambda_i \right)$. La part d'inertie expliquée par la droite Δ_2 est $\lambda_2 / \left(\sum_{i=1}^{k-1} \lambda_i \right)$. La part d'inertie expliquée par le plan Π est égale à $(\lambda_1 + \lambda_2) / \left(\sum_{i=0}^{k-1} \lambda_i - 1 \right)$.

Exemple

Par calcul matriciel on obtient la matrice $W = {}^t F \cdot D_h^{-1} \cdot F \cdot D_k^{-1} = \begin{pmatrix} 0{,}11130 & 0{,}10711 & 0{,}09441 \\ 0{,}00894 & 0{,}00878 & 0{,}00792 \\ 0{,}87976 & 0{,}88408 & 0{,}89766 \end{pmatrix}$, son polynôme caractéristique $\det(W - \lambda I_3) = \lambda^3 - 1{,}0177\lambda^2 + 0{,}0177015386\lambda - 0{,}0000015386$ puis ses racines : $\lambda_0 = 1$, $\lambda_1 = 0{,}01766$, $\lambda_2 = 0{,}000084$. L'inertie totale du nuage par rapport à l'origine O est égale à Trace $(W) = 1{,}0177$ et donc l'inertie à expliquer, si on analyse par rapport au point G_L, est Trace $(W) - 1 = 0{,}0177$.

À λ_1 correspond le vecteur $\overrightarrow{U}_1 = (x, y, z)$ défini par $\overrightarrow{U}_1 \cdot W = 0{,}01766 \overrightarrow{U}_1$; on trouve des solutions de type $x = -0{,}94260232z$ et $y = -0{,}05739768z$. Tenant compte de $\overrightarrow{U}_1 \cdot D_k^{-1} \cdot {}^t \overrightarrow{U}_1 = 1$ on choisit $\overrightarrow{U}_1 = (-0{,}287244,\ 0{,}017491,\ 0{,}304735)$ qui donne la direction du premier axe principal passant par G_L. Le même type de calcul pour $\lambda_2 = 0{,}000084$ donne $x = 3{,}11857739z$ et $y = -4{,}11857739z$ puis $\overrightarrow{U}_2 = (0{,}00140721, -0{,}00185844, 0{,}00045123)$.

Les h projections du nuage $N(L)$ sur l'axe orienté par le vecteur $\overrightarrow{U}_\ell$ sont représentées par la matrice colonne $\Psi_\ell = X \cdot D_k^{-1} \cdot {}^t \overrightarrow{U}_\ell$ où $X = D_h^{-1} F$.

La projection du nuage des points-lignes L_i sur le plan Π formé par les deux premiers axes factoriels Δ_1 et Δ_2 donne un schéma appelé carte factorielle des profils lignes. Il est à remarquer que $\sum_{i=1}^{h} f_{i\bullet} \Psi_{i,\ell} = 0$ et $\sum_{i=1}^{h} f_{i\bullet} \Psi_{i,\ell}^2 = \lambda_\ell$.

– Pour interpréter correctement les résultats, il faut calculer les *contributions absolues* des point-lignes qui expriment la part prise par une modalité de la variable dans l'inertie expliquée par l'axe. Pour les point-lignes, la part de la contribution de L_i à la variance prise en compte par l'axe de direction $\overrightarrow{U}_\ell$ est égale à $(f_{i\bullet} \Psi_\ell^2 / \lambda_\ell)$.

– Les cosinus-carrés, parfois appelés *contributions relatives* permettent de déterminer si un point-ligne est ou non bien représenté. Ainsi la qualité de la représentation du point-ligne L_{i_0} sur l'axe Δ_ℓ de direction $\overrightarrow{U}_\ell$ est évaluée par le cosinus de l'angle entre l'axe et le vecteur joignant le centre de gravité G_L au point L_{i_0} :

$$\cos^2(\overrightarrow{U}_\ell, \overrightarrow{G_L L_{i_0}}) = (\Psi_{i_0,\ell})^2 / ||L_{i_0} G_L||^2 = (\Psi_{i_0,\ell})^2 \left[\sum_{j=1}^{k} (f_{i_0 j} / f_{i_0 \bullet} - f_{\bullet j})^2 / f_{\bullet j} \right].$$

© Dunod. La photocopie non autorisée est un délit.

La qualité de la représentation sur un plan du point-ligne L_{i_0} est égale à la somme des valeurs exprimant la qualité de la représentation sur chacun des axes constituant la plan. Prenant en compte tous les axes possibles (en fait, $k-1$ possibilités) on a évidemment la somme des cosinus carrés des projections qui est égale à 1.

Exemple (suite)

Les projections des points lignes sur le premier axe et second sont respectivement $\Psi_1 = X \cdot D_k^{-1} \cdot {}^t\overrightarrow{U}_1 = \begin{pmatrix} -0{,}362 \\ -0{,}283 \\ 0{,}060 \end{pmatrix}$ et $\Psi_2 = \begin{pmatrix} 0{,}0601 \\ -0{,}0095 \\ 0{,}0003 \end{pmatrix}$, ainsi le premier point-ligne a dans le plan pour coordonnées $(-0{,}362, 0{,}0601)$, le deuxième point-ligne a pour coordonnées $(-0{,}283, -0{,}0095)$, le troisième $(0{,}057, 0{,}0003)$. La contribution du premier point-ligne (les grandes entreprises) à la variance prise en compte par l'axe de direction $\overrightarrow{U}_1$ est $(f_{1\bullet}\Psi^2_{L_1,1}/\lambda_1) = 0{,}0195 \times (-0{,}362)^2/0{,}01766 = 0{,}144$. S'intéressant à la qualité de la représentation du premier point-ligne (les grandes entreprises) sur le premier axe on constate que $\cos^2(\overrightarrow{U}_1, \overrightarrow{G_L L_1}) = (\Psi_{L_1,1})^2/||L_1 G_L||^2$

$$= (\Psi_{L_1,1})^2 \Big/ \left[\sum_{j=1}^{k} (f_{1j}/f_{1\bullet} - f_{\bullet j})^2 / f_{\bullet j} \right] = 0{,}973$$

car $f_{11} = 0{,}0038, f_{12} = 0{,}0004, f_{13} = 0{,}0153, f_{1\bullet} = 0{,}0195, f_{\bullet 1} = 0{,}0961, f_{\bullet 2} = 0{,}008,$ $f_{\bullet 3} = 0{,}8958$ et $\Psi_{L_1,1} = (-0{,}362)$. De même $\cos^2(\overrightarrow{U}_1, \overrightarrow{G_L L_2}) = (\Psi_{L_1,2})^2/||L_2 G_L||^2$ $= 0{,}027$. Ici $k-1 = 2$ donc la somme des deux cosinus carrés est nécessairement égale à 1.

3 Nuage des points-colonnes dans $\mathbb{R}^h$

Le tableau des profils en colonne a pour terme général $f_{ij}/f_{\bullet j}$ il est obtenu en divisant les fréquences relatives par les marges en colonne $f_{\bullet j}$, il se calcule de façon matricielle par $Y = F D_k^{-1}$.

Points-colonnes / ***X***	**Point C_1**	**…**	**Point C_j**	**…**	**Point C_K**	**Total**
a_1	$f_{11}/f_{\bullet 1}$	…	$f_{1j}/f_{\bullet j}$	…	$f_{1k}/f_{\bullet k}$	1
…	…	…	…	…	…	…
a_i	$f_{i1}/f_{\bullet 1}$	…	$f_{ij}/f_{\bullet j}$	…	$f_{ik}/f_{\bullet k}$	1
…	…	…	…	…	…	…
a_h	$f_{h1}/f_{\bullet 1}$	…	$f_{hj}/f_{\bullet j}$	…	$f_{hk}/f_{\bullet k}$	1
Total	1		1		1	

Le centre de gravité G_C du nuage $N(C)$ a pour coordonnées $(f_{1\bullet}, \ldots, f_{h\bullet})$. La matrice d'inertie est $W' = (F \cdot D_k^{-1} \cdot {}^t F) D_h^{-1}$. Les valeurs propres non nulles μ_ℓ de

W' sont identiques aux valeurs propres de W. On détermine les vecteurs propres associés à la matrice d'inertie du nuage des points colonnes $W' \cdot \overrightarrow{V}_\ell = \mu_\ell \overrightarrow{V}_\ell$ où $({}^t\overrightarrow{V}_\ell \cdot D_h^{-1} \cdot \overrightarrow{V}_\ell) = 1$ puis les axes principaux Δ'_ℓ.

Exemple

Le tableau des profils en colonne correspond à la division des fréquences relatives par les marges en ligne $f_{\bullet j} = n_{\bullet j}/n$, soit matriciellement $Y = FD_K^{-1} = \begin{pmatrix} 0{,}0397 & 0{,}0476 & 0{,}0171 \\ 0{,}2798 & 0{,}2286 & 0{,}1365 \\ 0{,}6804 & 0{,}7238 & 0{,}8464 \end{pmatrix}$.

Profil colonne

	Étranger majoritaire	Étranger minoritaire	National	Profil moyen
Grande	0,0397	0,0476	0,0171	0,0195
Moyenne	0,2798	0,2286	0,1365	0,151
Petite	0,6804	0,7238	0,8464	0,8295
Total	1	1	1	

Pour les points-colonnes le centre de gravité est $G_C = (0{,}096, 0{,}008, 0{,}896)$

$W' = \begin{pmatrix} 0{,}0221 & 0{,}0215 & 0{,}0191 \\ 0{,}1664 & 0{,}1631 & 0{,}1484 \\ 0{,}8115 & 0{,}8154 & 0{,}8325 \end{pmatrix}$ et les trois racines de son polynôme caractéristique sont identiques à celles de W : $\lambda_0 = 1$, $\lambda_1 = 0{,}01766$, $\lambda_2 = 0{,}000084$. Inertie totale : Trace $(W) - 1 = 0{,}0177$.

À λ_1 correspond le vecteur ${}^t\overrightarrow{V}_1 = (-0{,}0531, -0{,}3219, 0{,}3749)$.

À λ_2 correspond ${}^t\overrightarrow{V}_2 = (0{,}1276, \ -0{,}1568, 0{,}0292)$.

Les coordonnées factorielles correspondant à l'axe orienté par le vecteur $\overrightarrow{V}_\ell$, ou projections des différents points-colonnes sur Δ'_ℓ, sont $\Psi'_\ell = D_k^{-1} \times {}^tF \times D_h^{-1} \times \overrightarrow{V}_\ell$.

$$\cos^2(\overrightarrow{V}_\ell, \overrightarrow{G_C C'_{j0}}) = (\Psi'_{j0,\ell})^2 / ||C_{j0}, G_L||^2 = (\Psi'_{j0,\ell})^2 \left(\sum_{i=1}^{h} (f_{ij}/f_{\bullet j} - f_{i\bullet})^2 / f_{i\bullet} \right).$$

La carte factorielle des profils-colonnes correspond à la projection du nuage des points-colonnes C_j sur le plan de projection formé par les premiers axes factoriels pris deux à deux, soit le plan factoriel $\Pi(O, \overrightarrow{V}_1, \overrightarrow{V}_2)$.

4 Représentation simultanée

Les matrices W et W' ayant mêmes valeurs propres non nulles, l'on identifie les axes principaux Δ_ℓ et Δ'_ℓ et l'on réalise sur un même graphe les deux représentations. La proximité de deux points-lignes L_{i_0} et L_{i_1} traduit un comportement analo-

© Dunod. La photocopie non autorisée est un délit.

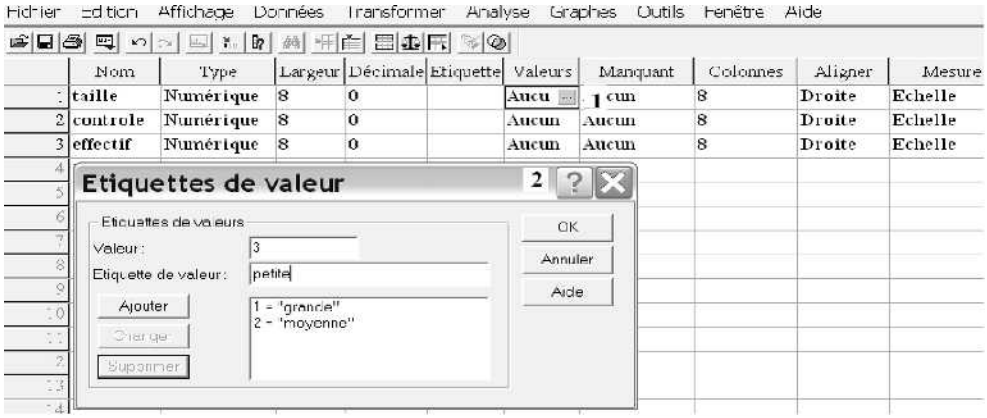
gue vis-à-vis de la variable Y. La proximité de deux points-colonnes traduit un comportement analogue vis-à-vis de la variable X. La proximité d'un point-ligne L_i et d'un point colonne C_j établit une forte interaction entre la modalité i du caractère X et la modalité j du caractère Y.

5 Procédures de traitement sous SPSS

Considérons l'exemple précédent où l'on souhaite établir une correspondance entre le type de contrôle exercé sur l'entreprise et sa catégorie de taille. Les valeurs prises par les variables doivent être numériques pour le logiciel. Aussi nous décidons d'attribuer pour la variable taille la valeur 1 aux grandes entreprises, la valeur 2 aux entreprises moyennes et 3 aux petites entreprises. Pour la variable nature du contrôle, nous attribuons la valeur 1 à une entreprise sous contrôle national , 2 à un contrôle étranger minoritaire et 3 à un contrôle étranger majoritaire.

Pour établir un lien direct et lisible entre les valeurs de la variable taille et les modalités du contrôle de l'entreprise il faut recoder les caractères en variables numériques, pondérer les observations et enfin utiliser le programme d'analyse factorielle des correspondances.

a) Les caractères doivent être recodés en variables numériques.

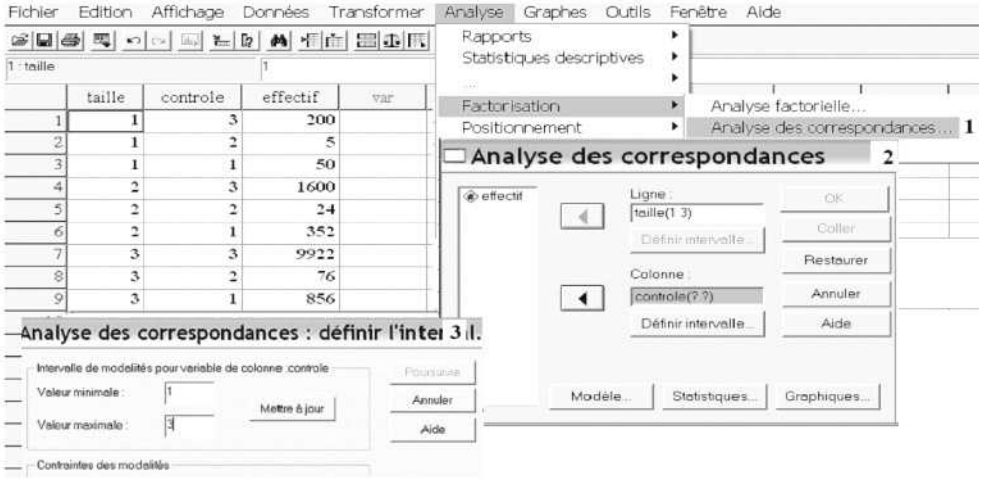


Procédure.

1. Cliquer sur (Affichage) puis sur (variables).

2. Cliquer dans la ligne de la variable taille sur (Valeurs) et dans le menu **étiquettes de valeurs** on affecte la valeur 1 à grande puis on clique sur ajouter, etc. On fait de même pour la variable taille.

b) Les données des variables étant pondérées par la variable effectif on peut effectuer l'analyse des correspondances.



Procédure.

1. On clique sur (Analyse) et on sélectionne dans le menu déroulant (Factorisation) puis (Analyse des correspondances).

2. Dans le menu analyse des correspondances on sélectionne en ligne la taille et en colonne le contrôle.

3. On précise en cliquant sur (Définir intervalle) les modalités de caractères étudiés pour la variable contrôle en mettant dans le menu qui apparaît la valeur 1 pour minimum et 3 pour maximum, on fait de même pour la variable taille. Cliquant sur (Mettre à jour) puis (Poursuivre) et **ok** on obtient les résultats suivants.

Résultats

	Valeur singulière	Inertie	Khi-deux	Sig.	Proportion d'inertie	
Dimension					**Expliquée**	**Cumulée**
1	0,1329	0,01766			0,9953	0,9953
2	0,0092	0,00008			0,0047	1
Total		0,01774	232,1906	0	1	1

Commentaires. Les deuxièmes et troisièmes valeurs propres λ_1 et λ_2 communes aux matrices d'inertie des points-lignes W et des points-colonnes W' sont respectivement 0,01766 et 0,00008. L'inertie totale associée au premier plan de projection dont les axes sont les dimensions 1 et 2 orientés par les vecteurs propres associés à λ_1 et λ_2 est égale à $0{,}01766 + 0{,}00008 = 0{,}01774$ (= somme des deux valeurs propres). La part d'inertie expliquée par le premier axe est (0,01766)/0,01744 = 0,9953. La valeur de l'inertie est un indicateur de la dispersion du nuage et mesure la liaison entre deux variables. Cette inertie totale (somme des valeurs propres différentes de 1) multipliée par l'effectif total donne la valeur 232,1906. Le test d'indépendance du Khi-deux ayant un niveau de signification inférieur 0,0000, on conclut au rejet de l'hypothèse d'indépendance entre la variable type de contrôle et taille des entreprises.

© Dunod. La photocopie non autorisée est un délit.

Les deux tableaux suivants donnent des précisions supplémentaires sur les profils-lignes et les profils colonnes.

Caractéristiques des points-lignes

	Masse	Score dans la dimension		Inertie	Contribution			
		1	2		De point à inertie de dimension		De dimension à inertie de point	
TAILLE					1	2	1	2
grande	0,019	– 0,992	0,627	0,003	0,144	0,836	0,973	0,027
moyenne	0,151	– 0,777	– 0,099	0,012	0,686	0,163	0,999	0,001
petite	0,829	0,165	0,003	0,003	0,169	0,001	1	0
Total actif	1			0,018	1	1		

La Masse désigne les $f_{i\bullet}$ ou marges en ligne (ou fréquences relatives associées aux différentes valeurs de la variable taille). Les scores dans la dimension 1 : $\begin{pmatrix} -0{,}992 \\ -0{,}777 \\ 0{,}165 \end{pmatrix}$ et dans la dimension 2 : $\begin{pmatrix} 0{,}627 \\ -0{,}099 \\ 0{,}003 \end{pmatrix}$ sont les projections des points lignes sur les axes 1 et 2. Il est à remarquer que ces projections correspondent à celles obtenues par calcul (p. 326) à un coefficient de dilatation = 1/(*valeur singulière*)^{0,5}, soit pour le premier axe $\Psi_1/(0{,}1329)^{0,5}$ ou de façon équivalente

$$\Psi_1 = (0{,}1329)^{0,5} \times \begin{pmatrix} -0{,}992 \\ -0{,}777 \\ 0{,}165 \end{pmatrix} = \begin{pmatrix} -0{,}362 \\ -0{,}283 \\ 0{,}060 \end{pmatrix}.$$

Contribution des point-lignes à la détermination des axes. Les contributions absolues expriment la part prise par une modalité de la variable dans l'inertie expliquée par un facteur. Par exemple, pour les point-lignes cette contribution des grandes firmes à la variance prise en compte par l'axe 1 est égale à $(f_{1\bullet}\Psi^2_{grande,1}/\lambda_1) = 0{,}019 \times [-0{,}992 \times (0{,}1329)^{0,5}]^2/0{,}0176 = 0{,}144$, etc. Ici c'est la modalité entreprise moyenne qui contribue le plus à l'inertie du nuage projeté des point-lignes sur l'axe 1. En revanche c'est la modalité grande-entreprise qui contribue le plus à l'inertie du nuage des point-lignes projeté sur l'axe 2.

Qualité de la représentation. Les cosinus carrés permettent d'apprécier si un point est bien représenté sur un sous-espace factoriel. Ainsi la qualité de la représentation du point-ligne ✎ grandes entreprises ✎ sur l'axe 1 est évaluée par le cosinus de l'angle entre l'axe et le vecteur joignant ce point au centre de gravité G_L : $\cos^2(\overrightarrow{U}_1, grande) = \Psi^2_{grande,1}/||grande, G_L||^2 = 0{,}973$, et la qualité de la représentation sur le plan est égale à la somme des cosinus carrés sur les deux axes.

– La même analyse peut être faite sur le nuage des points-colonnes

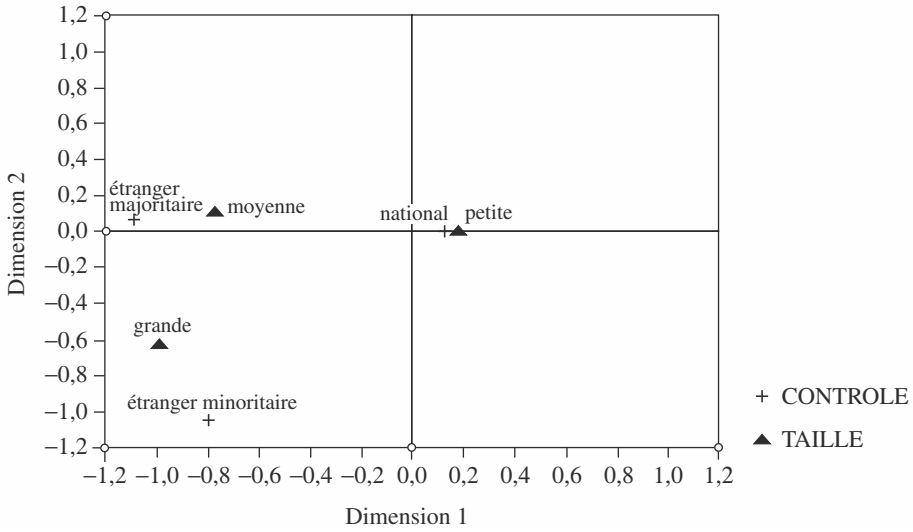
	Masse	Score dans la dimension		Inertie	Contribution			
		1	2		De point à inertie de dimension		De dimension à inertie de point	
Contrôle					1	2	1	2
Étranger majoritaire	0,096	– 1,089	0,066	0,015	0,858	0,046	1	0,0003
Étranger minoritaire	0,008	– 0,795	– 1,044	0,001	0,038	0,954	0,894	0,1065
National	0,896	0,124	0,002	0,002	0,104	0,001	1	0
Total actif	1			0,018	1	1		

Contribution des points-colonnes à la détermination des axes. La modalité « entreprises sous contrôle étranger majoritaire » contribue le plus à la variance prise en compte par l'axe 1, sa contribution étant égale à 0,858. La modalité « entreprises avec actionnaires étrangers minoritaires » contribue le plus à l'axe 2.

Commentaire du graphe présenté ci-après. L'analyse suivant le premier axe nous permet d'affirmer que la présence majoritaire d'actionnaires étrangers se réalise dans les entreprises de taille moyenne. L'analyse suivant l'axe 2 nous permet de voir que ce sont les modalités grandes entreprises, et contrôle étranger minoritaire qui contribuent le mieux à l'axe.

Sur le graphe on observe une proximité entre la modalité petite entreprise du caractère taille et la modalité nationale de contrôle de l'entreprise ainsi qu'une proximité entre la modalité taille moyenne et une présence majoritaire des actionnaires étrangers. Les petites entreprises ont tendance à être caractérisées par la présence de seuls actionnaires nationaux alors que nombre d'entreprises moyennes sont sous le contrôle d'actionnaires étrangers majoritaires. Les grandes entreprises seraient caractérisées par la présence d'actionnaires étrangers minoritaires.

© Dunod. La photocopie non autorisée est un délit.



Pour une étude approfondie de l'analyse des données le lecteur pourra utilement consulter l'ouvrage *Méthodes statistiques en gestion* de TENENHAUS (M.) Dunod 1996.

CORRECTIONS DES EXERCICES

CHAPITRE 1

Exercice 1

Le caractère x représente la note sur 20 attribuée à chacun des 100 candidats soumis à un test d'embauche. Valeurs, effectifs n_i, fréquences relatives f_i, fréquences relatives cumulées $F(a_i)$ où a_i est l'extrémité supérieure de la classe sont donnés par le tableau suivant :

Valeurs	**[0,5]**	**]5,8]**	**]8,10]**	**]10,12]**	**]12,15]**	**]15,20]**	**Total**
Effectifs n_i	25	18	12	20	15	10	100
Fréquences relatives $f_i = n_i/100$	0,25	0,18	0,12	0,2	0,15	0,1	1
Fréquences relatives cumulées $F(a_i) = \pi_i = f_1 + f_2 + \ldots f_i$	0,25	0,43	0,55	0,75	0,9	1	
Fréq rel, par amplitude de classe $y_i' = f_i/(a_i - a_{i-1})$	0,05 (=0,25/(5-0).	0,06 (=0,18/(8-5).	0,06	0,1	0,05	0,02	

1. La classe]10,12] est la classe modale car ayant le plus grand rapport $f_i/(a_i - a_{i-1})$ et le mode est le centre de classe : $x_{m_0} = 11$.

© Dunod. La photocopie non autorisée est un délit.

La *classe médiane* est la classe $[a_{i-1},a_i[$ qui doit satisfaire aux conditions « $F(a_i) \geqslant 0{,}5$ et $F(a_{i-1}) < 0{,}5$ ». Donc]8,10] est la classe médiane et la médiane m_e des valeurs de l'échantillon est estimée par

$$m'_e = 8 + (10 - 8)(0{,}5 - 0{,}43)/(0{,}55 - 0{,}43) = 9{,}166.$$

La *classe du premier quartile* est la classe $]a_{i-1},a_i]$ telle que $F(a_i) \geqslant 0{,}25$ et $F(a_{i-1}) < 0{,}25$ soit la classe [0, 5] et le premier quartile q_1 des valeurs de l'échantillon est estimée par

$$q'_1 = 0 + (5 - 0) \times (0{,}25 - 0)/(0{,}25 - 0) = 5.$$

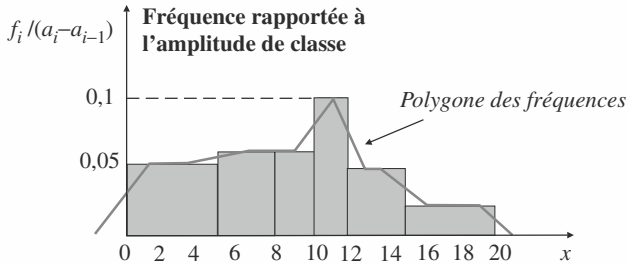
La *classe du troisième quartile* est la classe $]a_{i-1},a_i]$ telle que $F(a_i) \geqslant 0{,}75$ et $F(a_{i-1}) < 0{,}75$ soit la classe]10,12] et

$$q'_3 = a_{i-1} + (a_i - a_{i-1}) \times (0{,}75 - F(a_{i-1}))/(F(a_i) - F(a_{i-1})) = 12\,.$$

Intervalle interquartile =12- 5 = 7

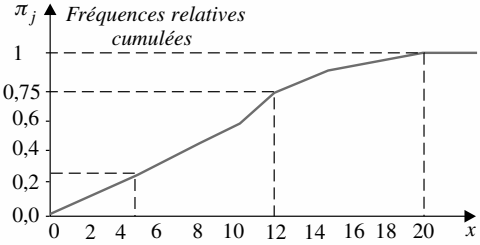
2. Histogramme et polygone des fréquences

Histogramme. Pour la première classe on porte en abscisse 0 et 5 et on porte en ordonnée $f_i/(a_i - a_{i-1}) = 0{,}05,\ldots$, pour la dernière classe on porte en abscisse 15 et 20 et en ordonnée $f_i/(a_i - a_{i-1}) = 0{,}02$.



Polygone des fréquences. L'amplitude de classe minimale est $\theta = 2$, on obtient le polygone en joignant $M_0 = (a_0 - \theta/2,0)$ avec $a_0 = 0$ d'où $M_0 = (-1, 0)$, $M_1 = (a_0 + \theta/2, y'_1) = (0 + 1, 0{,}05) = (1, 0{,}02)$; $M'_1 = (a_1 - \theta/2, y'_1) = (5 - 1, 0{,}05) = (4, 0{,}05),\ldots$, $M_6 = (15 + 1, 0{,}02)$, $M'_6 = (20 - 1, 0{,}02)$, $M_7 = (20 + 1, 0)$.

3. Courbe cumulative des fréquences



Exercice 2

Le caractère x représente les salaires mensuels des 400 salariés d'une entreprise.

1. Valeurs, effectifs n_i, fréquences relatives f_i, fréquences relatives cumulées $F(a_h)$, sommes S_h et proportion $\rho(a_h)$ des salaires ordonnés et cumulés sont donnés par le tableau suivant :

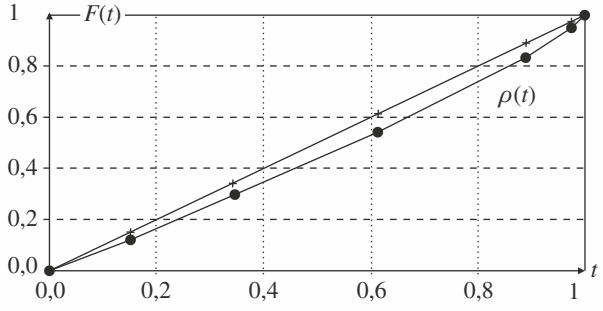
Valeurs en euros	**[750-800]**	**]800- 850]**	**]850-900]**	**]900-1 100]**	**]1 100-1 500]**	**]1 500-2 000]**	**Total**
Effectifs n_i	60	80	105	110	35	10	400
Fréquences relatives f_i	0,150	0,200	0,263	0,275	0,088	0,025	1
Fréq. rel. cum $F(a_h)$	0,150	0,350	0,613	0,888	0,975	1,000	
x_i^* centre de classe	775	825	875	1 000	1 300	1 750	
$n_i \times x_i^*$	46 500	66 000	91 875	110 000	45 500	17 500	377 375
$S_h = \sum_{i \leqslant h} n_i \times x_i^*$	46 500	112 500	204 375	314 375	359 875	377 375	
$\rho(a_h) = S_h/377\,375$	0,12	0,30	0,54	0,83	0,95	1,00	

2. Pour estimer la moyenne arithmétique $\overline{x}$ des valeurs de l'échantillon on détermine les centres de classe $\overline{x}_i^*$ (où $x_i^* = (a_{i-1} + a_i)/2$ est le centre de la i-ème classe) et l'on en déduit

$$\overline{x} \cong \overline{x}' = \frac{1}{n} \sum_{i=1}^{K} n_i x_i^* = (60 \times 775 + 80 \times 825 + \ldots 10 \times 1\,750)/400 = 943{,}4\,.$$

3. $V_x = \sigma^2 \cong \dfrac{1}{n} \sum_{i=1}^{K} n_i (x_i^*)^2 - \overline{x}'^2 = (60 \times 7\,75^2 + \ldots 10 \times 17\,50^2)/400 - 943{,}4^2$
$= 36\,558$ d'où $\sigma = 191{,}2$. Intervalle interquartile : $999{,}63 - 825 = 174{,}63$

4. En représentant les 6 points $P_h = (F(a_h); \rho(a_h))$: $P_1 = (0{,}15; 0{,}12)$, $P_2 = (0{,}35; 0{,}30), \ldots,$ $P_6 = (1; 1)$ on constate que la répartition des salaires $\rho(a_h)$ est très proche de l'équipartition parfaite représentée par la diagonale.



© Dunod. La photocopie non autorisée est un délit.

Exercice 3

La distribution est présentée dans le tableau suivant

Valeurs	10	12	13	14	15	16	17
Effectifs	2	5	5	6	3	2	2
Fréquences relatives f_i	0,08	0,2	0,2	0,24	0,12	0,08	0,08
Fréq. relatives cumulées F_i	0,08	0,28	0,48	0,72	0,84	0,92	1

1.

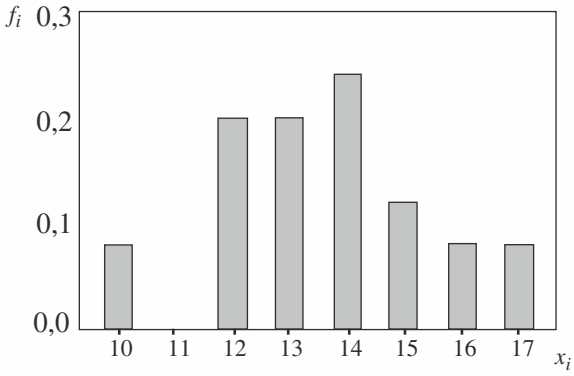


Diagramme en bâtons de la distribution

2. Le *mode* est 14 (valeur de x_i ayant le plus grand effectif).

La *médiane* : on observe que $F_3 = 0{,}48 < 0{,}5$ et $F_4 = 0{,}72 > 0{,}5$ donc la médiane est égale à x_4^* : $m_e = 14$. Le *premier quartile* : on observe que $F_1 = 0{,}08 < 0{,}25$ et $F_2 = 0{,}28 > 0{,}25$ donc le premier quartile q_1 est égal à x_2^* : $q_1 = 12$. Le *troisième quartile* : on observe que $F_4 = 0{,}72 < 0{,}75$ et $F_5 = 0{,}84 > 0{,}75$ donc le troisième quartile q_3 est égal à x_5^* : $q_3 = 15$.

La *moyenne* $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{K} n_i x_i^* = (2 \times 10 + 5 \times 12 + \ldots 2 \times 17)/25 = 13{,}6$.

3. La *variance* $V(x) = \dfrac{1}{n}\sum_{i=1}^{K} n_i (x_i^*)^2 - \overline{x}^2$

$= (2 \times 10^2 + 5 \times 12^2 + \ldots 2 \times 17^2)/25 - 13{,}6^2 = 31{,}493$.

L'*étendue* $= x_7^* - x_1^* = 17 - 10 = 7$;

l'intervalle interquartile $= q_3 - q_1 = \ 15 - 12 = 3$.

CHAPITRE 2

Exercice 1

1. a) La variable expliquée est y le nombre d'erreurs de saisie par page et la variable explicative est le nombre x de jours de stage. On vérifie que

$$\overline{x} = 10,\ \overline{y} = 31,\ \sigma(x) = 3{,}24,\ \sigma(y) = 8{,}17,\ \mathrm{Cov}\,(x,y) = -24{,}74$$

b) La relation affine estimée est : $y = a_0 x + b_0 = -2{,}357x + 54{,}57$ puisque $a_0 = \mathrm{Cov}\,(x,y)/V(x) = -24{,}74/3{,}24^2 = -2{,}357$ et que $b_0 = \overline{y} - a_0\overline{x} = 54{,}57$.

L'intensité de la relation peut être mesurée par $r(x,y) = \mathrm{Cov}\,(x,y)/(\sigma(x) \times (\sigma(y)) = -0{,}9348$ ce qui montre un bon ajustement.

c) $V_E/V(y) = r^2(x,y) = 0{,}874$ et $V_R = (1 - r^2(x,y)) \times V(y) = 8{,}41$

2. a) Un jour de stage supplémentaire se traduit en moyenne par 2,357 erreurs de moins.

b) Le nombre d'erreurs de saisie que l'on peut attendre d'un nouveau stagiaire correspond à la valeur prise par y pour $x = 0$ soit 54,57.

c) Le nombre x de jours de stages supplémentaires nécessaires pour que le nombre d'erreurs de saisie passe à 10 par page est obtenu par résolution de l'équation $y = 10 = -2{,}357x + 54{,}57$ d'où $x = 18{,}9$ soit en fait 19 jours.

Exercice 2

1. *Cf.* tableau ci-dessous. $\overline{t} = 3$, $V(t) = 2$, $\mathrm{Cov}\,(t,y) = 30{,}4$, $\overline{y} = 33$, $V(y) = 497{,}6$.

On a la relation affine $y = a_0 t + b_0$ avec $a_0 = \mathrm{Cov}\,(t,y)/V(t) = 30{,}4/2 = 15{,}2$ et $b_0 = \overline{y} - a_0\overline{t} = -12{,}6$. D'où l'ajustement linéaire : $y = 15{,}2t - 12{,}6$.

Année	1	2	3	4	5
CA (10^6 euros)	8	12	35	40	70

2. $V_E/V(y) = r^2(t,y) = \mathrm{Cov}^2(t,y)/[V(t) \times V(y)] = 0{,}9286$ d'où $V_E = 462{,}08$ et $V_R = (1 - r^2(t,y)) \times V(y) = V(y) - V_E(y) = 35{,}52$.

© Dunod. La photocopie non autorisée est un délit.

3.

y	8,00	12,00	35,00	40,00	70,00	$\overline{y} = 33$
t	1,00	2,00	3,00	4,00	5,00	$\overline{t} = 3$
$\ln(t)$	0,00	0,69	1,10	1,39	1,61	$\overline{\ln(t)} = 0{,}9575$

$y = a\ln(t) + b$ avec $a = \text{Cov}(\ln(t), y)/V(\ln(t)) = 11{,}38/0{,}323 = 35{,}22$
et $b = \overline{y} - a\overline{\ln(t)} = -0{,}721$, $r^2(\ln(t), y) = \text{Cov}^2(\ln(t), y)/(V(\ln(t)) \times V(y)) = 0{,}805$.

$r^2(\ln(t), y)$ étant inférieur à $r^2(t, y)$ on peut considérer que l'ajustement linéaire est préférable.

QCM : ②.

CHAPITRE 3

Exercice 1

Les prix pratiqués en euros et les quantités (obtenues en divisant le chiffre d'affaires par les prix) sont les suivants :

	Prix		Quantité	
	Année 2004	Année 2006	Année 2004	Année 2006
	$p_{n_0}^{(i)}$	$p_n^{(i)}$	$q_{n_0}^{(i)}$	$q_n^{(i)}$
PX1	800	700	25 000	30 000
PX2	720	680	100 000	50 000
VTS	650	650	50 000	75 000
VTR	360	290	150 000	200 000

1. L'indice des prix de Laspeyres $L_p(n/n_0) = \sum_i p_n^{(i)} q_{n_0}^{(i)} / \sum_i p_{n_0}^{(i)} q_{n_0}^{(i)}$ avec $n_0 = 2004$ et $n = 2006$. $L_p(2006/2004) = (700 \times 25\,000 + \ldots + 290 \times 150\,000)/$
$(800 \times 25\,000 + \ldots + 360 \times 150\,000) = 0{,}904$
soit 90,4 pour un indice base 100.

L'indice des prix de Paasche $P_p(n/n_0) = \sum_i p_n^{(i)} q_n^{(i)} / \sum_i p_{n_0}^{(i)} q_n^{(i)}$ avec $n_0 = 2004$ et $n = 2006$:
$P_p(2006/2004) = (700 \times 30\,000 + \ldots + 290 \times 200\,000)/$
$(800 \times 30\,000 + \ldots + 360 \times 200\,000) = 0{,}895$
soit 89,5 pour un indice base 100.

2. L'indice de volume de Laspeyres $L_q(n/n_0) = \sum_i q_n^{(i)} p_{n_0}^{(i)} / \sum_i q_{n_0}^{(i)} p_{n_0}^{(i)}$ avec $n_0 = 2004$ et $n = 2006$:

$$L_q(2006/2004) = (30\,000 \times 800 + \ldots + 20\,000 \times 360)/ (25\,000 \times 800 + \ldots + 150\,000 \times 360) = 1{,}0126$$

soit 101,26 pour un indice base 100.

L'indice de volume de Paasche $P_q(n/n_0) = \sum_i q_n^{(i)} p_n^{(i)} / \sum_i q_{n_0}^{(i)} p_n^{(i)}$ avec $n_0 = 2004$ et $n = 2006$:

$$P_q(2006/2004) = (30\,000 \times 700 + \ldots + 200\,000 \times 290)/ (25\,000 \times 700 + \ldots + 150\,000 \times 290) = 1{,}0015\,.$$

Exercice 2

1. Évolution du cours du dollar \$, base 100 en Juillet 2006 :

	janv-06	juil-06	oct-06	avr-07
$\mathbf{G_h}$ **valeur euro en \$**	1,1821	1,2792	1,2737	1,3369
base 100 en juil-06	92,41	100,00	99,57	104,44

Évolution base 100 juil-06 : $92{,}41 = (1{,}1821/1{,}2792) \times 100$; 100 ; 99,57 ; 104,44

– Le taux de variation global $= (1{,}3369/1{,}1821) - 1 = 0{,}131$ soit 13,1 %

– Le taux de variation moyen mensuel $\overline{\tau}_m$ du taux de change sur l'intervalle de 15 mois séparant le premier cours $\mathbf{G}_1$ du 16-ième $\mathbf{G}_{16}$ est défini par la relation :

$$\mathbf{G}_{16} = \mathbf{G}_1(1 + \overline{\tau}_m)^{15} \text{ c'est-à-dire } 1{,}3369 = 1{,}1821(1 + \overline{\tau}_m)^{15}\,.$$

Donc $\overline{\tau}_m = (1{,}3369/1{,}1821)^{1/15} - 1 = -0{,}00823$, soit $-0{,}823$ %

– Le taux de variation moyen trimestriel $\overline{\tau}_t$ sur les cinq trimestres de la période (t_1, t_6) est défini par la relation : $\mathbf{G}_6 = \mathbf{G}_1(1 + \overline{\tau}_t)^5$. Donc $1{,}3369 = 1{,}1821(1 + \overline{\tau}_t)^5$ d'où l'on déduit $\overline{\tau}_t = 0{,}0249$ soit 2,49 %. On peut également remarquer que $\overline{\tau}_t = (1 + \overline{\tau}_m)^3 - 1$.

2. Le taux de variation moyen mensuel $\overline{\tau}^*$ du cours du dollar entre juillet 2006 et octobre 2006 concerne l'intervalle de 3 mois (t_1, t_4). Il est donc défini par la relation $\mathbf{G}_4 = (1 + \overline{\tau}^*)^3 \mathbf{G}_1$ c'est-à-dire $1{,}2737 = 1{,}2792(1 + \overline{\tau}^*)^3$. Donc $\overline{\tau}^* = -0{,}001436$.

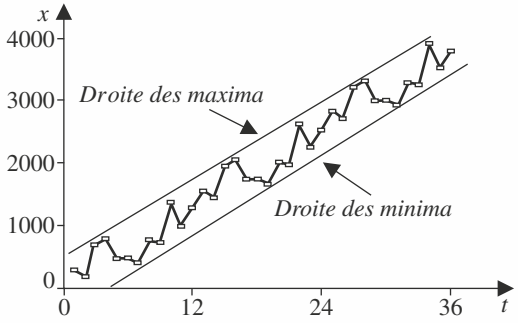
QCM 1 : ④ ; **QCM 2** : ⑤ car $t_m = 10$ %.

© Dunod. La photocopie non autorisée est un délit.

CHAPITRE 4

Exercice 1

1. Joignant les minima et les maxima de la série chronologique, on observe que les deux droites sont parallèles (voir ci-dessous). Il convient donc de retenir un modèle additif de type : $x_h = T_h + S_{r(h)} + A_h$ avec x_h désignant la *h-ème* observation, T_h le trend, $S_{r(h)}$ le facteur saisonnier et A_h le facteur accidentel.



2. On a 36 valeurs x_h : $h = 1,2,\ldots,36$. Le mouvement saisonnier a un nombre pair $\nu = 12$ $(= 2 \times p$ avec $p = 6)$ de périodes dans l'année.

À la $(p + 1)$-ème période est attribuée pour trend la valeur $T_{p+1} = (0{,}5x_1 + x_2 + \ldots + x_\nu + 0{,}5x_{\nu+1})/\nu$

soit $T_7 = (0{,}5 \times 300 + 195 + \ldots + 1280 + 0{,}5(1560)/12 = 755{,}08$;

de même $T_8 = (0{,}5 \times 195 + 685 + \ldots 1560 + 0{,}5 \times \ \ 1455)/12 = 860{,}08$, etc.

Tableau de T_h

an \ j	1	2	3	4	5	6	7	8	9	10	11	12
N-2							755,08	860,08	965,08	1070,08	1175,08	1280,08
N-1	1385,08	1490,04	1595,21	1700,42	1805,42	1910,42	2015,42	2120,42	2225,38	2330,33	2435,33	2540,33
N	2645,33	2750,42	2855,29	2960,08	3065,08	3170,08						

On calcule $S'_h = (x_h - T_h)$ et on en déduit les 12 coefficients saisonniers S_j . Ainsi $S'_7 = (x_7 - T_7) = 405 - 755{,}08 = -350{,}08, S'_8 = (x_8 - T_8) = 756 - 860{,}08$ $= -104{,}08, S'_{13} = (x_{13} - T_{13}) = 1560 - 1385{,}08 = 174{,}92.$

Tableau de calcul de $S'_h = (x_h - T_h)$ et de S_j

an \ j	1	2	3	4	5	6	7	8	9	10	11	12
N–2							– 350,08	– 104,08	– 245,08	279,92	– 175,08	– 0,08
N–1	174,92	– 35,04	349,79	349,58	– 70,42	– 175,42	– 350,42	– 105,42	– 240,38	279,67	– 175,33	– 0,33
N	174,67	– 35,42	348,71	349,92	– 70,08	– 175,08						
$\overline{S}'_j$	174,79	– 35,23	349,25	349,75	– 70,25	– 175,25	– 350,25	– 104,75	– 242,73	279,79	– 175,21	– 0,21
S_j	174,82	– 35,20	349,27	349,77	– 70,23	– 175,23	– 350,23	– 104,73	– 242,70	279,82	– 175,18	– 0,18

La moyenne par saison des écarts $S'_h = (x_h - T_h)$ est notée $\overline{S}'_j$.

Ainsi $\overline{S}'_1 = (174{,}92 + 174{,}67)/2 \quad = 174{,}79, \ldots, \overline{S}'_{12} = (-0{,}08 - 0{,}33)/2 = -0{,}21$.

Calculant la somme de $\sum_{j=1}^{12} \overline{S}'_j = (174{,}79 - 35{,}23 + \ldots - 0{,}21) = -0{,}29$ on déduit la moyenne $\overline{S} = -0{,}29/12 \cong -0{,}024$ ainsi que la valeur des coefficients saisonniers corrigés $S_j = \overline{S}'_j - \overline{S}$: $S_1 = 174{,}79 + 0{,}024 = 174{,}82, \ldots, S_{12} = -0{,}21 - (-0{,}024) = -0{,}18$.

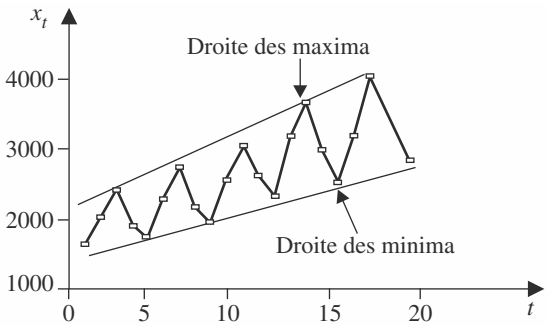
La série désaisonnalisée est obtenue en soustrayant aux valeurs x_h le coefficient saisonnier $S_{r(h)}$.

Tableau de la série désaisonnalisée $(x_h - S_j)$

an \ j	1	2	3	4	5	6	7	8	9	10	11	12
N – 2	125,1	230,0	335,2	440,4	545,4	650,4	755,4	861,4	960,4	1070,3	1175,3	1280,3
N – 1	1385,3	1490,4	1596,3	1700,1	1805,1	1910,1	1665,0	2015,0	1985,0	2610,0	2260,0	2540,0
N	2645,2	2750,2	2854,8	2960,3	3065,3	3170,3	3275,3	3381,8	3482,7	3590,2	3695,2	3800,2

Exercice 2

1. On constate que l'écart-type augmente d'année en année, aussi peut-on penser à un modèle multiplicatif. De même, lorsque l'on joint les minima et maxima de la série, on observe également que les deux droites s'écartent. La composante saisonnière augmentant avec le trend, il est donc préférable de recourir à un schéma multiplicatif.



© Dunod. La photocopie non autorisée est un délit.

2. Le modèle multiplicatif : $x_h = T_h \times S_{r(h)} \times (1 + A_h)$ (où $j = r(h)$ est le reste de la division de h par 4 si h n'est pas divisible par 4, $j = 4$ sinon) peut être reformulé en modèle additif. Si on utilise les logarithmes népériens et que l'on pose $x_h^* = \ln(x_h), T_h^* = \ln(T_h), S_{r(h)}^* = \ln(S_{r(h)}), A_h^* = \ln(1 + A_h)$, on obtient le schéma additif suivant : $x_h^* = T_h^* + S_{r(h)}^* + A_h^*$.

Tableau de $x_h^* = \ln(x_h)$ et du trend T_h^*

	x_h^*				T_h^*			
$r(h)$	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
2003	7,409	7,620	7,796	7,552			7,598	7,617
2004	7,441	7,737	7,918	7,691	7,647	7,679	7,711	7,738
2005	7,552	7,850	8,021	7,873	7,765	7,801	7,851	7,904
2006	7,764	8,067	8,204	8,004	7,954	7,993	8,020	8,030
2007	7,844	8,070	8,300	7,977	8,042	8,051		

Le mouvement saisonnier a un nombre pair ν de périodes dans l'année : $\nu = 4 = 2 \times p$ où $p = 2$.

À la $(p + 1)$-ème période est attribuée au trend la valeur
$T_{p+1}^* = (0{,}5x_1^* + x_2^* + \ldots + x_\nu^* + 0{,}5x_{\nu+1}^*)/\nu$: soit

$T_3^* = (0{,}5 \times 7{,}409 + 7{,}620 + 7{,}796 + 7{,}552 + 0{,}5 \times 7{,}441)/4 = 7{,}598$;

$T_4^* = (0{,}5x_2^* + x_3^* + x_4^* + x_5^* + 0{,}5x_6^*)/4 = 7{,}617$; etc.

On calcule les écarts $S'_h = (x_h^* - T_h^*)$: $S'_3 = (7{,}796 - 7{,}598) = 0{,}198$;

$S'_4 = 7{,}552 - 7{,}617 = -0{,}065$; $S'_5 = -0{,}206$; $\ldots$; $S'_{18} = 0{,}019$ et en résultent les coefficients saisonniers :

Écarts saisonniers $S'_h = (x_h^* - T_h^*)$ et coefficients saisonniers

Année **Saison** $j = r(h)$	**1**	**2**	**3**	**4**
2003			0,198	– 0,065
2004	– 0,206	0,058	0,207	– 0,048
2005	– 0,214	0,049	0,170	– 0,031
2006	– 0,190	0,073	0,184	– 0,026
2007	– 0,199	0,019		
$\overline{S}_j^*$ **moyenne par trimestre des écarts saisonniers**	– 0,202	0,050	0,190	– 0,042
$\sum_{j=1}^{4} \overline{S}_j^* = -0{,}004$ d'où $\overline{S} = \frac{1}{4} \sum_{j=1}^{4} \overline{S}_j^* \cong -0{,}001$				
Coefficients saisonniers corrigés $S_j^* = \overline{S}_j^* - \overline{S}$	– 0,201	0,051	0,191	– 0,041

On détermine alors la série désaisonnalisée $z_h^* = x_h^* - S_{r(h)}^*$. Ainsi, $z_1^* = x_1^* - S_1^* = 7{,}409 - (-0{,}201) = 7{,}610$, $z_{11}^* = 8{,}021 - 0.191 = 7{,}830$ ...

Tableau de la série désaisonnalisée z_h^*

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2003	7,610	7,569	7,604	7,593
2004	7,642	7,686	7,727	7,732
2005	7,753	7,799	7,830	7,914
2006	7,965	8,016	8,013	8,046
2007	8,045	8,019	8,108	8,018

3. On cherche par la méthode des moindres carrés ordinaires la droite d'équation $z_h^* = a_0 h + b_0$ avec $h = (i-1) \times \nu + j$, i désignant l'année ($i = 1$ pour 2003, ...,$i = 5$ pour 2007), j désignant le trimestre ($j = 1$ pour le premier trimestre, etc.). On a $a_0 = \text{Cov}(z_t^*, h)/V(h) = 1{,}08/33{,}25 = 0{,}0325$ et $b_0 = \overline{z}_h^* - a_0\overline{h} = 7{,}834 + 0{,}0325 \times 10{,}5 = 7{,}4933$ d'où la valeur estimée $z'_h = 0{,}0\,325 \times h + 7{,}4\,933$ puis $T_h = e^{z'_h}$.

Exercice 3

1. On détermine tout d'abord le trend sous la forme d'une fonction du temps $x_h = a_0 \times h + b_0$ avec $h = 4 \times (i-1) + j$. On trouve selon la méthode des moindres carrés ordinaires $b_0 = 135{,}1$ et $a_0 = 7{,}1647$. Les valeurs T_h estimées de x_h par cet ajustement correspondent au trend.

h	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_h	152	175	126	140	185	207	151	172	218	238	180	205	249	280	215	243
T_h	142,3	149,4	156,6	163,8	170,9	178,1	185,3	192,4	199,6	206,7	213,9	221,1	228,2	235,4	242,6	249,7

Réécrivant cette série dans un tableau à double entrée où figurent en ligne les années et en colonne les trimestres, on obtient :

Année i \ Trimestre j	1	2	3	4
$N-3$ soit $i=1$	142,3	149,4	156,6	163,8
$N-2$ soit $i=2$	170,9	178,1	185,3	192,4
$N-1$ soit $i=3$	199,6	206,7	213,9	221,1
N soit $i=4$	228,2	235,4	242,6	249,7

© Dunod. La photocopie non autorisée est un délit.

On détermine ensuite les rapports S_{ij} entre valeurs observées x_{ij} et valeurs ajustées T_{ij} (soit en base 100) : $S_{ij} = 100 \times (x_{ij}/T_{ij})$.
Ainsi $S_{11} = 100 \times (152/142{,}3) = 106{,}8$; etc.

Année i	Trimestres j 1	2	3	4
	S_{ij}			
$N-3$ soit $i=1$	106,8	117,1	80,5	85,5
$N-2$ soit $i=2$	108,2	116,2	81,5	89,4
$N-1$ soit $i=3$	109,2	115,1	84,1	92,7
N soit $i=4$	109,1	118,9	88,6	97,3
Moyenne des rapports	108,35	116,85	83,69	91,23
S_j	108,32	116,82	83,66	91,20

Puis, on calcule la moyenne de ces rapports par saison, soit $\overline{S}_{\bullet j} = \frac{1}{k}\sum_{i=1}^{k} S_{ij}$. Ainsi $\overline{S}_{\bullet j} = (106{,}8 + 108{,}2 + 109{,}2 + 109{,}1)/4 = 108{,}35$. Afin de respecter la condition selon laquelle la somme des coefficients saisonniers est égale à 100, on calcule la moyenne $\overline{S} = \sum_{i=1}^{\nu} = 100{,}03$ et on attribue finalement à S_j la valeur $S_j = \overline{S}_{\bullet j} - (\overline{S} - 100)$. Ainsi $S_1 = 108{,}35 - (100{,}03 - 100) = 108{,}32$; etc.

QCM 1 : ③ ; **QCM 2** : ④.

CHAPITRE 6

Exercice 1

1. Le premier cadre a quatre possibilités, le deuxième quatre possibilités, etc. Soit en tout 4^{20} possibilités.

2. Le nombre de combinaisons distinctes possibles pour un cadre est égal à C_4^2 soit pour l'ensemble des cadres : $(C_4^2)^{20}$.

Exercice 2

1. Le nombre n de répartitions est $\frac{20!}{8!7!5!}$. En effet, le nombre de groupes distincts que l'on peut attribuer à X est C_{20}^8 ; il reste à attribuer 7 clients à Y parmi les 12 restants, cette attribution pouvant se faire de façons distinctes. Donc $n = C_{20}^8 \times C_{12}^7 = \frac{20!}{8!7!5!}$.

2. Le nombre de répartitions possibles des 11 hommes clients est $\frac{11!}{5!4!2!}$, des 9 femmes clientes est $\frac{9!}{3!3!3!}$. La solution est donc le produit des deux termes.

Exercice 3

1. Il reste à répartir 80 euros de différentes façons. Le nombre de quadruplets d'entiers (n_1,n_2,n_3,n_4) tels que $n_i \geqslant 0\ \forall i = 1,2,3,4$ et $n_1+n_2+n_3+n_4 = 80$ est égal à $C_{n+p-1}^{p-1} = C_{80+4-1}^{4-1} = C_{83}^{3} = 91881$.

Exercice 4

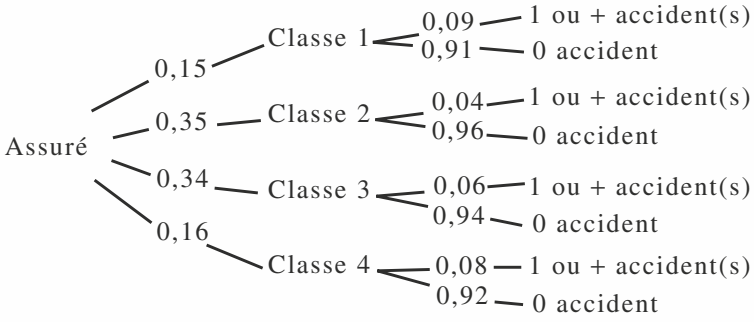
Il y a A_5^3 façons d'attribuer un chauffeur à chaque groupe. On « ajoute » un mannequin aux autres passagers et on les répartit en 3 groupes de $12!/(4! \times 4! \times 4!)$ façons distinctes. D'où le nombre de façons distinctes : $(A_5^3 \times 12!)/(4! \times 4! \times 4!)$.

QCM 1 : ② **QCM 2 :** ① **; QCM 3 :** ④

CHAPITRE 7

Exercice 1

1. On fait un graphe sur lequel apparaissent la probabilité qu'un assuré tiré au hasard appartienne à chacune des classes et les probabilités correspondantes de déclarer au moins un accident au cours de l'année.



Soit l'événement **A** « avoir un accident au moins »

$p(\mathbf{A}) = (0{,}15 \times 0{,}09) + (0{,}35 \times 0{,}04) + (0{,}34 \times 0{,}06) + (0{,}16 \times 0{,}08) = 0{,}0\,607$

© Dunod. La photocopie non autorisée est un délit.

2. Soit l'événement **B** « être âgé de + de 45 ans c'est-à-dire appartenir à la classe 3 ou 4 ».

$p(\mathbf{A}/\mathbf{B}) = p(\mathbf{A} \cap \mathbf{B})/p(\mathbf{B}) = [(0{,}34 \times 0{,}06) + (0{,}16 \times 0{,}08)]/(0{,}34 + 0{,}16) = 0{,}0664$

3. Soit l'événement **C** « être âgé de moins de 26 ans c'est-à-dire appartenir à la classe 1 ».

$$p(\mathbf{C}/\mathbf{A}) = p(\mathbf{C} \cap \mathbf{A})/p(\mathbf{A}) = (0{,}15 \times 0{,}09)/(0{,}0607) = 0{,}2224\,.$$

4. Soit les événements $\overline{A}$ « ne pas avoir déclaré d'accident » et **B** « être âgé de + de 45 ans, c'est-à-dire appartenir à la classe 3 ou 4 ».On cherche $p(\mathbf{B}/\overline{A})$ où $\overline{A}$ est l'événement contraire de **A**. On a : $p(\overline{A}) = 1 - p(\mathbf{A}) = 0{,}9393$. Donc $p(\mathbf{B}/\overline{A}) = p(\mathbf{B} \cap \overline{A})/p(\overline{A})$ $= [(0{,}34 \times 0{,}94) + (0{,}16 \times 0{,}92)]/0{,}9393 = 0{,}4969$.

Exercice 2

G désigne l'événement : « l'entreprise a lors de l'octroi du crédit un ratio **Z** supérieur à c », $\overline{G}$ désigne l'événement contraire. **D** désigne l'événement : « l'entreprise se révèle défaillante », **S** désigne l'événement : « l'entreprise se révèle saine ». Le référentiel est $E = \{(\mathbf{D},\overline{G}); (\mathbf{D},\mathbf{G}); (\mathbf{S}\overline{G},); (\mathbf{S},\mathbf{G})\}$

Une entreprise ayant un ratio **Z** supérieur à la valeur critique c (événement **G**), donc *a priori* considérée comme saine peut se révéler défaillante (événement **D**). Réciproquement, une entreprise considérée comme à risque (évènement $\overline{G}$) car son ratio **Z** est inférieur à la valeur critique c peut se révéler saine. Soit $p(\mathbf{D}/\mathbf{G})$ la probabilité qu'une entreprise soit défaillante alors que d'après le calcul du ratio elle aurait dû être saine.

$$p(\mathbf{D}/\mathbf{G}) = \frac{p(\mathbf{D} \cap \mathbf{G})}{p(\mathbf{G})} = \frac{p(\mathbf{D}) \times p(\mathbf{G}/\mathbf{D})}{p(\mathbf{D}) \times p(\mathbf{G}/\mathbf{D}) + p(\mathbf{S}) \times p(\mathbf{G}/\mathbf{S})}$$

$$= \frac{(0{,}05 \times 20)}{(0{,}05 \times 0{,}20) + (0{,}95 \times 0{,}90)} = \frac{0{,}01}{0{,}865} = 1{,}15\ \%\,.$$

On a 1,15 % de chances de se tromper lorsque à partir du ratio on qualifie une entreprise de saine.

Soit $p(\mathbf{S}/\overline{G})$ la probabilité qu' une entreprise soit saine alors que d'après le calcul du ratio elle était considérée comme à risque.

$$p(\mathbf{S}/\overline{G}) = \frac{p(\mathbf{S} \cap \overline{G})}{p(\overline{G})} = \frac{p(\mathbf{S}) \times p(\overline{G}/\mathbf{S})}{p(\overline{G})} \frac{(0{,}95 \times 0{,}10)}{(0{,}05 \times 0{,}8) + (0{,}10 \times 0{,}95)} = 70{,}3\ \%$$

On a 70 % de chances de se tromper lorsque l'on classe comme défaillante une entreprise dont le ratio est inférieur à c.

CHAPITRE 8

Exercice 1

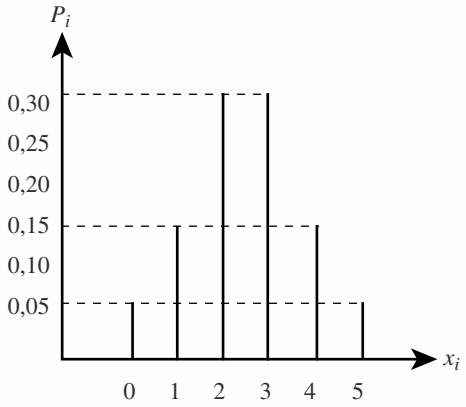
1. La distribution de la *v.a X* peut être présentée sous forme de tableau :

Demande x_i	0	1	2	3	4	5
Probabilité p_i	0,05	0,15	0,30	0,30	0,15	0,05
Probabilité cumulée $\mathbf{P}(X \leqslant x_i)$	0,05	0,20	0,50	0,80	0,95	1

De ce tableau se déduit l'expression de la fonction de répartition $F(t) = P(X \leqslant t)$:

$F(t) = 0$	$F(t) = 0{,}05$	$F(t) = 0{,}20$	$F(t) = 0{,}50$	$F(t) = 0{,}80$	$F(t) = 0{,}95$	$F(t) = 1$
$\forall t < 0$	si $0 \leqslant t < 1$	si $1 \leqslant t < 2$	si $2 \leqslant t < 3$	si $3 \leqslant t < 4$	si $4 \leqslant t < 5$	$\forall t \geqslant 5$

La distribution de probabilité est représentée ci-dessous



Distribution de probabilité : diagramme en bâtons

2. $E(X) = \sum_{i=1}^{6} p_i x_i = 0 \times 0{,}05 + 1 \times 0{,}15 + \ldots + 5 \times 0{,}05 = 2{,}5$

$V(X) = E(X^2) - (E(X))^2 = (0^2 \times 0{,}05 + 1^2 \times 0{,}15 + \ldots + 5^2 \times 0{,}05) - 2{,}5^2 = 1{,}45$

Il existe une valeur $x_{io} = 2$ de Δ_X telle que $\sum_{i=1}^{i_o} p(x_i) = 0{,}5$. Donc $(x_{i_0} + x_{i_0+1})/2 = (2+3)/2 = 2{,}5$ est la valeur médiane de la distribution

3. Il y a rupture de stock lorsque la demande est supérieure à 2 perceuses :

$$P(X > 2) = 1 - P(X \leqslant 2) = 1 - 0{,}5 = 0{,}5$$

© Dunod. La photocopie non autorisée est un délit.

4. Il faut déterminer le plus petit x_i tel que $P(X \leqslant x_i) \geqslant 0{,}9$.

Ici on observe que $x_i = 4$. Il faut donc avoir au moins 4 perceuses en stock en début de journée.

Exercice 2

1. $\Delta_X = [0,3]$ et $f(t) = k$ (constante) $\forall t \in \Delta_X$, donc $1 = \int_0^3 k\mathrm{d}x = [kx]_0^3 = 3k$ et par suite $f(t) = 1/3$. $F(t)$ désignant la fonction de répartition de X, on a :

$F(t) = 0 \quad \forall t \leqslant 0$; $F(t) = \int_0^t f(x)\mathrm{d}x = \int_0^t (1/3)\mathrm{d}x = [(1/3) \times x]_0^t = t/3$

si $0 < t < 3$; $F(t) = 1 \; \forall \, t \geqslant 3$; $E(X) = \int_0^3 xf(x)\mathrm{d}x = \int_0^3 \frac{1}{3}x\mathrm{d}x = [(1/3) \times x^2/2]_0^3$

$= 3/2$. $E(X^2) = \int_0^3 x^2 f(x)\mathrm{d}x = \int_0^3 \frac{1}{3}x^2\mathrm{d}x = 3$.

$V(X) = E(X^2) - [E(X)]^2 = 3 - (3/2)^2 = 0{,}75$. $F(t) = t/3$ $\forall t$ vérifiant $0 < t < 3$, donc $F(2) = 2/3$.

2. On sait que $\Delta_X = [0,3]$. Pour déterminer le domaine des valeurs de $Y = 2X + 3$ il suffit de remarquer que $X = 0$ implique $Y = 3$ et $X = 3$ implique $Y = 9$: $\Delta_Y = [3, 9]$.

Notons $G(t) = P(Y \leqslant t)$ la fonction de répartition de Y.

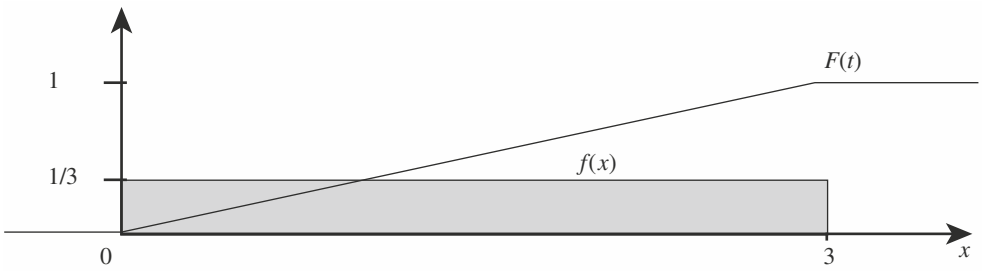
Pour $t < 3$ [resp. $t > 9$] on a $G(t) = 0$ [resp.1].

Pour $t \in \Delta_Y, G(t)$ se déduit de l'expression de la fonction de répartition F de X :

$$G(t) = P(Y \leqslant t) = P(2X + 3 \leqslant t) = P\left(X \leqslant \frac{t-3}{2}\right) = F\left(\frac{t-3}{2}\right).$$

Or pour $3 \leqslant t \leqslant 9$ on a $0 \leqslant \frac{t-3}{2} \leqslant 3$, donc $F\left(\frac{t-3}{2}\right) = \frac{1}{3} \times \frac{t-3}{2}$ et par suite $G(t) = \frac{t-3}{6}$.

$E(Y) = E(2X + 3) = 2E(X) + 3 = 6$; $V(Y) = V(2X + 3) = V(2X) = 2^2 V(X) = 3$.



Représentation de la densité de probabilité f(x) et de la fonction de répartition F(t)

QCM 1 : ③ ; QCM 2 : ③ ; QCM : ④

CHAPITRE 9

Exercice 1

L'erreur algébrique X du solde d'un compte client suit sensiblement la loi $\mathcal{N}(0; 5^2)$.

Donc $(X - m)/\sigma \cong \mathcal{N}_{0;1}$ où $m = 0$ et $\sigma = 5$ par suite $X/5 = \mathcal{N}_{0;1}$. Pour $t \geqslant 0$ on lit page 374 la valeur de $P(\mathcal{N}_{0;1}(t) = \Phi(t)$.

a) $P(X < 10) = P(X/5 < 2) = P(\mathcal{N}_{0;1} < 2) = \Phi(2) = 0{,}977$;

b) $P(X < 5) = \Phi(1) = 0{,}8413$;

c) $P(X < -5) = P(X/5 < -1) = P(\mathcal{N}_{0;1} < -1) = \Phi(-1) = 1 - \Phi(1) = 0{,}1357$;

d) $P(-5 \leqslant X \leqslant 5) = P(-1 < X/5 < 1) = P(-1 < \mathcal{N}_{0;1} < 1) = \Phi(1) - \Phi(-1)$
$= 2\Phi(1) - 1 = 0{,}06826$

Exercice 2

1. La demande aléatoire X du produit suit la loi normale de moyenne $m = 5\,000$ et d'écart type $\sigma = 1\,000$. Donc $P(X > 5\,500) = 1 - P(X \leqslant 5\,500)$
$= 1 - \Phi\left(\dfrac{5\,500 - m}{\sigma}\right) = 1 - \Phi(0{,}5) = 0{,}3085$

Conclusion : il y a 30,85 chances sur 100 pour que la demande du produit soit supérieure à 5 500.

2. t^* désigne le niveau de stock devant être maintenu pour que la demande soit satisfaite dans 90 % des cas. On a donc $0{,}90 = P(X \leqslant t^*) = \Phi\left(\dfrac{t^* - 5\,000}{1\,000}\right)$. D'après la table statistique de la loi normale centrée réduite, on a : $\dfrac{t^* - 5\,000}{1\,000} = 1{,}28$ et donc $t^* = 6\,280$.

Conclusion : pour satisfaire la demande dans 90 % des cas, le niveau du stock doit être maintenu à raison de 6 280 produits au début de chaque période.

Exercice 3

5 % des pièces sorties d'un atelier d'usinage sont défectueuses.

1. X désignant le nombre aléatoire de pièces défectueuses : $X = B_{15;0,05}$

a) $P(X = 1) = C_{15}^1 \times 0{,}05^1 \times 0{,}95^{14} = 15 \times 0{,}05 \times 0{,}95^{14} = 0{,}3658$

b) $P(X \geqslant 2) = 1 - \{P(X = 0) + P(X = 1)\} = 1 - P(X \leqslant 1) = 1 - P(B_{15;0,05} \leqslant 1)$
$= 1 - 0{,}8290 = 0{,}171.$

2. Pour un lot de 1 000 pièces, la condition $np(1 - p) = 1\,000 \times 0{,}05 \times 0{,}95$
$= 47{,}5 > 10$ étant satisfaite, on peut approximer la loi binomiale par une loi normale :

© Dunod. La photocopie non autorisée est un délit.

$(X - np)/\left(\sqrt{np(1-p)}\right) \cong \mathcal{N}_{0;1}$ et donc $(X - 50)/6{,}89 \cong \mathcal{N}_{0;1}$ à condition d'utiliser la correction de continuité.

$P(X \leqslant 30) = P(X < 30{,}5) = P[(X - 50)/6{,}89) < (30{,}5 - 50)/6{,}89)] \cong$
$P(\mathcal{N}_{0;1} \leqslant -2{,}83) = 1 - \Phi(2{,}83) = 0{,}0023$

Exercice 4

À la suite de l'expérience élémentaire de l'appel téléphonique, on a soit un échec avec une probabilité : $1 - p = 1 - 0{,}01 = 0{,}99$ soit un succès avec une probabilité $p = 0{,}01$. On renouvelle alors cette expérience autant de fois que nécessaire pour être sûr à 95 % d'avoir un succès. Le nombre X d'échecs qui précède l'obtention de la première vente suit la loi géométrique $\mathcal{G}(0{,}01)$: $X \sim> \mathcal{G}(0; 0{,}1)$. On cherche le plus petit entier h tel que $P(X \leqslant h - 1) \geqslant 0{,}95$ soit $P(X = 0) + P(X = 1) + \ldots.\ P(X \leqslant h - 1) = 0{,}01 \times 0{,}99^{\circ}$ $+0{,}01 \times 0{,}99^{1} + \ldots.0{,}01 \times 0{,}99^{h-1} \geqslant 0{,}95$.

Or $\sum_{i=0}^{h-1}(0{,}01 \times 0{,}99^{i}) = 1 - 0{,}99^{h}$ puisque qu'il s'agit d'une suite géométrique de raison 0,99, donc on cherche h tel que $1 - 0{,}99^{h} \geqslant 0{,}95$ ou de façon équivalente $h \ln(0{,}99) \leqslant \ln(0{,}05)$. Aussi $h = 299$. Il faut donc au moins 299 appels pour obtenir un succès avec une probabilité de 95 %.

Exercice 5

X suit une loi de Poisson $\mathcal{P}(17)$.

a) On cherche le plus petit entier n_0 tel que $\mathcal{P}(X \leqslant n_0) \geqslant 0{,}96$. Pour $\lambda > 10$ on peut utiliser l'approximation normale $2\left(\sqrt{P_\lambda} - \sqrt{\lambda}\right) \cong \mathcal{N}_{0;1}$ après avoir remarqué que $\mathrm{P}(P_\lambda \leqslant n_0) \cong \Phi\left[2\left(\sqrt{n_0 + 0{,}5} - \sqrt{\lambda}\right)\right] = 0{,}96$ par lecture de table de la loi normale $2\left(\sqrt{n_0 + 0{,}5} - \sqrt{17}\right) = 1{,}76$ donc $n_0 = 25$.

b) Le nombre N d'absences dans cette équipe suit une loi Binomiale $\mathcal{B}(25,\ 0{,}04)$ qui peut être approchée par une loi de Poisson de paramètre $np = 25 \times 0{,}04 = 1$ (car $np < 3$ et n considéré comme grand, bien qu'inférieur au seuil usuel de 50) : $N \cong P_1$

Il y aura interruption de la production lorsque $X + N > 25$. Or $X + N \cong P_{17} + P_1 = P_{17+1}$ (car les variables sont indépendantes), donc :

$\mathcal{P}(X + N > 25) \cong \mathcal{P}(P_{18} > 25) = 1 - \mathcal{P}(P_{18} \leqslant 25) \cong 1 - 0{,}9474 = 0{,}0526$. En effet pour $\lambda > 10$, $\mathcal{P}(P_\lambda \leqslant h) \cong \Phi\left[2\left(\sqrt{h + 0{,}5} - \sqrt{\lambda}\right)\right]$ pour $h = 25$ soit $\Phi\left[2\left(\sqrt{25 + 0{,}5} - \sqrt{18}\right)\right] = \Phi(1{,}614) = 0{,}9474.$

CHAPITRE 10

Exercice 1

1. La population statistique ***P*** est l'ensemble des véhicules des assurés ayant une cylindrée supérieure à 1 200 cm^3

2. La moyenne aléatoire $\overline{D}$ de l'échantillon prend pour valeur $\overline{d} = 15$ et l'écart-type standard *S,* la valeur $s = 5$.

La statistique utilisée : $\dfrac{\overline{D} - m_p}{S/\sqrt{n}} \cong \mathcal{N}_{0;1}$ avec $n = 900$. L'intervalle de confiance étant de 95 %, on a $0{,}95 = \mathrm{P}(-1{,}96 \leqslant \mathcal{N}_{0;1} \leqslant 1{,}96)$, donc $\mathrm{P}\left(-1{,}96 \leqslant \dfrac{\overline{D} - m_p}{S/30} \leqslant 1{,}96\right) \cong 0{,}95$ ou de façon équivalente $\mathrm{P}(\overline{D} - 1{,}96 \times S/30 \leqslant m_p \leqslant \overline{D} + 1{,}96 \times S/30) \cong 0{,}95$.

Réalisation de l'intervalle de confiance : $14{,}67 \leqslant m_p \leqslant 15{,}33$

3. Pour trouver un intervalle de confiance de σ_P on utilise l'approximation « $\dfrac{S^2 - \sigma_P^2}{\sqrt{(\hat{\mu}_4 - S^4)}/n} \cong \mathcal{N}_{0;1}$ »

Le risque d'erreur α étant de 10 %, on cherche $a_{\alpha/2}$ et $b_{\alpha/2}$ tels que $\mathrm{P}(\mathcal{N}_{0;1} \leqslant a_{\alpha/2}) = 0{,}05$ et $\mathrm{P}(\mathcal{N}_{0;1} \geqslant b_{\alpha/2}\} = 0{,}05$. On obtient $a_{\alpha/2} = -1{,}64$ et $b_{\alpha/2} = 1{,}64$.

Donc $0{,}90 = \mathrm{P}(-1{,}64 \leqslant \mathcal{N}_{0;1} \leqslant 1{,}64) \cong \mathrm{P}(S^2 - 1{,}64 \times \sqrt{(\hat{\mu}_4 - S^4)}/30 \leqslant \sigma_p^2 \leqslant S^2 + 1{,}64\sqrt{(\hat{\mu}_4 - S^4)}/30) \cong 0{,}90$.

Réalisation de l'intervalle de confiance : $5^2 - 1{,}64 \times \sqrt{(800 - 5^4)}/30 \leqslant \sigma_p^2 \leqslant 5^2 + 1{,}64 \times \sqrt{(800 - 5^4)}/30$ et donc $4{,}93 \leqslant \sigma_p \leqslant 5{,}07$.

Il y a 90 % de chance que σ_p soit compris entre 4,93 et 5,07.

Exercice 2

a) La distribution étant normale, pour obtenir un intervalle de confiance de m_P on utilise la propriété : $\dfrac{\overline{X} - m_P}{S/\sqrt{n}} = \boldsymbol{t}_{n-1}$ (variable de Student) où $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ et $S = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$.

Ici $n = 10$, $\overline{X}$ prend pour valeur $\overline{x} = (20{,}1 + 19{,}9 + \ldots + 19{,}9)/10 = 20$ et l'écart standard S prend pour la valeur observée $s = \sqrt{\dfrac{1}{10-1}\sum_{i=1}^{10}(x_i - 20)^2} = 0{,}1826$

© Dunod. La photocopie non autorisée est un délit.

Se fixant un risque d'erreur $\alpha = 0{,}05$ on cherche $a_{\alpha/2}$ et $b_{\alpha/2}$ tels que $P(t_9 < a_{\alpha/2}) = \alpha/2 = 0{,}025$ et $P(t_9 > b_{\alpha/2}) = \alpha/2 = 0{,}025$. Par lecture de la table de la loi de Student, on lit $a_{\alpha/2} = -2{,}262$ et $b_{\alpha/2} = 2{,}262$. Donc $0{,}95 = P(-2{,}262 \leqslant t_9 \leqslant 2{,}262)$ $= P(-2{,}262 \leqslant \dfrac{\overline{X} - m_p}{S/\sqrt{n}} \leqslant 2{,}262)$ ou de façon équivalente :

$P(\overline{X} - 2{,}262 \times S/\sqrt{10} \leqslant m_p \leqslant \overline{X} + 2{,}262 \times S/\sqrt{10}) = 0{,}95$.

Réalisation de l'intervalle de confiance : $19{,}58 \leqslant m_p \leqslant 20{,}41$.

b) La distribution étant normale, pour trouver un intervalle de confiance de σ_P on utilise la propriété « $\dfrac{(n-1)S^2}{\sigma_P^2} = \chi^2_{n-1}$ » où $n = 10$. Prenant un risque d'erreur $\alpha = 0{,}1$ on cherche $a_{\alpha/2}$ et $b_{\alpha/2}$ tels que $P(\chi^2_9 < a_{\alpha/2}) = \alpha/2 = 0{,}05$ et $P(\chi^2_9 > b_{\alpha/2}) = \alpha/2 = 0{,}05$. Par lecture de table de la loi de Khi-deux à 9 degrés de libertés, on constate que $a_{\alpha/2} = 3{,}33$ et $b_{\alpha/2} = 16{,}92$. Aussi

$$P(3{,}33 \leqslant \chi^2_9 \leqslant 16{,}92) = P(3{,}33 \leqslant \frac{(10-1)S^2}{\sigma_P^2} \leqslant 16{,}92) = 0{,}90,$$

et donc $P(9S^2/16{,}92 \leqslant \sigma_p^2 \leqslant 9S^2/3{,}33) = 0{,}90$

Réalisation de l'intervalle de confiance : $0{,}018 \leqslant \sigma_p^2 \leqslant 0{,}09$.

Exercice 3

La classification par rang croissant donne $x_{(1)} = 16$, $x_{(2)} = 18$, …, $x_{(10)} = 60$:

rang i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$x_{(i)}$	16	18	28	32	35	45	45	48	52	60

Niveau de confiance : $1 - \alpha = 0{,}85$ donc $\alpha/2 = 0{,}075$.

Le quantile considéré étant $\xi_{0,5}$ on cherche le plus grand entier i_0 tel que $P(B_{10;0,5} \leqslant i_0 - 1) \leqslant 0{,}075$

i	**0**	**1**	**2**	**3**
$P(B_{10;0.5} \leqslant i)$	0,00098	0,01074	0,05468	0,17187

Sur la table statistique l'on observe que $P(B_{10;0,5} \leqslant 2) = 0{,}05468$ et $P(B_{10;0,5} \leqslant 3) = 0{,}1787$ donc $i_0 - 1 = 2$ et $j_0 = n - i_0 + 1 = 8$. Par suite $P(3 \leqslant B_{10;0,5} < 8) = 1 - 2 \times 0{,}05468 \cong 0{,}89 > 0{,}85$.

Avec un niveau de confiance de 89 % on a $x_{(3)} \leqslant \xi_{0,5} \leqslant x_{(8)}$, c'est-à-dire $28 \leqslant \xi_{0,5} \leqslant 48$.

Exercice 4

1. Le nombre aléatoire K d'employés qui sur l'échantillon possèdent le caractère étudié « être satisfait de la formation FT », suit une loi binomiale $\mathcal{B}(50;\, p)$ où p est la proportion des employés qui, parmi les 200, sont satisfaits : $K = B_{50;\,p}$.

La statistique utilisée : $n = 50$ étant supérieur à 30 on peut utiliser l'approximation « $\dfrac{F-p}{\sqrt{\frac{N-n}{N-1}F(1-F)/n}} \cong \mathcal{N}_{0;1}$ » où $N = 200$ et F prend la valeur $f = 40/50 = 0{,}8$.

On a : $\mathrm{P}(-1{,}64 \leqslant \mathcal{N}_{0;1} \leqslant 1{,}64) = 0{,}90$.

Donc $\mathrm{P}(-1{,}64 \leqslant \dfrac{F-p}{\sqrt{\frac{N-n}{N-1}F(1-F)/n}} \leqslant 1{,}64) \cong 0{,}90$ ou de façon équivalente :

$$\mathrm{P}\left(F - 1{,}64\sqrt{\frac{200-50}{200-1}F(1-F)/50} \leqslant p \leqslant F + 1{,}64\sqrt{\frac{200-50}{200-1}F(1-F)/50}\right) \cong 0{,}90.$$

Réalisation de l'intervalle de confiance :

$$0{,}719 \leqslant p \leqslant 0.880$$

2. *Statistique employée* : $\dfrac{\overline{X} - m_p}{\frac{S}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}} \cong \mathcal{N}_{0;1}$.

L'intervalle de confiance recherché étant de 95%, $\Phi(b_{\alpha/2}) = 1 - \alpha/2 = 0{,}975$.

Aussi, $0{,}95 = \mathrm{P}(-1{,}96 \leqslant \mathcal{N}_{0,1} \leqslant 1{,}96) \cong \mathrm{P}(-1{,}96 \leqslant \dfrac{\overline{X} - m_p}{\frac{S}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}} \leqslant 1{,}96)$ ou

$$\mathrm{P}(\overline{X} - 1{,}96 \times \frac{S}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \leqslant m_P \leqslant \overline{X} + 1{,}96 \times \frac{S}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}) = 0{,}95\,.$$

Réalisation de l'intervalle de confiance :

$$68 - 1{,}96 \times \frac{24}{\sqrt{50}} \times \sqrt{\frac{200-50}{200-1}} \leqslant m_P \leqslant 68 + 1{,}96 \times \frac{24}{\sqrt{50}} \times \sqrt{\frac{200-50}{200-1}}\,,$$

soit $62{,}22 \leqslant m_P \leqslant 73{,}77$.

Exercice 5

1. On sait que $\mathrm{E}(\overline{X}^{(i)}) = m_i$ et $V(\overline{X}^{(i)}) = \sigma_i^2/n_i.(i = 1{,}2)$. Donc $\mathrm{E}(Z) = \dfrac{\nu_1}{\nu}m_1 + \dfrac{\nu_2}{\nu}m_2 = m$. $\overline{X}^{(1)}$ et $\overline{X}^{(2)}$ étant indépendants car définis sur des populations distinctes on a $\mathrm{V(Z)} = \mathrm{V}\left(\dfrac{\nu_1}{\nu}\overline{X}^{(1)}\right) + \mathrm{V}\left(\dfrac{\nu_2}{\nu}\overline{X}^{(2)}\right) = \dfrac{\nu_1^2}{\nu^2} \times \dfrac{\sigma_1^2}{n_1} + \dfrac{\nu_2^2}{\nu^2} \times \dfrac{\sigma_2^2}{n_2} \longrightarrow 0$ lorsque $n_1 \longrightarrow \infty$ et $n_2 \longrightarrow \infty$. Par suite Z est bien un ECSB de m.

© Dunod. La photocopie non autorisée est un délit.

2. On cherche le minimum de $\frac{\nu_1^2}{\nu^2} \times \frac{\sigma_1^2}{n_1} + \frac{\nu_2^2}{\nu^2} \times \frac{\sigma_2^2}{n_2}$ sous la contrainte $n_1 + n_2 - n = 0$. En utilisant le Lagrangien $L(n_1,n_2,\lambda) = \frac{\nu_1^2}{\nu^2} \times \frac{\sigma_1^2}{n_1} + \frac{\nu_2^2}{\nu^2} \times \frac{\sigma_2^2}{n_2} + \lambda(n_1 + n_2 - n = 0)$ on constate que les conditions $\frac{\partial L}{\partial n_1} = 0$, $\frac{\partial L}{\partial n_2} = 0$, $\frac{\partial L}{\partial \lambda} = 0$ sont satisfaites lorsque $n_i = \frac{\nu_i \sigma_i}{\nu_1 \sigma_1 + \nu_2 \sigma_2} n \quad \forall i = 1,2.$

Généralisation. La population P est partagée en k classes : $P = P_1 \cup P_2 \cup \ldots P_k$ où card.$P_h = \nu_h$ et où la moyenne et la variance de la classe P_h sont respectivement notées m_h et σ_h^2. On prélève dans chaque classe P_h un échantillon avec remise $(X_1^{(h)},\ldots,X_{nh}^{(h)})$ de taille n_h et l'on note $\overline{X}^{(h)} = \frac{1}{n_h} \sum_{i=1}^{n_h} X_i^{(h)}$. La taille globale n de l'ensemble des k échantillons étant fixée ($\sum_{h=1}^{k} n_h = n$ entier fixé)

$$n_h = \frac{\nu_h \sigma_h}{\sum_{h=1}^{k} \nu_h \sigma_h} n \quad \forall h = 1, \ldots, k.$$

Exercice 6

1. Lorsque qu'un échantillon de grande taille n est extrait au hasard et de façon non exhaustive d'une population binomiale $\boldsymbol{P}$, on utilise l'approximation $\frac{F - p}{\sqrt{F(1-F)/(n-1)}} \cong \mathcal{N}_{0;1}$ où F désigne la proportion dans l'échantillon. L'intervalle de confiance de p^* étant de niveau 0.95 on cherche a_α tel que $0{,}95 = P(|\mathcal{N}_{0;1}| \leqslant a_\alpha) = 2\Phi(a_\alpha) - 1$. Par lecture de la table de la loi normale standard on lit $a_\alpha = 1.96$.

Donc $0.95 = P(|\mathcal{N}_{0;1}| \leqslant 1.96) \cong P\left(\frac{|F - p^*|}{\sqrt{F(1-F)/(2500-1)}} \leqslant 1.96\right)$.

La proportion F dans l'échantillon prend la valeur $f = 375/2500 = 0.15$ donc, avec un risque d'erreur de 5 %, on peut affirmer que : $\frac{|0.15 - p^*|}{\sqrt{0.15(1-0.15)/(250-1)}} \leqslant 1.96$ et donc $0.136 \leqslant p^* \leqslant 0.164$

2. e désignant un électeur choisi au hasard dans la population des votants, considérons les *évènements* J « e est né en janvier » et E « e votera pour la liste EXT ». On a $p(J \cup E) = p^*$, $p(E) = p$, $p(J) \cong 1/12$ et $p(J \cap E) = p(J) \times p(E)$ car les évènements sont indépendants.

De la formule de Poincaré $p(J \cup E) = p(J) + p(E) - p(J \cap E)$ on déduit $p^* = (1/12) + p - p/12$ puis $p = (12p^* - 1)/11$ et par suite $0.057 \leqslant p^* \leqslant 0.088$.

CHAPITRE 11

Exercice 1

1. Sachant que l'entreprise doit verser une amende si elle dépasse la limite de 6 mg/kg de polluant dans les déchets de l'usine, le test le plus approprié est le test H_0^* « $m_p \leqslant 6$ » contre H_1 « $m_p > 6$ ». La réalisation pratique de ce test nécessite de tester H_0 « $m_p = 6$ » contre H_1 « $m_p > 6$ ».

Statistique utilisée. $X_1, \ldots, X_{12}$ étant un échantillon iid d'une loi $\mathcal{N}(m, \sigma^2)$, si H_0 est vraie, on a

$T_0 = \dfrac{(\overline{X} - 6)}{S/\sqrt{12}} = \boldsymbol{t}_{12-1}$ (variable de Student). Sous l'hypothèse alternative H_1 on a $\overline{X} \xrightarrow{\text{pr}} m_P > 6$ donc T_0 tend à prendre des valeurs positives grandes. Le domaine de rejet de H_0 est de type $[c_\alpha, \infty[$ où c_α est tel que $\alpha = 0{,}05 = P(T_0 \geqslant c_\alpha) = P(\boldsymbol{t}_{11} > c_\alpha)$, soit $c_\alpha = 1{,}796$. T_0 prenant la valeur $t_0^* = \dfrac{(6{,}8 - 6)}{\sqrt{3{,}6}/\sqrt{12}} = 1{,}46 < 1{,}796$ on ne peut rejeter H_0 et donc *a fortiori* H_0^*. Le niveau de signification observé $\alpha_c = P(\boldsymbol{t}_{11} \geqslant 1{,}46) = 0{,}086$ est le risque minimum associé au rejet à tort de H_0^*.

2. Testons H_0 « $\sigma_p = 2$ » contre $\overline{H}_0$ « $\sigma_p \neq 2$ ».

Statistique utilisée. Sous l'hypothèse H_0 « $\sigma_p = 2$ », on a $T_0 = \dfrac{(12-1)S^2}{\sigma^2} = \chi^2_{12-1}$.

Règle de décision. Test bilatéral avec $\alpha = 5\%$. La zone de rejet est de type $]0, c_{1,\alpha/2}] \cup [c_{2,\alpha/2}, \infty[$ où $c_{1,\alpha/2}$ est tel que $\alpha/2 = 0{,}025 = P_0(T_0 \leqslant c_{1,\alpha/2}) = P(\chi^2_{11} \leqslant c_{1,\alpha/2})$ d'où $c_{1,\alpha/2} = 3{,}82$ et $c_{2,\alpha/2}$ est tel que $\alpha/2 = 0{,}025 = P_0(T_0 \geqslant c_{2,\alpha/2}) = P(\chi^2_{11} \geqslant c_{2,\alpha/2})$ d'où $c_{2,\alpha/2} = 21{,}9$.

La valeur $t_0^* = \dfrac{(n-1)S^2}{\sigma_0^2} = \dfrac{(12-1)3{,}6}{4} = 9{,}9$ prise par T_0 n'appartenant pas à la zone de rejet $]0,\ 3{,}82] \cup [21{,}9,\ \infty[$ on décide d'accepter l'hypothèse H_0.

Exercice 2

Il s'agit de tester H_0^* « $p \geqslant 0{,}30$ » (et donc le lancement de la campagne publicitaire n'est pas nécessaire) contre H_1' « $p < 0{,}30$ » (auquel cas la société doit accélérer la campagne publicitaire). Techniquement, il suffit de tester H_0 « $p = 0{,}30$ » contre H_1' en utilisant la statistique appropriée.

$F = K/n$ est un estimateur de p où $n = 500$ est la taille de l'échantillon et K est le nombre aléatoire d'individus qui, sur l'échantillon, possèdent le caractère étudié.

Propriété utilisée. « $\dfrac{(F-p)}{\sqrt{F(1-F)/n}} \cong \mathcal{N}_{0;1}$ » puisque $n = 500$ est grand et que l'on teste H_0^*.

© Dunod. La photocopie non autorisée est un délit.

Statistique utilisée. Sous H_0 « $p = 0{,}30$ », $T_0 = \dfrac{(F - 0{,}30)}{\sqrt{F(1-F)/n}} \cong \mathcal{N}_{0;1}$ »

Règle de décision. Si H_1' est vraie, alors $F \xrightarrow[n\to\infty]{\text{pr}} p < 0{,}30$ donc T_0 tend à prendre des valeurs négatives. La zone de rejet de H_0 est $]-\infty, c'_\alpha]$ où c'_α est tel que $\alpha = 0{,}10 = P_0(T_0 \leqslant c'_\alpha) \cong P(\mathcal{N}_{0;1} \leqslant c'_\alpha)$, d'où $c'_\alpha = -1{,}28$. F prend la valeur $f = 100/500 = 0{,}20$ donc T_0 prend la valeur $t_0^* = \dfrac{(0{,}20 - 0{,}30)}{\sqrt{0{,}20(1-0{,}20)/500}}$

$= -5{,}58 < -1{,}28$

On rejette donc H_0 et *a fortiori* H_0^* avec un risque d'erreur inférieur à 10 %. Compte tenu du niveau de signification observé $\alpha_c = P_0(T_0 \leqslant -5{,}58) \cong P(\mathcal{N}_{0;1} \leqslant -5{,}58) = 1{,}2 \times 10^{-8}$ on peut, avec un risque d'erreur négligeable, affirmer que la société doit accélérer le lancement de sa campagne publicitaire afin que le manque de notoriété de ses produits n'entraîne pas des pertes éventuelles de marché.

Exercice 3

1. Il s'agit de tester H_0 : « $m_p = 4$ » contre $\overline{H}_0$ « $m_p \neq 4$ ». Considérant l'échantillon comme étant de grande taille, si H_0 est vraie on a $T_0 = \dfrac{(\overline{X} - 4)}{S/\sqrt{n}} \cong \mathcal{N}_{0;1}$.

Règle de décision. Si $\overline{H}_0$ est vraie, le domaine de rejet est de la forme :
$]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$ où $\Phi(c_{\alpha/2}) = 1 - \alpha/2 = 0{,}975$. En effet, $P_0(T_0 \leqslant -c_{\alpha/2}) = P_0(T_0 \geqslant c_{\alpha/2}) = \alpha/2 = 0{,}025$ implique $P(\mathcal{N}_{0;1} \leqslant -c_{\alpha/2}) = P(\mathcal{N}_{0;1} \geqslant c_{\alpha/2}) \cong \alpha/2 = 0{,}025$. Par lecture de la table page 374 on constate que $c_{\alpha/2} = 1{,}96$.

On a ici $n = 60$, $\overline{x} = 3.8$ et $s = 2{,}1$, donc T_0 prend la valeur

$t_0^* = (3{,}8 - 4)/\ (2{,}1/\sqrt{60}) = -0{,}737$. Cette valeur étant comprise entre $-1{,}96$ et 1.96, on ne rejette donc pas l' hypothèse H_0.

2. Considérant l'échantillon de grande taille, si H_0 est vraie on a $T_0 = \dfrac{(S^2 - (2{,}2)^2)}{\sqrt{\hat{\mu}_4 - S^4}/\sqrt{n}} \cong \mathcal{N}_{0;1}$.

Règle de décision. Sous l'hypothèse alternative : $S_n^2 \xrightarrow[n\to\infty]{\text{pr}} \sigma_P^2 < (2.2)^2$, donc T_0 tend donc à prendre de grandes valeurs négatives. Le domaine de rejet de H_0 est du type $]-\infty, c'_\alpha]$ où c'_α est tel que :

$\alpha = 0{,}10 = P_0(T_0 \leqslant c'_\alpha) \cong P(\mathcal{N}_{0;1} \leqslant c'_\alpha) = \Phi(c'_\alpha)$, d'où $c'_\alpha = -1{,}28$.

T_0 prend la valeur $t_0^* = (2{,}1^2 - 2{,}2^2)/[\sqrt{(19{,}55 - 2{,}1^4)}/\sqrt{60}] = -10{,}43$.

On prend la décision D_1 de rejeter H_0 au profit de H_1' puisque $t_0^* < -1{,}28$.

CHAPITRE 12

Exercice 1

1. On doit tester H_0 « $m_P = m_Q$ » contre H_0 « $m_P \neq m_Q$ ».
Variable statistique utilisée. Les deux échantillons pouvant être considérés de grande taille, sous H_0 la *v.a.* $T_0 = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_P} + \frac{S_Y^2}{n_Q}}}$ suit sensiblement la loi $\mathcal{N}_{(0;1)}$.

Domaine de rejet de H_0 *et règle de décision*. Le domaine de rejet de H_0 est de type $]-\infty,-c_{\alpha/2}] \cup [c_{\alpha/2},\infty[$. Ayant fait choix du niveau de signification $\alpha = 0{,}10$ on cherche $c_{\alpha/2}$ défini par $1-\alpha = P_0(-c_{\alpha/2} < T_0 < c_{\alpha/2}) \cong P(-c_{\alpha/2} < \mathcal{N}_{0;1} < c_{\alpha/2})$ c'est-à-dire $\Phi(c_{\alpha/2}) = 1-\alpha/2 = 0{,}95$, soit $c_{\alpha/2} = 1{,}64$.
La valeur $t_0^* = \dfrac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_P} + \frac{s_y^2}{n_Q}}} = \dfrac{42{,}6 - 37{,}8}{\sqrt{\frac{4{,}62^2}{50} + \frac{4{,}45}{60}}} = 5{,}51$ appartient à la zone de rejet.
On décide donc de rejeter l'hypothèse H_0.

2. On doit tester l'hypothèse H_0 « $\sigma_P = \sigma_Q$ » contre $\overline{H}_0$ « $\sigma_P \neq \sigma_Q$ ». Sous l'hypothèse H_0, la statistique utilisée $T_0 = \dfrac{S_X^2 - S_Y^2}{\sqrt{\frac{\hat{\mu}_{4,X} - S_X^4}{n_P} + \frac{\hat{\mu}_{4,Y} - S_Y^4}{n_Q}}}$ suit sensiblement la loi normale centrée réduite : $T_0 \cong \mathcal{N}_{0;1}$. Le domaine de rejet de H_0 est donc du type $]-\infty,-c_{\alpha/2}] \cup [c_{\alpha/2},\infty[$ où $c_{\alpha/2} = 1{,}64$ car solution de $\Phi(c_{\alpha/2}) = 1-\alpha/2 = 0{,}95$.
La valeur $t_0^* = \dfrac{4{,}62^2 - 4{,}45^2}{\sqrt{(625 - 4{,}62^4)/50 + (600 - 4{,}45^4)/60}} = 0{,}589$ prise par T_0 étant comprise entre $-1{,}64$ et $1{,}64$ on prend la décision de ne pas rejeter H_0.

Exercice 2

On doit tester H_0 « X et Y ont même distribution ». Le domaine $\Delta = \{0,1,2,3,\ldots\}$ des valeurs possibles de X et Y est partagé en 7 classes : $C_1 = \{0\}$; $C_2 = \{2\}$; …; $C_7 = \{6\}$ et l'on constate que tous les effectifs observés sont supérieurs à 3 : $n_1 = 5$, $n_1' = 15$; $n_2 = 6$, $n_2' = 16$; …; $n_7 = 10$, $n_7' = 6$.
On utilise pour indicateur de proximité des distributions la variable statistique
$Z = n \times n' \times \sum_{i=1}^{h} \dfrac{(N_i/n - N_i'/n\prime)^2}{N_i + N_i'}$ (où h le nombre de classes est égal à 7,$n = 50$, $n' = 60$) qui prend la valeur $z = nn' \sum_{i=1}^{h} \dfrac{(n_i/n - n_i'/n')^2}{n_i + n_i'} = 10{,}8$.

© Dunod. La photocopie non autorisée est un délit.

Valeurs de X et Y	0	1	2	3	4	5	6	**Total**
Effectif n_i de X	5	6	9	8	6	6	10	$n = 50$
Effectif n'_i de Y	15	16	9	6	4	4	6	$n' = 60$
$\dfrac{(n_i/n - n'_i/n')^2}{n_i + n'_i}$	0,0011	0,0010	0,0001	0,0003	0,0003	0,0003	0,0006	0,0036

Sous l'hypothèse H_0 (et compte tenu de $n_i \geqslant 3$ et $n'_i \geqslant 3$) on a $Z \cong \chi^2_{7-1}$. *Règle de décision*. Le domaine de rejet de H_0 est de type $[c_\alpha, \infty[$. Pour un niveau de signification de $\alpha = 5\,\%$, on cherche c_α tel que $P(\chi^2_6 > c_\alpha) = \alpha = 0{,}05$ et on lit sur la table statistique que $c_\alpha = 12{,}592$.

Z prenant la valeur 10,8 qui est inférieure à 12,592 on ne peut rejeter H_0. Toutefois la valeur du niveau de signification observé $\alpha_c = P_0(Z \geqslant 10{,}8) \cong P(\chi^2_6 \geqslant 10{,}8) = 0{,}095$ incite au rejet de H_0.

Exercice 3

On dispose de 7 valeurs x_h et de 9 valeurs y_h. Après classement par ordre de valeurs croissantes de ces 16 valeurs on obtient le rang r_h de x_h et le rang s_k de y_k.

Rang	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	– 0,50	– 0,32	– 0,30	– 0,30	– 0,20	– 0,18	– 0,13	0,1	0,11	0,17	0,31	0,40	0,41	0,55	0,70	0,73
Échantillon	**1**	2	2	2	2	2	2	**1**	2	**1**	2	**1**	**1**	**1**	2	**1**

Pour tester H_0 « X et Y ont même distribution » contre H_1 « $F < G$ c'est-à-dire X est stochastiquement supérieur à Y « utilisons la variable statistique $W_{7,9}$ de Wilcoxon qui prend la valeur $w = \sum_{h=1}^{7} r_h = 1 + 8 + 10 + 12 + 13 + 14 + 16 = 74$. Pour un niveau de signification $\alpha = 10\,\%$ on cherche le plus petit entier c_α tel que $P_0(W_{7,9} \geqslant c_\alpha) \geqslant 0{,}10$. On lit $c_\alpha = 2W - c'_\alpha = 119 - 46 = 73$. La *v.a.* $W_{7,9}$ prenant une valeur $w = 74$ supérieure à 73, on décide de rejeter l'hypothèse H_0 « $G = F$ » au profit de H_1.

Exercice 4

1. Il s'agit de tester H_0 « $p_P = p_Q$ » contre $\overline{H}_0$ « $p_P \neq p_Q$ » avec un risque d'erreur de 5 %.

Variable statistique utilisée. Si l'hypothèse H_0 est vraie :

$$T_0 = \frac{F_1 - F_2}{\sqrt{\overline{F}(1 - \overline{F})} \times \sqrt{\frac{1}{n_P} + \frac{1}{n_Q}}} \cong \mathcal{N}_{0;1}$$

F_1 et F_2 désignant les proportions aléatoires dans chaque échantillon et $\overline{F} = (n_P F_1 + n_q F_2)/(n_P + n_Q)$

Le domaine de rejet de H_0 est du type $]-\infty,-c_{\alpha/2}] \cup [c_{\alpha/2},\infty[$ où $c_{\alpha/2} = 1{,}96$ car défini par la condition $\Phi(c_{\alpha/2}) = 1 - \alpha/2 = 0{,}975 \cdot F_1$, F_2 et $\overline{F}$ prennent respectivement les valeurs $f_1 = 10/120 = 0{,}0833, f_2 = 7/80 = 0{,}0875, \overline{f} = (10+7)/200 = 0{,}085$. Donc la variable statistique T_0 prend la valeur $t^* = -0{,}1035$. Cette valeur étant comprise entre $-1{,}96$ et $1{,}96$ on prend la décision de ne pas rejeter H_0 .

2. Soit respectivement p_1, p_2 et p_3 les proportions d'individus qui possèdent le caractère considéré dans les 3 populations binomiales P_1, P_2, P_3. On extrait de chaque population P_i un échantillon de taille ν_i. Le nombre aléatoire N_i d'éléments de cet échantillon qui possèdent le caractère considéré suit la loi $B(\nu_i; p_i)$. On a ici $\nu = \sum_{h=1}^{h} \nu_i \;\; = 120 + 80 + 100 = 300$ et on se propose de tester H_0 « $p_1 = p_2 = p_3$ ($= p$ valeur commune inconnue) » contre $\overline{H}_0$.

Statistique utilisée. Notons, n_i la réalisation de N_i et soit $\hat{p} = \left(\sum_{i=1}^{h} N_i \right) / \nu$ qui prend pour valeur $p^* = \left(\sum_{i=1}^{h} N_i \right) / \nu = (10 + 7 + 15)/300 = 0{,}1066$. Sous H_0 (et si $\nu_i\, p^* \geqslant 4\ \forall i$) on a $Z = \sum_{i=1}^{3} \dfrac{(N_i - \nu_i \hat{p})^2}{\nu_i \hat{p}(1-\hat{p})} \cong \chi^2_{3-1}$

ν_i	120	80	100	Total 300
n_i	10	7	15	Total 32
$(n_i - \nu_i p^*)^2/\nu_i p^*(1-p^*)$	0,682	0,306	1,978	Total 2.966

Règle de décision. Sous l'hypothèse alternative $\overline{H}_0$, la variable statistique Z tend à prendre des valeurs élevées. Le domaine de rejet de H_0 est donc de type $[c_\alpha,\infty[$. Pour un niveau de signification α on détermine c_α tel que $P(\chi^2_2 \geqslant c_\alpha) = \alpha = 0{,}05$ car alors $P_0(Z \geqslant c_\alpha) \cong 0{,}05$ et on rejette H_0 si la valeur $z = \sum_{i=1}^{3} \dfrac{(n_i - \nu_i p^*)^2}{\nu_i p^*(1-p^*)} = 2{,}966$ est supérieure à c_α. Page 377 on lit $c_\alpha = 5{,}99$ donc on ne peut rejeter H_0. Remarquer que $\alpha_c = P_0(Z \geqslant 2{,}966) \cong P(\chi^2_2 \geqslant 2{,}966) \cong e^{-2{,}966/2} = 0{,}227$.

QCM : ④

© Dunod. La photocopie non autorisée est un délit.

CHAPITRE 13

Exercice 1

On teste H_0 « X et Y indépendants et donc $\rho(X,Y) = 0$ « contre H_1 » $\rho(X,Y) > 0$, les valeurs prises par X et Y ont tendance à être concordantes ».

À chaque couple aléatoire (X_i, Y_i) on peut associer le couple aléatoire d'entiers (R_i, S_i) qui prend pour valeur le couple d'entiers (r_i, s_i) où r_i est le rang de x_i dans $x_1, \ldots, x_n$ et s_i le rang de y_i dans $y_1, \ldots, y_n$. La variable aléatoire $\rho_s = \text{Cov}(R,S)/\sigma(R) \times \sigma(S)$ est un estimateur de $\rho(X,Y)$.

Sous l'hypothèse H_0 et pour $n = 8$ *cf.* p. 381 des valeurs critiques de la distribution de ρ_S.

Règle de décision. Sous l'hypothèse alternative $H_1 : \rho_S \xrightarrow{\text{pr}} \rho(X,Y) > 0$ donc ρ_S tend à prendre des valeurs strictement positives et par suite le domaine de rejet de H_0 est de type $[c_\alpha, 1]$ où c_α est tel que $P_0(\rho_S \geqslant c_\alpha) \cong \alpha = 0{,}05$. Par lecture de table on obtient $c_\alpha = 0{,}643$. Pour calculer la valeur prise par ρ_S réécrivons les 8 couples de résultats selon les valeurs croissantes de la première composante x_i et faisons apparaître les rangs r_i et s_i.

Rang r_i	1	2	3	4	5	6	7	8	total = 36
x_i	– 10,79	– 1,18	– 0,85	– 0,6	– 0,23	0,11	0,47	3,18	
y_i	5	32	57	71	130	213	162	19	
rang s_i	1	3	4	5	6	8	7	2	total = 36
$(r_i - s_i)^2$	0	1	1	1	1	4	0	36	total = 44

$$\rho_S^* = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(r_i - s_i)^2 = 1 - \frac{6}{8^3 - 8} \times 44 = 0{,}48$$

On prend la décision de ne pas rejeter H_0 au profit de H_1 puisque $\rho_S^* = 0{,}48 < 0{,}643$.

Exercice 2

À chaque individu e est associé son âge $X(e)$ et son opinion $Y(e)$ qui est de type 1, 2 ou 3. On se propose de tester H_0 « X et Y indépendants » contre $\overline{H}_0$ où X désigne l'âge des clients potentiels et Y désigne leur intention d'achat. Δ_X est partagé en 4 classes ($h = 4$) et Δ_Y en 3 classes ($k = 3$) ainsi que l'indique le tableau ci-dessous. Les 400 couples de valeurs sont donc répartis entre les 12 classes $\mathbf{C_{ij}}$. Dans le tableau ci-après figurent les effectifs théoriques $n_{ij}^* = n_{i\bullet} \times n_{\bullet j}/n$, soit pour la première classe $100 \times 200/400 = 50$ etc..

Y / X		oui	non	ne peut pas se prononcer	Total
< 30 ans	Effectifs	65	27	8	100
	Effectifs théoriques	50	37,5	12,5	
de 30 à 45 ans	Effectifs	50	19	11	80
	Effectifs théoriques	40	30	10	
de 45 à 60 ans	Effectifs	35	24	11	70
	Effectifs théoriques	35	26,25	8,75	
plus de 60 ans	Effectifs	50	80	20	150
	Effectifs théoriques	75	56,25	18,75	
	Total	200	150	50	400

Statistique utilisée. Tous les effectifs théoriques étant supérieurs à 5, sous l'hypothèse H_0 . la *v.a.* $Z = \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{(N_{ij} - N_{i\bullet} \times N_{\bullet j}/400)^2}{N_{i\bullet} \times N_{\bullet j}/400}$ suit sensiblement la loi $\chi^2_{(4-1)(3-1)}$: $Z \cong \chi^2_{(6)}$.

Règle de décision. Sous l'hypothèse alternative $\overline{H}_0$, la variable statistique Z tend à prendre des valeurs élevées. Le domaine de rejet de H_0 est donc de type $[c_\alpha, \infty[$ où c_α est tel que $P_0(Z > c_\alpha) \cong P(\chi^2_{(6)} > c_\alpha) = \alpha = 0{,}05$, soit $c_\alpha = 12{,}59$. La *v.a.* Z prend pour valeur $z = (65-50)^2/50 + (27-37{,}5)^2/37{,}5 + \ldots \quad +(20-18{,}75)^2/18{,}75 = 34{,}9$ qui est supérieure à 12,59 donc on rejette l'hypothèse d'indépendance.

CHAPITRE 14

Exercice 1

1. Sur l'échantillon on observe que $\overline{x} = 0{,}800$ est proche de la variance standard $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0{,}855$ aussi peut-on penser à une loi de Poisson.

2. X désignant le nombre aléatoire de destinations saturées, on teste l'hypothèse H_0 « X suit la loi de Poisson de paramètre $\lambda = 0{,}8$ ». Pour cela on répartit les 30 valeurs entre les 4 classes que sont naturellement $(X = 0)$, $(X = 1)$, $(X = 2)$ et $(X \geqslant 3)$ puis on calcule successivement les probabilités $p_i = P_0(X \in C_i)$, les effectifs théoriques np_i, et les termes $(n_i - np_i)^2/np_i$. L'effectif théorique de la dernière classe $np_4 = 30(1 - p_1 - p_2 - p_3) = 1{,}42$ étant insuffisant on regroupe les deux dernières classes et on obtient le tableau ci-après.

Ainsi $p_1 = P_0(X \in C_1) = P(X = 0/X \rightsquigarrow \mathcal{P}(0{,}8)) = e^{-0{,}8}(0{,}8^0/(0!) = 0{,}449$ d'où $np_1 = 30 \times 0{,}449 = 13{,}48$ et $(n_1 - np_1)^2/np_1 = (12 - 13{,}48)^2/13{,}48 = 0{,}162$ etc.

© Dunod. La photocopie non autorisée est un délit.

Valeurs de X	0	1	$X \geqslant 2$	Total
Proba $p_i = \mathrm{P}_0(X \in C_i)$	0,449	0,359	0,191	1
Effectif théorique np_i	13,480	10,784	5,736	30
Effectif réel n_i	12	12	6	30
$(n_i - np_i)^2/np_i$	0,162	0,137	0,012	0,312

La statistique $Z = \sum_{i=1}^{3} \frac{(N_i - 30p_i)^2}{30p_i}$ prend la valeur $z = 0{,}312$. Sous H_0, la *v.a.* Z suit sensiblement la loi khi-deux à $(3-1) = 2$ degrés de liberté. Prenant un risque d'erreur de première espèce $\alpha = 0{,}05$, on cherche c_α tel que $0{,}05 = \mathrm{P}_0(Z \geqslant c_\alpha) \cong \mathrm{P}(c_\alpha)$. Page 377 on lit $c_\alpha = 5{,}99$.

z étant inférieur à 5.99 on accepte H_0.

Exercice 2

Les 12 valeurs x_i ayant été ordonnées par ordre de valeurs croissantes : $x_{(1)} = 15 < x_{(2)} = 18 < \ldots < x_{(12)} = 25$, la fonction de répartition empirique $G^*_{12}(x_i) = i/12$. On veut tester H_0 « X suit une loi normale $\mathcal{N}(m; \sigma^2)$ » contre $\overline{\mathrm{H}}_0$ en utilisant le test d'ajustement de Kolmogorov-Smirnov. Sous l'hypothèse nulle H_0, estimant m par $\overline{x} = 20{,}25$ et σ par $s = 2{,}58$, on a la fonction de répartition $F(x_i) = \Phi[(x_{(i)} - 20{,}25)/2{,}58]$. Aussi $d_i^+ = (i/n) - \Phi[(x_{(i)} - \overline{x})/s]$ et $d_i^- = \Phi[(x_{(i)} - \overline{x})/s] - (i-1)/n$:

Clients	4	5	10	6	9	1	8	7	11	2	12	3
Rang i	1	2	3	4	5	6	7	8	9	10	11	12
Score x_i	15	18	18,5	19	19,5	20	20,5	21	21,5	22	23	25
(i/n), $i/12$	0,083	0,167	0,250	0,333	0,417	0,500	0,583	0,667	0,750	0,833	0,917	1,000
$\Phi[(x_{(i)} - \overline{x})/s]$	0,021	0.192	0,249	0,314	0,386	0,461	0,539	0,614	0,686	0,751	0,857	0,967
d_i^+	0,062	– 0,025	0,001	0,019	0,031	0,039	0,045	0,052	0,064	0,082	0,060	0,033
$((i-1)/n)$, $(i-1)/12$		0,083	0,167	0,250	0,333	0,417	0,500	0,583	0,667	0,750	0,833	0,917
d_i^-	0,021	0,108	0,082	0,064	0,052	0,045	0,039	0,031	0,019	0,001	0,023	0,051

La statistique D de Kolmogorov-Smirnov prend la valeur $d = max\ \{d^+, d^-\} = 0{,}108$ et $n = 12$. Prenant un risque d'erreur $\alpha = 0{,}05$, on rejette H_0 si $(n^{1/2} + 0{,}85 \times n^{-1/2} - 0{,}01)d > 0{,}895$.

Ici $(12^{1/2} + 0{,}85(12^{-1/2}) - 0{,}01) \times 0{,}108 = 0{,}399$ est inférieur à $0{,}895$, aussi ne peut-on rejeter H_0.

CHAPITRE 15

Exercice

1. *Le test d'analyse de variance de Fisher*. Notons X_{ij} le score d'implication du j-ème vendeur qui a le mode de rémunération G_i ($i = 1,2$ ou 3). Si on admet que X_{ij} fluctue de façon normale autour de sa valeur moyenne m_i qui caractérise l'effet du mode de rémunération G_i, on peut appliquer le test associé à l'analyse de variance de Fisher. Celle-ci peut dans cette application, être représentée par le tableau suivant :

Observation	1	2	3	4	5	$\overline{x}_i$	n_i	$\Sigma_j (x_{ij} - \overline{x}_i)^2$	$n_i \times (\overline{x}_i - \overline{x})^2$
G_1	$x_{11} = 60$	$x_{12} = 65$	$x_{13} = 85$	$x_{14} = 80$	$x_{15} = 75$	73	5	430	40,50
G_2	$x_{21} = 52$	$x_{22} = 66$	$x_{23} = 82$	$x_{24} = 77$		69,25	4	530,75	3,27
G_3	$x_{31} = 36$	$x_{32} = 58$	$x_{33} = 92$	$x_{34} = 84$		67,5	4	1955	28,17
Total								2 915,75	71,94

On a $n = 13$ observations et $k = 3$ sous-populations. La moyenne de l'ensemble des observations $\overline{x} = 70{,}15 = (60 + \ldots + 84)/13$. On constate que F prend la valeur $f = \big[(71{,}94/(3-1)\big]/\big[(2915{,}75/(13-3)\big] \cong 0{,}12$. Sous l'hypothèse « G_1, G_2 et G_3 ont la même distribution supposée normale » alors F suit la loi de Fisher-Snédécor $F(k-1; n-k)$ c'est-à-dire ici $F(2; 10)$. Avec un niveau de signification de 5 % on cherche w_α tel que $0{,}05 = \mathrm{P}_0(F \geqslant c_\alpha) = \mathrm{P}(F^2_{10} \geqslant c_\alpha)$. On obtient $c_\alpha = 4{,}10$ (*cf.* page 379). w étant inférieur à 4,10 on accepte donc l'hypothèse selon laquelle les différents types de promotions ont le même impact.

2.1. *Le test de Jonckheere-Terpstra*. On réalise une étude comparative des échantillons :
– aux premier et deuxième échantillons d'effectifs respectifs 5 et 4, on associe la variable de Mann-Whitney U_{12} qui est égale au nombre de couples (h,ℓ) tels que $X_{1h} > X_{2\ell}$: U_{12} prend donc pour valeur $u_{12} = 1 + 1 + 4 + 3 + 2 = 11$;
– aux premier et troisième échantillon, on associe la valeur $u_{13} = 2 + 2 + 3 + 2 + 2 = 11$;
– aux deuxième et troisième échantillon, on associe la valeur $u_{23} = 1 + 2 + 2 + 2 = 7$.

La statistique J étant définie comme étant la somme des variables U_{ij} (avec $i < j$), on en déduit que la statistique J prend la valeur $29 = (11 + 11 + 7)$. La distribution de J étant tabulée sous l'hypothèse de base H_0 « $m_1 = m_2 = m_3$ » on cherche sur la table statistique le plus petit entier θ_α tel que $\mathrm{P}_0(J \geqslant \theta_\alpha) \leqslant 5\,\%$. Sur la table statistique (*cf.* p. 385 et puisque $n_1 = 5$, $n_2 = 4$ et $n_3 = 4$ on lit pour (4, 4, 5) et $\alpha = 0{,}05$, $\theta_\alpha = 41$. La valeur prise par J étant inférieure à on ne peut rejeter l'hypothèse de base.

2.2. *Test utilisant la statistique KW de Kruskal-Wallis*. On range les 13 valeurs numériques par ordre croissant et on attribue à chacune d'elles son rang parmi l'ensemble des observations :

© Dunod. La photocopie non autorisée est un délit.

Échantillon extrait de G_1	4	5	7	9	12	$R_{1\bullet} = 37$	$\overline{R}_{1\bullet} = 7{,}4$
Échantillon extrait de G_2	2	6	8	10		$R_{2\bullet} = 26$	$\overline{R}_{2\bullet} = 6{,}5$
Échantillon extrait de G_3	1	3	11	13		$R_{3\bullet} = 28$	$\overline{R}_{3\bullet} = 7$

La statistique $KW = \dfrac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\overline{R}_{i\bullet} - \dfrac{n+1}{2} \right)^2$ prend la valeur $kw^* = [12/(13 \times 14)] \times [5 \times (7{,}4 - 7)^2 + 4 \times (6{,}5 - 7)^2 + 4(7 - 7)^2] = 0{,}119$. Pour $\alpha = 0{,}05$, on cherche c_α tel que $\mathrm{P}_0(KW \geqslant c_\alpha) \cong 0{,}05$ et on obtient $c_\alpha = 5{,}62$ (*cf.* p. 383). On a $kw^* < 5{,}62$ donc on ne peut rejeter H_0.

CHAPITRE 16

Exercice

1. La droite d'ajustement a pour équation « $y'_t = 3{,}564 \times t + 129{,}3$ » car $\overline{t} = 8{,}5$, $\mathrm{V}(t) = 21{,}25$, $\overline{y} = 159{,}6$; $\mathrm{Cov}\,(t, y_t) = 75{,}75$ et $a_0 = \mathrm{Cov}(t, y_t)/\mathrm{V}(t) = 3{,}564$, $b_0 = \overline{y} - a_0\overline{t}$.

Les résidus $e_i = y'_t - y_t$ figurent ci-après :

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_t	132,89	136,45	140,02	143,58	147,15	150,71	154,28	157,84	161,41	164,97	168,54	172,10	175,67	179,23	182,80	186,36
e_i	– 32,89	– 16,45	– 9,02	56,42	– 12,15	– 20,71	– 13,28	52,16	38,59	– 29,97	– 17,54	47,90	34,33	– 34,23	– 21,80	– 21,36

On a $\mathrm{V}(y_t) = 1\,292{,}6$, $\mathrm{V}_R(y_t) = \dfrac{1}{16} \sum_{t=1}^{16} e_t^2 = 1\,022{,}58 = (1 - r^2)\mathrm{V}(y_t)$ d'où $r^2 = 0{,}209$

La valeur de r^2 étant proche de 0,2 on peut estimer que l'ajustement linéaire est de piètre qualité.

2. Partant du modèle $\varepsilon_i = \rho \times \varepsilon_{i-1} + \nu_i$ on teste l'hypothèse H_0 « $\rho = 0$ » contre $\overline{\mathrm{H}}_0$ « $\rho \neq 0$ ».

$d^* = \sum_{i=2}^{16} (e_i - e_{i-1})^2 / \sum_{i=1}^{16} e_i^2 = 1{,}716$ est la valeur prise par la statistique de Durbin-Watson. Par lecture de table p. 389, on obtient les seuils suivants : $d_2 = 1{,}371$ et $d_1 = 1{,}106$. On a $d_2 < d^* \leqslant 4 - d_2$, aussi accepte-t-on l'hypothèse d'indépendance des résidus.

3. Pour tester H_0 « $a = 0$ » contre $\overline{\mathrm{H}}_0$ « $a \neq 0$ » avec un niveau de signification $\alpha = 0{,}10$ on utilise sous l'hypothèse H_0, la variable statistique $T_0 = (\hat{a} - 0) \times \sqrt{(n-2)\mathrm{V}(t)}/\hat{\sigma}_\varepsilon = t_{14}$. Le domaine de rejet de H_0 est du type $]-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty[$

où $c_{\alpha/2}$ est défini par $\alpha/2 = \mathrm{P}_0(T_0 \geqslant c_{\alpha/2}) = \mathrm{P}(\boldsymbol{t}_{14} \geqslant c_{\alpha/2})$. Pour $\alpha = 0{,}10$ on trouve $c_{\alpha/2} = 1{,}76$.

$t_0^* = 3{,}56 \times \sqrt{14 \times 21{,}25}/\sqrt{1022{,}58} = 1{,}92$ qui est supérieure à 1,76 et par suite on rejette H_0.

4. Pour déterminer un intervalle de confiance de a de niveau 0,95 on utilise à nouveau la propriété « $(\hat{a} - a) \times \sqrt{(n-2)\mathrm{V}(t)}/\sigma_\varepsilon = \boldsymbol{t}_{n-2}$ ». On constate que : $\mathrm{P}(-2{,}14 \leqslant \boldsymbol{t}_{14} \leqslant 2{,}14) = 0{,}95$ donc $0{,}95 = \mathrm{P}(-2{,}14 \leqslant (\hat{a} - a) \times \sqrt{(n-2)V(\mathrm{t})}/\hat{\sigma}_\varepsilon \leqslant 2{,}14)$ ou de façon équivalente :

$$0{,}95 = \mathrm{P}(\hat{a} - 2{,}14 \times \hat{\sigma}_\varepsilon/\sqrt{(n-2)\mathrm{V}(t)} \leqslant a \leqslant \hat{a} + 2{,}14 \times \hat{\sigma}_\varepsilon/\sqrt{(n-2)\mathrm{V}(t)}) .$$

Donc avec un niveau de confiance de 95 % on peut considérer que l'inégalité ci-dessous est vraie

$-2{,}14 \times \sqrt{1\,022{,}58}/\sqrt{14 \times 21{,}25} + 3{,}56 \leqslant a \leqslant 2{,}14 \times \sqrt{1\,022{,}58}/\sqrt{14 \times 21{,}25} + 3{,}56$ soit $-0{,}37 \leqslant a \leqslant 7{,}49$.

5. Dans le modèle considéré ici (*cf.* § 3.2) $n = 16$ et $k = 4$. Il peut s'exprimer sous la forme matricielle $Y = \hat{\mathbf{a}}X + \varepsilon$

Tableau des estimations et tests sur les coefficients

	Valeurs $a_i^0 \downarrow$	$\theta_n\sqrt{\omega'_{ii}}$	$t_0^* = (a_i^0 - 0)/\theta_n^*\sqrt{\omega'_{ii}}$	$\alpha_c =$ $2 \times \mathrm{P}(\boldsymbol{t}_{n-k-1} > \lvert t_0^* \rvert)$	**Borne inférieure de l'IC à 0,95**	**Borne supérieure de l'IC à 0,95**
a_1	– 28,500	19,577	– 1,456	0,173	– 71,588	14,588
a_2	– 60,250	19,285	– 3,124	0,01	– 102,696	– 17,804
a_3	– 49,750	19,108	– 2,604	0,025	– 91,806	– 7,694
a_4	3,000	1,506	1,992	0,072	– 0,314	6,314
a_5	168,750	20,204	8,352	0	124,282	213,218

CHAPITRE 17

Exercice 1

1. Soit f_{ij} la fréquence relative observée sur la classe $A_i \times B_j$ et soit f_{ij}^* la valeur estimée de cette classe par la méthode du maximum de vraisemblance. Pour tester l'hypothèse d'absence H_0 d'interaction « $\lambda_{11}^{\mathrm{AB}} = 0$ », utilisons la propriété selon laquelle si l'hypothèse H_0 est vraie et si $nf_{ij}^* > 3 \ \forall i, j,$ la *v.a.* $Y^2 = 2n\Sigma_{ij} f_{ij} \, [\ln(f_{ij}) - \ln(f_{ij}^*)]$ suit sensiblement la loi $\chi^2(1)$ (le nombre de degrés de liberté correspondant au nombre de paramètres supposés nuls).

© Dunod. La photocopie non autorisée est un délit.

Partant des données, on calcule les fréquences relatives : $f_{11} = 50/400 = 0{,}125$ etc.

	B_1	B_2	**Total** $f_{i\bullet}$
A_1	$f_{11} = 0{,}125$	$f_{12} = 0{,}25$	$f_{1\bullet} = 0{,}375$
A_2	$f_{21} = 0{,}50$	$f_{22} = 0{,}125$	$f_{2\bullet} = 0{,}625$
Total $f_{\bullet j}$	$f_{\bullet 1} = 0{,}625$	$f_{\bullet 2} = 0{,}375$	$f_{\bullet\bullet} = 1$

Les valeurs estimées par la méthode du maximum de vraisemblance des inconnues principales μ, λ_1^A, λ_1^B sont $\mu^* = (1/2)\ln(f_{1\bullet} \times f_{2\bullet} \times f_{3\bullet} \times f_{4\bullet}) = (1/2)\ln(0{,}375 \times 0{,}625 \times 0{,}625 \times 0{,}375) = -1{,}45$; $\lambda_1^{A^*} = (1/2)\ln(f_{1\bullet}/f_{2\bullet}) = (1/2)\ln(0{,}375/0{,}625) = -0{,}255$; $\lambda_1^{B^*} = (1/2)\ln(f_{\bullet 1}/f_{\bullet 2}) = (1/2)\ln(0{,}625/0{,}375) = 0{,}255$.

Les fréquences relatives estimées sont : $f_{11}^* = \exp[\mu^* + \lambda_1^{A^*} + \lambda_1^{B^*}] = 0{,}234$, $f_{12}^* = \exp[\mu^* + \lambda_1^{A^*} - \lambda_1^{B^*}] = 0{,}141$, $f_{21}^* = \exp[\mu^* - \lambda_1^{A^*} + \lambda_1^{B^*}] = 0{,}391$, $f_{22}^* = \exp[\mu^* - \lambda_1^{A^*} - \lambda_1^{B^*}] = 0{,}234$. Les effectifs attendus $n_{ij}^* = 400 f_{ij}^*$ sont indiqués ci-après :

	B_1	B_2
A_1	93,6	56,4
A_2	156,4	93,6

Pour tester l'hypothèse « $\lambda_{11}^{AB} = 0$ », on détermine la valeur prise par Y^2 :

$y^{*2} = 2 \times 400[0{,}125 \times (\ln(0{,}125) - \ln(0{,}234)) + \ldots + 0{,}125((\ln(0{,}125) - \ln(0{,}234))]$

$\cong 88$. Le niveau de signification du test ou risque de rejet à tort de l'hypothèse H_0 est $P_0(Y^2 \geqslant 88) = P(\chi_1^2 \geqslant 88) \cong 0{,}00$. Ce niveau de risque étant négligeable on rejette l'hypothèse de non interaction.

Exercice 2

Fonction de vraisemblance : $FV = P(Y_1 = y_1 \text{ et } \ldots Y_8 = y_8) = \prod_{i=1}^{8} p_i^{y_i}(1 - p_i^{(1-y_i)})$ où $p_i = ax_i$. Donc $L(a) = \ln[FV] = \Sigma_i y_i \ln(ax_i) + (1 - y_i)\ln(1 - ax_i)$. La valeur a^* déduite de la méthode du maximum de vraisemblance maximise $L(a)$ et donc annule sa dérivée $L'(a)$:

$L'(a) = 4/a - 0{,}24/(1 - 0{,}24a) - 0{,}35/(1 - 0{,}35a) - 0{,}42/(1 - 0{,}42a)$

$-0{,}63/(1 - 0{,}63a) = 0 \Longrightarrow a^* = 1{,}0892$.

Or la dérivée seconde $L''(1{,}0892) = -1{,}9273 < 0$ donc $L(a)$ est maximum en a^*.

Valeurs x_i	0,24	0,31	0,35	0,42	0,52	0,63	0,68	0,83
Valeurs y_i	0	1	0	0	1	0	1	1
p_i	0,26	0,34	0,38	0,46	0,57	0,69	0,74	0,90
$p_i \times (1 - p_i)$	0,19	0,22	0,24	0,25	0,25	0,22	0,19	0,09

ANNEXES

Section 1

LOIS CONTINUES BIVARIÉES

1 Caractérisation

On dit que le couple aléatoire (X,Y) suit une loi bivariée continue caractérisée par son support D, partie convexe de $\mathbb{R}^2$ et sa fonction de densité $\varphi(x,y) : D \longrightarrow \mathbb{R}$ supposée continue, lorsque pour tout domaine δ inclus dans D on a $P[(X,Y) \in \delta] = \iint_\delta \varphi(x,y)\mathrm{d}x\mathrm{d}y$

Remarque. « $(X,Y) \in D$ » étant un événement certain on a nécessairement $\iint_D \varphi(x,y)\mathrm{d}x\mathrm{d}y = 1$.

Fonction de répartition du couple (X,Y). La fonction de répartition $H(t,u) = P(X \leqslant t \text{ et } Y \leqslant u)$ est déterminée par l'expression suivante :

$H(t,u) = \iint_{\delta(t,v)} \varphi(x,y)\mathrm{d}x\mathrm{d}y$ où $\delta(t,u) = \{(x,y) \in D / x \leqslant t \text{ et } y \leqslant u\}$.

De plus $\dfrac{\partial^2 H}{\partial x \partial y}(x,y) = \varphi(x,y)\ \forall (x,y) \in D^\circ$ (intérieur de D)

Lois marginales. La projection du domaine D sur l'axe $x'Ox$ est un intervalle (a,b) : $(a,b) = \{x \in \mathbb{R} / (x,y) \in D\}$. De même on définit l'intervalle $(c,d) = \{y \in \mathbb{R} / (x,y) \in D\}$. Par suite $D \subset (a,b) \times (c,d)$.

Pour chaque $x_0 \in (a,b)$, $\mathrm{I}(x_0) = \{y \in \mathbb{R} | (x_0,y) \in D\}$ est un intervalle.

© Dunod. La photocopie non autorisée est un délit.

Loi de X. La *v.a. X* suit la loi continue caractérisée par son support $[a,b]$ et sa densité de probabilité $f(x) = \int_{I(x)} \varphi(x,y)\mathrm{d}y$.

Loi de Y. La *v.a. Y* suit la loi continue caractérisée par on support $[c,d]$ et sa densité de probabilité $g(y) = \int_{J(y)} \varphi(x,y)\mathrm{d}x$ où $J(y_o) = \{x \in \mathbb{R}/(x,y_o) \in D\}$

Moments. $\psi(x,y)$ désigne une fonction à valeur numérique définie et continue sur le support D de la loi que suit le couple aléatoire (X,Y).

On définit $\mathrm{E}[\psi(X,Y)] = \iint_D \psi(x,y)\varphi(x,y)\mathrm{d}x\mathrm{d}y$

En particulier, r et s désignant des entiers naturels on définit

• le moment d'ordre r en X et s en Y : $m_{r,s} = \mathrm{E}(X^r Y^s) = \iint_D x^r y^s \varphi(x,y)\mathrm{d}x\mathrm{d}y$

• le moment centré d'ordre r en X et s en Y :

$\mu_{r,s} = \mathrm{E}[(X-m_1)^r (Y-m_2)^s] = \iint_D (x-m_1)^r (y-m_2)^s \varphi(x,y)\mathrm{d}x\mathrm{d}y$

où $m_1 = \mathrm{E}(X) = \int_a^b xf(x)\mathrm{d}x$, $m_2 = \mathrm{E}(Y) = \int_c^d yg(y)\mathrm{d}y$

• la covariance : $\mathrm{Cov}\,(X,Y) = \mu_{1,1} = \mathrm{E}[(X-m_1)(Y-m_2)]$

$$= \iint_D (x-m_1)\ (y-m_2)\mathrm{d}x\mathrm{d}y.$$

Formule de Koenig : $\mathrm{Cov}\,(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)E(Y)$.

2 Convergences en loi et convergence en probabilité

Soit une suite de couples (on dit aussi *vecteurs*) aléatoires (X_n,Y_n) qui suivent une loi bivariée dont la fonction de répartition est notée H_n : $H_n(t,u) = P(X_n \leqslant t \text{ et } Y_n \leqslant u)$.

2.1 Convergence en loi

On dit que la suite (X_n,Y_n) converge en loi vers le couple aléatoire (X,Y) lorsque, pour $n \longrightarrow \infty$, on a $H_n(t,u) \longrightarrow H(t,u) = P(X \leqslant t \text{ et } Y \leqslant u)$ en tout point (t,u) où H est continue. On écrit : $(X_n,Y_n) \xrightarrow{\text{loi}} (X,Y)$.

Soit un domaine δ de $\mathbb{R}^2$ tel que sa frontière $\partial\delta$ ait un poids nul pour la distribution H. Alors, $P[(X_n,Y_n) \in \delta] \longrightarrow P[(X,Y) \in \delta]$ lorsque $n \longrightarrow \infty$.

Théorème central limite. Soit une suite de couples aléatoires indépendants (X_1,Y_1), (X_2,Y_2), ... , (X_n,Y_n), ... qui suivent une même loi $\mathcal{L}(m,W)$ de valeur moyenne $m = (m_1,m_2)$ et dont la matrice des covariances W est inversible. Notons $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ et $\overline{Y}_n \dfrac{1}{n}\sum_{i=1}^{n} Y_i$. Lorsque $n \longrightarrow \infty$, la distribution du couple $\left(\sqrt{n}(\overline{X}_n - m_1), \sqrt{n}(\overline{Y}_n - m_2)\right)$ converge en loi vers $\mathcal{N}_2(\mathrm{O},W)$. *Cf.* § 3.

2.2 Convergence en probabilité

On dit que la suite (X_n,Y_n) converge en probabilité vers le couple de nombres réels (k_1,k_2) lorsque, $\forall\varepsilon > 0$ et pour $n \longrightarrow \infty$, on a $P(k_1 - \varepsilon \leqslant X_n \leqslant k_1 + \varepsilon$ et $k_2 - \varepsilon \leqslant Y_n \leqslant k_2 + \varepsilon) \longrightarrow 1$. On écrit : $(X_n,Y_n) \xrightarrow{\text{pr}} (k_1,k_2)$.

Propriété. (X_n,Y_n) converge en probabilité vers (k_1,k_2) si et seulement si X_n converge en probabilité vers k_1 et Y_n converge en probabilité vers k_2.

3 Lois normales bivariées

3.1 Lois normales centrées $\mathcal{N}_2(\mathrm{O};\, W)$ de dimension 2

Définition. W désignera la matrice des covariances de tout couple aléatoire de réels (X,Y) qui suit une loi continue dont le support dans $\mathbb{R}^2$ a un intérieur non vide :

$$W = \begin{bmatrix} \mathrm{V}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{V}(Y) \end{bmatrix} \text{ avec } W \text{ inversible.}$$

On dit qu'un couple aléatoire (X,Y) suit une loi normale centrée non dégénérée lorsqu'il peut prendre n'importe quelle valeur de $\mathbb{R}^2$ avec une densité de probabilité de type suivant :

$$\varphi(x,y) = \frac{\sqrt{\det.A}}{(\sqrt{2\pi})^2} \mathrm{e}^{-\frac{1}{2}q(x,y)} \quad \text{où} \quad q(x,y) = ax^2 + 2bxy + cy^2 \quad \text{avec} \quad a > 0,$$

$\det.A > 0$ où $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix} = W^{-1}$.

Théorème.

(i) X [resp. Y] suit une loi normale centrée. Plus généralement, $Z = \alpha X + \beta Y$ suit la loi normale centrée $\mathcal{N}_2(\mathrm{O};\, \mathrm{V}(Z))$ où $\mathrm{V}(Z) = (\alpha,\beta) \times W \times {}^t(\alpha,\beta)$

(ii) X et Y indépendants $\Longleftrightarrow \mathrm{Cov}(X,Y) = 0$.

(iii) Soit une matrice carrée inversible $B = \begin{bmatrix} \alpha & \gamma \\ \beta & \delta \end{bmatrix}$ à termes réels. Le couple (X',Y') défini par « $X' = \alpha X + \beta Y$, $Y' = \gamma X + \delta Y$ » suit la loi normale centrée $\mathcal{N}_2(\mathrm{O},W^*)$ où $W^* = {}^tB \times W \times B$. En effet $W^* = \mathrm{E}[{}^t(X',Y').(X',Y')]$ $= \mathrm{E}[{}^t B.{}^t(X,Y).(X,Y).B] = {}^tB.\mathrm{E}[{}^t(X,Y).(X,Y)].B = {}^t B \cdot W \cdot B$

3.2 Loi normale $\mathcal{N}_2(m,W)$ de dimension 2

On dit que (X^*,Y^*) suit la loi normale $\mathcal{N}_2(m,W)$ où $m = (m_1,m_2)$ est un couple de réels, lorsque $(X^* - m_1, Y^* - m_2)$ suit la loi $\mathcal{N}_2(\mathrm{O},W)$. On a évidemment $\mathrm{E}(X^*) = m_1$ et $\mathrm{E}(Y^*) = m_2$.

© Dunod. La photocopie non autorisée est un délit.

Une loi normale bivariée $\mathcal{N}_2(m,W)$ est caractérisée par une densité de probabilité du type $\varphi(x,y) = k\mathrm{e}^{-\frac{1}{2}q(x,y)}\ \forall (x,y) \in \mathbb{R}^2$ où k est une constante et $q(x,y) = ax^2 + 2bxy + cy^2 + 2\alpha x + 2\beta y$ avec $a > 0$, det. $A > 0$ où $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$.

Alors (m_1,m_2) est solution du système linéaire : $\dfrac{\partial q}{\partial x}(x,y) = 0$, $\dfrac{\partial q}{\partial y}(x,y) = 0$ et

$$A = W^{-1} = (1/2)\begin{bmatrix} \frac{\partial^2 q}{\partial x^2} & \frac{\partial^2 q}{\partial x \partial y} \\ \frac{\partial^2 q}{\partial y \partial x} & \frac{\partial^2 q}{\partial y^2} \end{bmatrix}.$$

Théorème de la somme. (X,Y) suit la loi $\mathcal{N}_2(m,W)$, (X',Y') suit la loi $\mathcal{N}_2(m',W')$, de plus (X,Y) et (X',Y') sont indépendants. Alors $(X+X',Y+Y')$ suit la loi $\mathcal{N}_2(m+m',W+W')$.

3.3 Régression de la *v.a. Y* sur la *v.a. X* dans le cas d'une distribution normale de dimension deux, non dégénérée (det.$W \neq 0$)

On a la décomposition $(Y - m_2)/\sigma(Y) = r(X - m_1)/\sigma(X) + \sqrt{1-r^2}\mathcal{N}_{0;1}$ où la variable normale $\mathcal{N}_{0;1}$ est indépendante de X.

Généralisation. Soit un triplet aléatoire (X_1,X_2,X_3) qui suit la loi normale non dégénérée $\mathcal{N}_3(m; W)$ où $m = (m_1,m_2,m_3)$ et W désigne la matrice des covariances. Si det.$W \neq 0$ notons a_{ij} la terme général de la matrice $A = W^{-1}$.

On a la décomposition : $X_3 = -(a_{13}/a_{33})X_1 - (a_{23}/a_{33})X_2 + (1/\sqrt{a_{33}})\mathcal{N}_{0;1}$ où la variable $\mathcal{N}_{0;1}$ est indépendante du couple (X_1,X_2).

3.4 Test d'indépendance et intervalles de confiance

Soit un échantillon de n couples aléatoires indépendants (X_1,Y_1), (X_2,Y_2), ... , (X_n,Y_n) qui suivent la loi normale non dégénérée $\mathcal{N}_2(m,W)$. On note $m_F = \dfrac{1}{2}\ln\dfrac{1+r}{1-r}$ où r désigne le coefficient de corrélation linéaire qui lie X et Y

$$R_n = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}}$$ et on considère la *v.a.* $Z = \dfrac{1}{2}\ln\dfrac{1+R_n}{1-R_n}$

appelée *transformée de Fisher*.

Pour des valeurs modérées de n ($n > 12$), Z suit sensiblement la loi normale de valeur moyenne $m_0 = m_F + r/2(n-1)$ et de variance $(n-3)^{-1}$: $\sqrt{n-3}(Z - m_0) \cong \mathcal{N}_{0;1}$. Cette propriété permet de trouver un intervalle de confiance de m_0 et donc de r.

Théorème. Sous l'hypothèse « *X et Y indépendants* » c.a.d. « $r = 0$ » on a $\dfrac{R_n\sqrt{n-2}}{\sqrt{1-R_n^2}} = \boldsymbol{t}_{n-2}$ (variable de Student), propriété qui permet de tester l'indépendance.

Section 2

MÉTHODES DU MAXIMUM DE VRAISEMBLANCE

1 Exemples introductifs

1/ Ayant lancé 50 fois une pièce de monnaie dans des conditions identiques, on obtient 30 fois pile et 20 fois face. Le nombre de fois X où l'on obtient pile suit la loi binomiale $\mathcal{B}(50,p)$ où $p =$ « probabilité d'obtenir pile à l'issue d'un lancé » : $P(X = 30) = \mathrm{C}_{50}^{30} p^{30}(1-p)^{20}$ noté $\varphi(p)$.

Cette probabilité sera maximale lorsque la dérivée $\varphi'(p)$ est nulle ou de façon équivalente lorsque la dérivée de $\ln[\varphi(p)] = \ln(\mathrm{C}_{50}^{30}) + 30\ln(p) + 20\ln(1-p)$ est nulle. On constate que l'estimateur optimal $\hat{p}$ de p prend la valeur 30/50. Autrement dit, la proportion aléatoire de succès F est, par ce procédé, l'estimateur optimal de p.

2/ Le résultat X d'une expérience suit une loi exponentielle non décalée de paramètre a : $P(0 \leqslant X \leqslant x) = 1 - \mathrm{e}^{-ax}$. Les résultats de 3 expériences réalisées dans des conditions identiques sont les suivants : $x_1 = 1{,}2$, $x_2 = 0{,}8$, $x_3 = 2{,}2$. Pour obtenir un estimateur $\hat{\mathbf{a}}$ de a par la méthode du maximum de vraisemblance on cherche la valeur $\hat{a}$ qui maximise $P(x_1 \leqslant X_1 < x_1 + \mathrm{d}x_1$ et $x_2 \leqslant X_2 < x_2 + \mathrm{d}x_2$ et $x_3 \leqslant X_3 < x_3 + \mathrm{d}x_3) = (a\,\mathrm{e}^{-ax_1}\mathrm{d}x_1)(a\,\mathrm{e}^{-ax_2}\mathrm{d}x_2)(a\,\mathrm{e}^{-ax_3}\mathrm{d}x_3)$ où $\mathrm{d}x_1, \mathrm{d}x_2$ et $\mathrm{d}x_3$ sont des constantes considérées comme infiniment petites. On constate que $\hat{a}$ maximise $a^3\,\mathrm{e}^{-a(x_1+x_2+x_3)}$ la densité de probabilité du triplet de *v.a.* indépendantes (X_1, X_2, X_3) et on en déduit en considérant la dérivée de $\ln(a^3\,\mathrm{e}^{-a(x_1+x_2+x_3)})$ que $\hat{a} = 3/(x_1+x_2+x_3) = 1/1{,}4$.

2 Estimateurs ponctuels par la méthode du maximum de vraisemblance

Soit un échantillon $X_1,...,X_n$ *iid* d'une loi continue $\mathcal{L}$ caractérisée par son support Δ qui est un intervalle et par sa densité de probabilité $f(x,\ell)$ où la valeur du paramètre ℓ est inconnue. Le n-uple $(X_1,...,X_n)$ a pour support Δ^n et pour densité de probabilité $\varphi(x_1,...,x_n;\ell) = f(x_1,\ell) \times f(x_2,\ell) \times ... \times f(x_n,\ell)$.

Connaissant les valeurs numériques $x_1,...,x_n$ prises par $X_1,...,X_n$, la méthode consiste à rechercher la valeur numérique ℓ^* qui maximise $\varphi(x_1,...,x_n;\ell)$. Elle est

© Dunod. La photocopie non autorisée est un délit.

solution de l'équation $\frac{\partial \varphi}{\partial \ell}(x_1,...,x_n;\ell^*) = 0$ ou de façon équivalente de $\frac{\partial}{\partial \ell}[\ln \varphi] = \sum_{i=1}^{n} \frac{\partial f(x_i,\ell)}{\partial \ell} / f(x_i,\ell) = 0.$

Lorsque $X_1,...,X_n$ est un échantillon *iid* d'une loi discrète $\mathcal{L}$ caractérisée par son support $\Delta = \{x_1^*, x_2^*, \ldots, x_{n_0}^*\}$ et $P(X = x_i^*) = p(x_i^*;\ell)$ on cherche la valeur ℓ^* qui maximise $\psi(x_1,...,x_n;\ell) = P(X_1 = x_1 \text{ et } ...X_n = x_n) = P(X_1 = x_1)...P(X_n = x_n)$ ou $\ln[\psi]$. Donc ℓ^* est solution de $\sum_{i=1}^{n} \frac{\partial p(x_i,\ell)}{\partial \ell} / p(x_i,\ell) = 0.$

3 Estimateurs efficaces

Les notations sont celles du paragraphe précédent et soit $\hat{\ell} = \theta(X_1,...,X_n)$ un estimateur sans biais de ℓ : $\mathrm{E}(\hat{\ell}) = \ell$. Sous certaines conditions de régularités on a

$$\mathrm{V}(\hat{\ell}) \geqslant 1/n\mathrm{E}\left[\left(\frac{\partial \ln[f(X,\ell)]}{\partial \ell}\right)^2\right] \text{ ou } \mathrm{V}(\hat{\ell}) \geqslant 1/n\Sigma_i\left[p(x_i^*,\ell)\left(\frac{\partial \ln[p(x_i^*,\ell)]}{\partial \ell}\right)^2\right]$$

(inégalité de Fréchet).
Un estimateur dont la variance est égale à la borne inférieure est dit efficace.

TABLES STATISTIQUES

	Pages
Fonction de répartition de la loi normale centrée réduite $\mathcal{N}_{(0;1)}$	374
Fonction de répartition de la loi binomiale $\mathcal{B}(m_0, p)$	375
Fonction de répartition de la loi de Poisson $\mathcal{P}(\lambda)$	376
Percentiles des lois khi-deux $\chi^2(n)$	377
Table de nombres au hasard	377
Percentiles de la loi de Student-Fisher $\mathcal{S}t(n)$	378
Seuil à 5 % de la distribution de Fisher-Snedecor	379
Valeurs critiques de la statistique $W_{n,n'}$ de Wilcoxon-Mann-Whitney	380
Valeurs critiques de la statistique rhô ρ_S de Spearman	381
Valeurs critiques de la statistique $W_{k,n}$ de Kendall pour k classements	382
Valeurs critiques de la statistique de Kruskal-Wallis KW pour trois échantillons	383
Distribution de la statistique W_n^+ des rangs signés de Wilcoxon	384
Valeurs critiques de la statistique J de Jonckheere-Terpstra (k échantillons)	385
Valeurs critiques de la statistique $F_{k,n}$ de Friedman	386
Valeurs critiques de la statistique D_n de Kolmogorov-Smirnov	386
Valeurs critiques du Skewness d'une distribution normale	387
Valeurs critiques du Kurtosis d'une distribution normale	388
Table de la statistique Durbin-Watson pour un seuil de signification 5 %	389

© Dunod. La photocopie non autorisée est un délit.

Fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0;1)$

Valeurs de $\Phi(t) = \mathcal{P}(\mathcal{N}_{0;1} \leqslant t)$
Pour t négatif utiliser $\mathcal{P}(\mathcal{N}_{0;1} \leqslant t) = 1 - \mathcal{P}(\mathcal{N}_{0;1} \leqslant -t)$

t	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8801	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8954	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9380	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474,	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649,	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Les valeurs de t exprimées avec une décimale se lisent sur la première colonne, les centièmes se lisent sur la première ligne. À l'intersection de la ligne et colonne on lit $\Phi(t)$. Ainsi pour $t = 1{,}34$ on lit 1,3 dans la première colonne et 0,04 dans la première ligne, à l'intersection de la ligne où figure 1,3 et de la colonne où figure 0,04 on lit $\Phi(1{,}34) = \mathcal{P}(\mathcal{N}_{0;1} \leqslant 1{,}34) = 0{,}9099$.
On en déduit $\Phi(-1{,}34) = 1 - \Phi(1{,}34) = 1 - 0{,}9099 = 0{,}0901$

Table pour les grandes valeurs de t

t	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0
$\Phi(t)$	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,99984	0,99993	0,99997

Fonction de répartition de la loi binomiale $\mathcal{B}(m_0, p)$

Pour divers choix de m_0 et de p, on lit les valeurs de $\mathcal{P}(\mathcal{B}_{m_0,p} \leqslant h)$.
Exemple : $\mathcal{P}(\mathcal{B}_{5,\,0,1} \leqslant 2) = 0{,}9914$
Lorsque $p > 0{,}5$ et donc $q = 1 - p < 0{,}5$ on peut utiliser $\mathcal{P}(\mathcal{B}_{m_0,p} \leqslant h) = 1 - \mathcal{P}(\mathcal{B}_{m_0,q} \leqslant n - h - 1)$

Lois binomiales avec $m_0 = 5$
$p = 0{,}01, 0{,}02, \ldots, 0{,}50$

h	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,25	0,50
0	0,9510	0,9039	0,8587	0,8154	0,7738	0,7339	0,6957	0,6591	0,6240	0,5905	0,2373	0,0313
1	0,9990	0,9962	0,9915	0,9852	0,9774	0,9681	0,9575	0,9456	0,9326	0,9185	0,6328	0,1875
2	1,0000	0,9999	0,9997	0,9994	0,9988	0,9980	0,9969	0,9955	0,9937	0,9914	0,8965	0,5000
3	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9998	0,9997	0,9995	0,9844	0,8125
4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9990	0,9688
5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Lois binomiales avec $m_0 = 10$
$p = 0{,}01, 0{,}02, \ldots, 0{,}50$

h	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,25	0,50
0	0,9044	0,8171	0,7374	0,6648	0,5987	0,5386	0,4840	0,4344	0,3894	0,3487	0,0563	0,0010
1	0,9957	0,9838	0,9655	0,9418	0,9139	0,8824	0,8483	0,8121	0,7746	0,7361	0,2440	0,0107
2	0,9999	0,9991	0,9972	0,9938	0,9885	0,9812	0,9717	0,9599	0,9460	0,9298	0,5256	0,0547
3	1,0000	1,0000	0,9999	0,9996	0,9990	0,9980	0,9964	0,9942	0,9912	0,9872	0,7759	0,1719
4	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9997	0,9994	0,9990	0,9984	0,9219	0,3770
5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9803	0,6230
6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9965	0,8281
7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9996	0,9453
8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9893
9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9990
10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Lois binomiales avec $m_0 = 15$
$p = 0{,}01, 0{,}02, \ldots, 0{,}50$

h	0.01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,1	0,25	0,50
0	0,8601	0,7386	0,6333	0,5421	0,4633	0,3953	0,3367	0,2863	0,2430	0,2059	0,0134	0,0000
1	0,9904	0,9647	0,9270	0,8809	0,8290	0,7738	0,7168	0,6597	0,6035	0,5490	0,0802	0,0005
2	0,9996	0,9970	0,9906	0,9797	0,9638	0,9429	0,9171	0,8870	0,8531	0,8159	0,2361	0,0037
3	1,0000	0,9998	0,9992	0,9976	0,9945	0,9896	0,9825	0,9727	0,9601	0,9444	0,4613	0,0176
4	1,0000	1,0000	0,9999	0,9998	0,9994	0,9986	0,9972	0,9950	0,9918	0,9873	0,6865	0,0592
5	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9997	0,9993	0,9987	0,9978	0,8516	0,1509
6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9997	0,9434	0,3036
7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9827	0,5000
8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9958	0,6964
9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9992	0,8491
10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9408
11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9824
12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9963
13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995
14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

© Dunod. La photocopie non autorisée est un délit.

Fonction de répartition de la loi de Poisson $\mathcal{P}(\lambda)$

Pour certains choix de λ on peut lire la valeur de $\mathcal{P}(\mathcal{P}_\lambda \leqslant h)$ où h est un entier.

Exemple. Pour $\lambda = 0{,}5$ et $h = 2$ on lit $\mathcal{P}(\mathcal{P}_\lambda \leqslant 2) = 0{,}9856$

$\lambda = 0{,}1,\ 0{,}2,\ \ldots,\ 0{,}9$

h	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066
1	0,9953	0,9825	0,9631	0,9384	0,9098	0,8781	0,8442	0,8088	0,7725
2	0,9998	0,9989	0,9964	0,9921	0,9856	0,9769	0,9659	0,9526	0,9371
3	1	0,9999	0,9997	0,9992	0,9982	0,9966	0,9942	0,9909	0,9865
4	1	1	1	0,9999	0,9998	0,9996	0,9992	0,9986	0,9977
5	1	1	1	1	1	1	0,9999	0,9998	0,9997
6	1	1	1	1	1	1	1	1	1

$\lambda = 1,\ 1{,}5,\ \ldots,5$

h	1	1,5	2	2,5	3	3,5	4	4,5	5
0	0,3679	0,2231	0,1353	0,0821	0,0498	0,0302	0,0183	0,0111	0,0067
1	0,7358	0,5578	0,4060	0,2873	0,1991	0,1359	0,0916	0,0611	0,0404
2	0,9197	0,8088	0,6767	0,5438	0,4232	0,3208	0,2381	0,1736	0,1247
3	0,9810	0,9344	0,8571	0,7576	0,6472	0,5366	0,4335	0,3423	0,2650
4	0,9963	0,9814	0,9473	0,8912	0,8153	0,7254	0,6288	0,5321	0,4405
5	0,9994	0,9955	0,9834	0,9580	0,9161	0,8576	0,7851	0,7029	0,6160
6	0,9999	0,9991	0,9955	0,9858	0,9665	0,9347	0,8893	0,8311	0,7622
7	1	0,9998	0,9989	0,9958	0,9881	0,9733	0,9489	0,9134	0,8666
8	1	1	0,9998	0,9989	0,9962	0,9901	0,9786	0,9597	0,9319
9	1	1	1	0,9997	0,9989	0,9967	0,9919	0,9829	0,9682
10	1	1	1	0,9999	0,9997	0,9990	0,9972	0,9933	0,9863
11	1	1	1	1	0,9999	0,9997	0,9991	0,9976	0,9945
12	1	1	1	1	1	0,9999	0,9997	0,9992	0,9980
13	1	1	1	1	1	1	0,9999	0,9997	0,9993
14	1	1	1	1	1	1	1	0,9999	0,9998
15	1	1	1	1	1	1	1	1	0,9999
16	1	1	1	1	1	1	1	1	1

$\lambda = 5{,}5\ ,\ 6,\ \ldots,\ 8$

h	5,5	6	6,5	7	7,5	8
0	0,0041	0,0025	0,0015	0,0009	0,0006	0,0003
1	0,0266	0,0174	0,0113	0,0073	0,0047	0,003
2	0,0884	0,0620	0,0430	0,0296	0,0203	0,0138
3	0,2017	0,1512	0,1118	0,0818	0,0591	0,0424
4	0,3575	0,2851	0,2237	0,1730	0,1321	0,0996
5	0,52890	0,4457	0,3690	0,3007	0,2414	0,1912
6	0,6860	0,6063	0,5265	0,4497	0,3782	0,3134
7	0,8095	0,7440	0,6728	0,5987	0,5246	0,4530
8	0,8944	0,8472	0,7916	0,7291	0,6620	0,5925
9	0,9462	0,9161	0,8774	0,8305	0,7764	0,7166
10	0,9747	0,9574	0,9332	0,9015	0,8622	0,8159
11	0,9890	0,9799	0,9661	0,9467	0,9208	0,8881
12	0,9955	0,9912	0,9840	0,9730	0,9573	0,9362
13	0,9983	0,9964	0,9929	0,9872	0,9784	0,9658
14	0,9994	0,9986	0,9970	0,9943	0,9897	0,9827
15	0,9998	0,9995	0,9988	0,9976	0,9954	0,9918
16	0,9999	0,9998	0,9996	0,9990	0,998	0,9963
17	1	0,9999	0,9998	0,9996	0,9992	0,9984
18	1	1	0,9999	0,9999	0,9997	0,9993
19	1	1	1	1	0,9999	0,9997
20	1	1	1	1	1	0,9999
21	1	1	1	1	1	1

Table des percentiles des lois khi-deux $\chi^2(n)$

La table donne pour diverses valeurs de n et pour divers choix de α, la valeur de t correspondant à $\alpha = \mathcal{P}(\chi^2_n \geqslant t)$.

Exemple. On lit $0{,}99 = \mathcal{P}(\chi^2_7 \geqslant 1{,}24)$. Inversement on a $1 - \alpha = \mathcal{P}(\chi^2_n \leqslant t)$ soit $0{,}01 = \mathcal{P}(\chi^2_7 \leqslant 1{,}24)$

$\alpha \longrightarrow$ / $n \downarrow$	0,99	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0002	0,003	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84	5,41	6,64	10,83
2	0,02	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	7,82	9,21	13,82
3	0,12	0,35	0,58	1,01	1,42	2,37	3,67	4,64	6,25	7,82	9,84	11,34	16,27
4	0,30	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	11,67	13,28	18,47
5	0,55	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	13,39	15,09	20,51
6	0,87	1,64	2,20	3,07	3,83	5,35	7,23	8,56	10,65	12,59	15,03	16,81	22,46
7	1,24	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	16,62	18,48	24,32
8	1,65	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51	18,17	20,09	26,12
9	2,09	3,33	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	19,68	21,67	27,88
10	2,56	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31	21,16	23,21	29,59
11	3,05	4,58	5,58	6,99	8,15	10,34	12,90	14,63	17,28	19,68	22,62	24,73	31,26
12	3,57	5,23	6,30	7,81	9,03	11,34	14,01	15,81	18,55	21,03	24,05	26,22	32,91
13	4,11	5,89	7,04	8,63	9,93	12,34	15,12	16,98	19,81	22,36	25,47	27,69	34,53
14	4,66	6,57	7,79	9,47	10,82	13,34	16,22	18,15	21,06	23,69	26,87	29,14	36,12
15	5,23	7,26	8,55	10,31	11,72	14,34	17,32	19,31	22,31	25,00	28,26	30,58	37,70
16	5,81	7,96	9,31	11,15	12,62	15,34	18,42	20,47	23,54	26,30	29,63	32,00	39,25
17	6,41	8,67	10,09	12,00	13,53	16,34	19,51	21,61	24,77	27,59	31,00	33,41	40,79
18	7,00	9,39	10,87	12,86	15,44	17,34	20,60	22,76	25,99	28,87	32,35	34,81	42,31
19	7,63	10,12	11,65	13,72	15,35	18,34	21,69	23,90	27,20	30,14	33,69	36,19	43,82
20	8,26	10,85	12,44	14,58	16,27	19,34	22,78	25,04	28,41	31,41	35,02	37,57	45,31
22	9,54	12,34	14,04	16,31	18,10	21,34	24,94	27,30	30,81	33,92	37,66	40,29	48,27
24	10,86	13,85	15,66	18,06	19,94	23,34	27,10	29,55	33,20	36,42	40,27	42,98	51,18
26	12,20	15,38	17,29	19,82	21,79	25,34	29,25	31,79	35,56	38,89	42,86	45,64	54,05
28	13,57	16,93	18,94	21,59	23,65	27,34	31,39	34,03	37,92	41,34	45,42	48,28	56,89
30	14,95	18,49	20,60	23,36	25,51	29,34	33,53	36,25	40,26	43,77	47,96	50,89	59,70
32	16,36	20,07	22,27	25,15	27,37	31,34	35,67	38,47	42,59	46,19	50,49	53,49	62,49
34	17,79	21,66	23,95	26,94	29,24	33,34	37,80	40,68	44,90	48,60	53,00	56,06	65,25
36	19,23	23,27	25,64	28,73	31,11	35,34	39,92	42,88	47,21	51,00	55,49	58,62	67,98
38	20,69	24,88	27,34	30,54	32,99	37,34	42,04	45,08	49,51	53,38	57,97	61,16	70,70
40	22,16	26,51	29,05	32,34	34,87	39,34	44,17	47,27	51,81	55,76	60,44	63,69	73,40

Lorsque $n > 30$, utiliser l'approximation normale $\sqrt{2\chi^2_n} - \sqrt{2n-1} \cong \mathcal{N}_{0;1}$

Table de nombres au hasard

13407	62899	78937	90525	25033	56358	78902	47008	72488
50270	63237	94083	93634	71652	02656	57532	60307	91619
84980	62458	09703	78397	66179	46982	67619	39254	90763
22116	33646	17545	31321	65772	86506	09811	82848	92211
68645	15068	56898	87021	40115	27524	42221	88293	67592
26518	39122	96561	56004	50260	68648	85591	83979	09041
35493	41666	27871	71329	69212	57932	65281	57233	07732
77402	12994	59892	85581	70823	53338	34405	67080	16568
83679	97154	40341	84741	08967	73287	94952	59008	95774
71802	39356	02981	89107	79788	51330	37129	31898	34011
57494	72484	22676	44311	15356	05348	03582	66183	68392
73364	38416	93128	10297	11419	82937	84389	88273	96010
14499	83965	75403	18002	45068	54257	18085	92625	60911
40747	03084	07734	88940	88722	85717	73810	79866	84853
42237	59122	92855	62097	81276	06318	81607	00565	56626

© Dunod. La photocopie non autorisée est un délit.

Percentiles de la loi de Student-Fisher $\mathcal{S}t(n)$

La table donne pour diverses valeurs de n et pour divers choix de α, la valeur de t correspondant à $1-\alpha = \mathcal{P}(\boldsymbol{t}_n \leqslant t)$. Inversement on en déduit que $\alpha = \mathcal{P}(\boldsymbol{t}_n \geqslant t)$.
Pour t négatif utiliser $\mathcal{P}(\boldsymbol{t}_n \leqslant t) = 1 - \mathcal{P}(\boldsymbol{t}_n \leqslant -t)$
Exemple. On lit $0{,}99 = \mathcal{P}(\boldsymbol{t}_2 \leqslant 6{,}965)$.
On en déduit que $\mathcal{P}(\boldsymbol{t}_2 \leqslant -6{,}965) = 1 - \mathcal{P}(\boldsymbol{t}_2 \leqslant 6{,}965) = 0{,}01$

$1-\alpha$ / $n\downarrow$	0,5	0,6	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,000	0,325	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	318,3	636,6
2	0,000	0,289	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	22,33	31,6
3	0,000	0,277	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	0,000	0,271	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,000	0,267	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	0,000	0,265	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,000	0,263	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	0,000	0,262	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,000	0,261	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,000	0,260	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,000	0,260	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,000	0,259	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,000	0,259	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,000	0,258	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,000	0,258	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,000	0,258	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,000	0,257	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,000	0,257	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	0,000	0,257	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,000	0,257	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,000	0,257	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,000	0,256	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,000	0,256	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	0,000	0,256	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,000	0,256	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,000	0,256	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,000	0,256	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,000	0,256	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,000	0,256	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,000	0,256	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,385	3,646

Pour $n > 120$ on utilise l'approximation $\boldsymbol{t}_n \cong \mathcal{N}_{0,1}$, autrement dit pour tout t réel :

$$\mathcal{P}(\boldsymbol{t}_n \leqslant t) \cong \mathcal{P}(\mathcal{N}_{0,1} \leqslant t) = \Phi(t).$$

Table du seuil à 5 % de la distribution de Fisher-Snedecor

La table donne pour diverses valeurs de m et n pour $\alpha = 0{,}05$, la valeur de t correspondant à $\alpha = \mathcal{P}(F_n^m \geqslant t) = 0{,}05$. On a $\mathcal{P}(F_n^m \leqslant t') = \mathcal{P}(F_m^n \geqslant 1/t')$.
Exemple. Pour $m = 4$ et $n = 5$ on lit $\mathcal{P}(F_5^4 \geqslant 5{,}19) = 0{,}05$ ou évidemment $\mathcal{P}(F_4^5 \leqslant 1/5{,}19) = 0{,}05$

m / n	1	2	3	4	5	6	8	9	10	11	12	13	20	30	38	44	52
1	161	200	216	225	230	234	239	241	242	243	244	245	248	250	251	251	252
2	18,5	19,0	19,2	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,84	8,81	8,79	8,76	8,74	8,73	8,66	8,62	8,60	8,59	8,58
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	6,00	5,96	5,94	5,91	5,89	5,80	5,75	5,72	5,71	5,70
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,77	4,74	4,70	4,68	4,66	4,56	4,50	4,47	4,46	4,44
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,10	4,06	4,03	4,00	3,98	3,87	3,81	3,78	3,76	3,75
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,68	3,64	3,60	3,57	3,55	3,44	3,38	3,35	3,33	3,32
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,39	3,35	3,31	3,28	3,26	3,15	3,08	3,05	3,03	3,02
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,18	3,14	3,10	3,07	3,05	2,94	2,86	2,83	2,82	2,80
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	3,02	2,98	2,94	2,91	2,89	2,77	2,70	2,67	2,65	2,63
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,90	2,85	2,82	2,79	2,76	2,65	2,57	2,54	2,52	2,50
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,80	2,75	2,72	2,69	2,66	2,54	2,47	2,43	2,41	2,40
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,71	2,67	2,63	2,60	2,58	2,46	2,38	2,35	2,33	2,31
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,65	2,60	2,57	2,53	2,51	2,39	2,31	2,27	2,25	2,24
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,59	2,54	2,51	2,48	2,45	2,33	2,25	2,21	2,19	2,17
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,54	2,49	2,46	2,42	2,40	2,28	2,19	2,16	2,14	2,12
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,49	2,45	2,41	2,38	2,35	2,23	2,15	2,11	2,09	2,07
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,46	2,41	2,37	2,34	2,31	2,19	2,11	2,07	2,05	2,03
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,42	2,38	2,34	2,31	2,28	2,16	2,07	2,03	2,01	1,99
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,39	2,35	2,31	2,28	2,25	2,12	2,04	2,00	1,98	1,96
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,37	2,32	2,28	2,25	2,22	2,10	2,01	1,97	1,95	1,93
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,34	2,30	2,26	2,23	2,20	2,07	1,98	1,95	1,93	1,90
23	4,28	3,42	3,03	2,08	2,64	2,53	2,38	2,32	2,27	2,24	2,20	2,18	2,05	1,96	1,92	1,90	1,88
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,30	2,25	2,22	2,18	2,15	2,03	1,94	1,90	1,88	1,86
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,28	2,24	2,20	2,16	2,14	2,01	1,92	1,88	1,86	1,84
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,27	2,22	2,18	2,15	2,12	1,99	1,90	1,86	1,84	1,82
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,25	2,20	2,17	2,13	2,10	1,97	1,88	1,84	1,82	1,80
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,24	2,19	2,15	2,12	2,09	1,96	1,87	1,83	1,81	1,79
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,22	2,18	2,14	2,10	2,08	1,94	1,85	1,81	1,79	1,77
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,21	2,16	2,13	2,09	2,06	1,93	1,84	1,80	1,78	1,76
36	4,11	3,26	2,87	2,63	2,48	2,36	2,21	2,15	2,11	2,07	2,03	2,00	1,87	1,78	1,73	1,71	1,69
38	4,10	3,24	2,85	2,62	2,46	2,35	2,19	2,14	2,09	2,05	2,02	1,99	1,85	1,76	1,72	1,69	1,67
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,12	2,08	2,04	2,00	1,97	1,84	1,74	1,70	1,68	1,65
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	2,04	1,99	1,95	1,92	1,89	1,75	1,65	1,60	1,58	1,55
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,96	1,91	1,87	1,83	1,80	1,66	1,55	1,50	1,48	1,45

© Dunod. La photocopie non autorisée est un délit.

Valeurs critiques de la statistique $W_{n,n'}$ de Wilcoxon pour deux échantillons

Valeur critique inférieure w_α. Pour $\alpha = 0{,}01$, $0{,}05$, et $0{,}10$ on trouve sur la table le plus grand entier w_α tel que $\mathcal{P}_0(W_{n,n'} \leqslant w_\alpha) \leqslant \alpha$.
Valeur critique supérieure w'_α : $w'_\alpha = 2W - w_\alpha$ est le plus petit entier tel que $\mathcal{P}_0(W_{n,n'} \geqslant w'_\alpha) \leqslant \alpha$.
Exemple. Pour $n = 4$ et $n' = 5$ on lit $\mathcal{P}_0(W_{4,5} \leqslant 12) \leqslant 0{,}05$,
donc $w'_\alpha = 2W - w_\alpha = 40 - 12 = 28$ et par suite $\mathcal{P}_0(W_{4,5} \geqslant 28) \leqslant 0{,}05$.

	$n=3$				$n=4$				$n=5$				$n=6$				$n=7$				$n=8$			
	0,01	0,05	0,1	$2W$	0,01	0,05	0,1	$2W$	0,01	0,05	0,1	2W	0,01	0,05	0,1	$2W$	0,01	0,05	0,1	2W	0,01	0,05	0,1	$2W$
$n' \downarrow$																								
3		6	7	21																				
4		6	7	24		11	13	36																
5		7	8	27	10	12	14	40	16	19	20	55												
6		8	9	30	11	13	15	44	17	20	22	60	24	28	30	78								
7	6	8	10	33	11	14	16	48	18	21	23	65	25	29	32	84	34	39	41	105				
8	6	9	11	36	12	15	17	52	19	23	25	70	27	31	34	90	35	41	44	112	45	51	55	136
9	7	10	11	39	13	16	19	56	20	24	27	75	28	33	36	96	37	43	46	119	47	54	58	144
10	7	10	12	42	13	17	20	60	21	26	28	80	29	35	38	102	39	45	49	126	49	56	60	152
11	7	11	13	45	14	18	21	64	22	27	30	85	30	37	40	108	40	47	51	133	51	59	63	160
12	8	11	14	48	15	19	22	68	23	28	32	90	32	38	42	114	42	49	54	140	53	62	66	168
13	8	12	15	51	15	20	23	72	24	30	33	95	33	40	44	120	44	52	56	147	56	64	69	176
14	8	13	16	54	16	21	25	76	25	31	35	100	34	42	46	126	45	54	59	154	58	67	72	184
15	9	13	16	57	17	22	26	80	26	33	37	105	36	44	48	132	47	56	61	161	60	69	75	192
16	9	14	17	60	17	24	27	84	27	34	38	110	37	46	50	138	49	58	64	168	62	72	78	200
17	10	15	18	63	18	25	28	88	28	35	40	115	39	47	52	144	51	61	66	175	64	75	81	208
18	10	15	19	66	19	26	30	92	29	37	42	120	40	49	55	150	52	63	69	182	66	77	84	216
19	10	16	20	69	19	27	31	96	30	38	43	125	41	51	57	156	54	65	71	189	68	80	87	224
20	11	17	21	72	20	28	32	100	31	40	45	130	43	53	59	162	56	67	74	196	70	83	90	232
21	11	17	21	75	21	29	33	104	32	41	47	135	44	55	61	168	58	69	76	203	72	85	92	240
22	12	18	22	78	21	30	35	108	33	43	48	140	45	57	63	174	59	72	79	210	74	88	95	248
23	12	19	23	81	22	31	36	112	34	44	50	145	47	58	65	180	61	74	81	217	76	90	98	256
24	12	19	24	84	23	32	38	116	35	45	51	150	48	60	67	186	63	76	84	224	78	93	101	264
25	13	20	25	87	23	33	38	120	36	47	53	155	50	62	69	192	64	78	86	231	81	96	104	272

	$n=9$				$n=10$				$n=11$				$n=12$				$n=13$				$n=14$			
$n' \downarrow$	0,01	0,05	0,1	$2W$	0,01	0,05	0,1	$2W$	0,01	0,05	0,1	2W	0,01	0,05	0,1	$2W$	0,01	0,05	0,1	2W	0,01	0,05	0,1	$2W$
9	59	66	70	171																				
10	61	69	73	180	74	82	87	210																
11	63	72	76	189	77	86	91	220	91	100	106	253												
12	66	75	80	198	79	89	94	230	94	104	110	264	109	120	127	300								
13	68	78	83	207	82	92	98	240	97	108	114	275	113	125	131	312	130	142	149	351				
14	71	81	86	216	85	96	102	250	100	112	118	286	116	129	136	324	134	147	154	364	152	166	174	406
15	73	84	90	225	88	99	106	260	103	116	123	297	120	133	141	336	138	152	159	377	156	171	179	420
16	76	87	93	234	91	103	109	270	107	120	127	308	124	138	145	348	142	156	165	390	161	176	185	434
17	78	90	97	243	93	106	113	280	110	123	131	319	127	142	150	360	146	161	170	403	165	182	190	448
18	81	93	100	252	96	110	117	290	113	127	135	330	131	146	155	372	150	166	175	416	170	187	196	462
19	83	96	103	261	99	113	121	300	116	131	139	341	134	150	159	384	154	171	180	429	174	192	202	476
20	85	99	107	270	102	117	125	310	119	135	144	352	138	155	164	396	158	175	185	442	178	197	207	490
21	88	102	110	279	105	120	128	320	123	139	148	363	142	159	169	408	162	180	190	455	183	202	213	504
22	90	105	113	288	108	123	132	330	126	143	152	374	145	163	173	420	166	185	195	468	187	207	218	518
23	93	108	117	297	110	127	136	340	129	147	156	385	149	168	178	432	170	189	200	481	192	212	224	532
24	95	111	120	306	113	130	140	350	132	151	161	396	153	172	183	444	174	194	205	494	196	218	229	546
25	98	114	123	315	116	134	144	360	136	155	165	407	156	176	187	456	178	199	211	507	200	223	235	560

Valeurs critiques de la statistique ρ_S de Spearman en fonction de n

Pour $4 \leqslant n \leqslant 30$ on lit ci-dessous la valeur critique c_α telle que $\mathcal{P}_0(\rho_S \geqslant c_\alpha) \cong \alpha$. Notons que la valeur critique c'_α telle que $\mathcal{P}_0(\rho_S \leqslant -c'_\alpha) \cong \alpha$ est égale à $-c_\alpha$.

Exemple. Pour $n = 10$ et $\alpha = 0{,}05$ on lit $\mathcal{P}_0(\rho_S \geqslant 0{,}564) \leqslant 0{,}05$ et donc $\mathcal{P}_0(\rho_S \leqslant -0{,}564) \leqslant 0{,}05$

$n \downarrow \alpha \longrightarrow$	0,25	0,1	0,05	0,025	0,01	0,005	0,0025	0,001
4	0,6	1	1					
5	0,500	0,800	0,900	1	1			
6	0,371	0,657	0,829	0,886	0,943	1	1	
7	0,321	0,571	0,714	0,786	0,893	0,929	0,964	1
8	0,310	0,524	0,643	0,738	0,833	0,881	0,905	0,95
9	0,267	0,483	0,600	0,700	0,783	0,833	0,867	0,92
10	0,248	0,455	0,564	0,648	0,745	0,794	0,830	0,88
11	0,236	0,427	0,536	0,618	0,709	0,755	0,800	0,85
12	0,224	0,406	0,503	0,587	0,671	0,727	0,776	0,83
13	0,209	0,385	0,484	0,560	0,648	0,703	0,747	0,80
14	0,200	0,367	0,464	0,538	0,622	0,675	0,723	0,78
15	0,189	0,354	0,443	0,521	0,604	0,654	0,700	0,75
16	0,182	0,341	0,429	0,503	0,582	0,635	0,679	0,73
17	0,176	0,328	0,414	0,485	0,566	0,615	0,662	0,71
18	0,170	0,317	0,401	0,472	0,550	0,600	0,643	0,70
19	0,165	0,309	0,391	0,460	0,535	0,584	0,628	0,68
20	0,161	0,299	0,380	0,447	0,520	0,570	0,612	0,66
21	0,156	0,292	0,370	0,435	0,508	0,556	0,599	0,65
22	0,152	0,284	0,361	0,425	0,496	0,544	0,586	0,63
23	0,148	0,278	0,353	0,415	0,486	0,532	0,573	0,62
24	0,144	0,271	0,344	0,406	0,476	0,521	0,562	0,61
25	0,142	0,265	0,337	0,398	0,466	0,511	0,551	0,60
26	0,138	0,259	0,331	0,390	0,457	0,501	0,541	0,59
27	0,136	0,255	0,324	0,382	0,448	0,491	0,531	0,58
28	0,133	0,250	0,317	0,375	0,440	0,483	0,522	0,57
29	0,130	0,245	0,312	0,368	0,433	0,475	0,513	0,56
30	0,128	0,240	0,306	0,362	0,425	0,467	0,504	0,55

Pour $n > 30$ utiliser l'approximation normale $\sqrt{n-1}\rho_S \cong \mathcal{N}_{0;1}$ avec correction de continuité sur $D_{S,n}$.

© Dunod. La photocopie non autorisée est un délit.

Valeurs critiques de la statistique $W_{k,n}$ de Kendall pour k classements

Valeurs critiques w de $W_{k,n}$ pour le niveau de signification $\alpha = 0.05$

Exemple. Pour $n = 5$ et $k = 3$ $\alpha = 0.05$ on lit $P_0(W_{k,n} \geqslant 0{,}716) \leqslant 0.05$

$n \downarrow k \longrightarrow$	3	4	5	6
3	1	0.750	0.600	0.500
4	0.822	0.619	0.500	0.421
5	0.716	0.553	0.449	0.377
6	0.660	0.512	0.418	0.351
7	0.626	0.484	0.395	0.332
8	0.595	0.461	0.378	0.319
9	0.576	0.447	0.365	0.307
10	0.560	0.434	0.354	0.299
11	0.548	0.425	0.346	0.287
12	0.535	0.415	0.336	0.287
13	0.527	0.409	0.332	0.280
14	0.520	0.402	0.327	0.275
15	0.514	0.395	0.322	0.272
20	0.49	0.37	0.30	0.25
40	0.43	0.33	0.26	0.22
60	0.41	0.31	0.25	0.21
100	0.38	0.29	0.24	0.20
∞	0.33	0.25	0.20	0.17

Lorsque $k \geqslant 7$, la statistique $(k-1)W_{k,n}/(1-W_{k,n})$ suit sensiblement la loi de Fisher dont les degrés de liberté sont respectivement $\nu_1 = (n-1) - 2/k$ et $\nu_2 = (k-1)\nu_1$:

Lorsque $k \geqslant 20$ on utilise l'approximation $k(n-1)W_{k,n} \approx \chi^2_{n-1}$

Valeurs critiques de la statistique KW de Kruskal-Wallis

Valeurs critiques pour la statistique KW pour $k = 3$ échantillons.
On lit ci-dessous la valeur critique c_α telle que $\alpha \cong \mathcal{P}_0(KW \geqslant c_\alpha)$

Exemple. Pour trois échantillons de taille respective 4, 2 et 1 on lit $0,1 \cong \mathcal{P}_0(KW \geqslant 4,5)$

Tailles des échantillons			α				
n_1	n_2	n_3	0,10	0,05	0,01	005	0,001
2	2	2	4,25				
3	2	1	4,29				
3	2	2	4,71	4,71			
3	3	1	4,57	5,14			
3	3	2	4,56	5,36			
3	3	3	4,62	5,6	7,2	7,2	
4	2	1	4,5				
4	2	2	4,46	5,33			
4	3	1	4,06	5,22			
4	3	2	4,51	5,44	6,44	7	
4	3	3	4,71	5,73	6,75	7,32	8,02
4	4	1	4,17	4,97	6,67		
4	4	2	4,55	5,45	7,04	7,28	
4	4	3	4,55	5,6	7,14	7,59	8,32
4	4	4	4,65	5,69	7,66	8	8,65

Pour des tailles d'échantillon $n_i \geqslant 4 \; \forall i = 1,2,...,k$ utiliser $KW \cong \chi^2_{(k-1)}$ où k est le nombre d'échantillons.

© Dunod. La photocopie non autorisée est un délit.

Distribution de la statistique W_n^+ des rangs signés de Wilcoxon

Exemple. Lire $P_0(W_3^+ \geqslant 4) = 0.375$.

On a $P_0(W_3^+ \leqslant h) = P_0(W_3^+ \geqslant n(n+1)/2 - h)$

$w \downarrow n \longrightarrow$	3	4	5	6	7	8	9	10	11	$w \downarrow n \longrightarrow$	12	13	14	15
3	.625									39	.515			
4	.375									40	.485			
5	.250	.562								41	.455			
6	.125	.437								42	.425			
7		.312								43	.396			
8		.187	.500							44	.367			
9		.125	.406							45	.339			
10		.062	.312							46	.311	.500		
11			.219	.500						47	.285	.473		
12			.156	.422						48	.259	.446		
13			.094	.344						49	.235	.420		
14			.062	.281	.531					50	.212	.393		
15			.031	.219	.469					51	.190	.368		
16				.156	.406					52	.170	.342		
17				.109	.344					53	.151	.318	.500	
18				.078	.289	.527				54	.133	.294	.476	
19				.047	.234	.473				55	.117	.271	.452	
20				.031	.187	.422				56	.102	.249	.428	

Pour $n > 20$ utiliser l'approximation

$P(W_n^+ \leqslant h) \cong \Phi\left[(h + 0.5 - n(n+1)/4)/\sqrt{n(n+1)(2n+1)/24}\right]$

Valeurs critiques de la statistique de Jonckheere-Terpstra pour k échantillons

Valeurs critiques de la statistique J de Jonckheere-Terpstra pour k échantillons où n_i désigne la taille du i-ème échantillon. Les échantillons étant ordonnés par rang de taille $(n_1 \leqslant n_2 \leqslant n_3 ...)$ on lit $\mathcal{P}_0(J \geqslant \theta_\alpha) \cong \alpha$.

Exemple. Pour $n_1 = 2$, $n_2 = 3$ et $n_4 = 4$ on lit $\mathcal{P}_0(J \geqslant 20) \leqslant 0{,}1$

$n_i \downarrow$ $\alpha \longrightarrow$	0,1	0,05	0,01	0,005	$n_i \downarrow$ $\alpha \longrightarrow$	0,1	0,05	0,01	0,005
222	10	11	12		337	36	38	42	44
223	13	14	15	16	338	40	42	47	49
224	16	17	19	20	344	29	31	34	36
225	18	20	22	23	345	33	35	39	41
226	21	23	25	27	346	38	40	44	46
227	24	26	29	30	347	42	45	49	52
228	27	29	32	33	348	47	50	55	57
233	16	18	19	20	355	38	41	45	47
234	20	21	23	25	356	43	46	51	53
235	23	25	27	29	357	48	51	57	59
236	26	28	31	33	358	53	57	63	65
237	30	32	35	37	366	49	52	57	60
239	33	35	39	41	367	54	58	64	67
244	24	25	28	29	368	60	64	70	73
245	27	29	33	34	377	61	64	71	74
246	31	34	37	39	378	67	71	78	81
247	35	38	42	44	388	74	78	86	89
248	39	42	46	49	444	34	36	40	42
255	32	34	38	40	445	39	41	45	48
256	36	39	43	45	446	44	47	51	54
257	41	44	48	51	447	49	52	57	60
258	45	48	53	56	448	54	57	63	66
266	42	44	49	51	455	44	47	52	55
267	47	50	55	57	456	50	53	58	61
268	52	55	61	64	457	56	59	65	68
277	52	56	61	64	458	61	65	71	75
278	58	62	68	71	466	56	60	66	69
288	64	68	75	78	467	62	66	73	76
333	20	22	24	25	468	68	73	80	83
334	24	26	29	30	477	69	73	81	84
335	28	30	33	35	478	76	80	88	92
336	32	34	38	40	488	83	88	97	100

© Dunod. La photocopie non autorisée est un délit.

Valeurs critiques de la statistique de Friedman $F_{k,n}$

Pour k échantillons et n facteurs on lit ci-dessous la valeur critique c_α telle que $\mathcal{P}_0(F_{k,n} \geqslant c_\alpha) \leqslant \alpha$

Exemple. Pour $k = 4$ échantillons et $n = 5$ facteurs on lit $\mathcal{P}_0(F_{4,5} \geqslant 7{,}8) \leqslant 0{,}05$

$k=3$			$k=4$			$k=5$			$k=6$		
$n \downarrow \alpha \longrightarrow$	0,05	0,01	$n \downarrow \alpha \longrightarrow$	0,05	0,01	$n \downarrow \alpha \longrightarrow$	0,05	0,01	$n \downarrow \alpha \longrightarrow$	0,05	0,01
2	-	-	2	6,000	-	2	7,600	8,000	2	9,357	9,929
3	6,000	-	3	7,400	9,000	3	8,533	10,13	3	9,857	11,76
4	6,500	8,000	4	7,800	9,600	4	8,800	11,20	4	10,39	12,82
5	6,400	8,400	5	7,800	9,960	5	8,960	11,68	5	7,680	10,68
6	7,000	9,000	6	7,600	10,20	6	9,067	11,87	6	7,691	10,75
7	7,143	8,857	7	7,800	10,54	7	9,143	12,11	7	7,700	10,80
8	6,250	9,000	8	7,650	10,50	8	9,200	12,30	8	7,800	10,85
9	6,222	9,556	9	7,667	10,73	9	9,244	12,44	9	7,714	10,89
10	6,200	9,600	10	7,680	10,68						
11	6,545	9,455	11	7,691	10,75						
12	6,500	9,500	12	7,700	10,80						
13	6,615	9,385	13	7,800	10,85						
14	6,143	9,143	14	7,714	10,89						
15	6,400	8,933	15	7,720	10,92						
20	6,300	9,300	20	7,800	11,10						

Statistique D_n de Kolmogorov-Smirnov pour un échantillon de taille n

Valeurs critiques d_α de $D_n = \sup_t\{|G_n(t) - F(t)|\}$
On lit ci-dessous la valeur critique d_α telle que $\alpha \cong \mathcal{P}_0(D_n \geqslant d_\alpha)$

Exemple. Lire pour un échantillon de taille $n = 5$: $\mathcal{P}_0(D_5 \geqslant 0{,}565) \leqslant 0{,}05$

$n \downarrow \alpha \longrightarrow$	0,20	0,15	0,1	0,05	0,01
1	0,9	0,925	0,95	0,975	0,995
2	0,684	0,726	0,776	0,842	0,929
3	0,565	0,597	0,642	0,706	0,828
4	0,494	0,525	0,564	0,624	0,733
5	0,446	0,474	0,51	0,565	0,669
6	0,41	0,436	0,47	0,521	0,618
7	0,381	0,405	0,438	0,486	0,577
8	0,358	0,381	0,411	0,457	0,543
9	0,339	0,36	0,388	0,432	0,514
10	0,322	0,342	0,368	0,41	0,49
11	0,307	0,32	0,352	0,391	0,468
12	0,295	0,313	0,338	0,375	0,45
13	0,284	0,302	0,325	0,361	0,433
14	0,274	0,292	0,314	0,349	0,418
15	0,266	0,283	0,304	0,338	0,404
16	0,258	0,274	0,295	0,328	0,392
17	0,25	0,266	0,286	0,318	0,381
18	0,244	0,259	0,278	0,309	0,371
19	0,237	0,252	0,272	0,301	0,363
20	0,231	0,246	0,264	0,294	0,356
25	0,21	0,22	0,24	0,27	0,32
30	0,19	0,2	0,22	0,24	0,29
35	0,18	0,19	0,21	0,23	0,27
>35	$1{,}07/n^{1/2}$	$1{,}14/n^{1/2}$	$1{,}22/n^{1/2}$	$1{,}36/n^{1/2}$	$1{,}63/n^{1/2}$

Valeurs critiques du Skewness $\gamma_{1,n}$ d'une distribution normale

Valeurs critique γ_α. Pour $\alpha = 0.002$; 0.01 ; 0.02 ; 0.05 ; 0.10 et 0.20 on trouve sur la table la valeur γ_α telle que $P_0(|\gamma_{1,n}| \geqslant \gamma_\alpha) = \alpha$

Exemple. Lire $P_0(|\gamma_{1,10}| \geqslant 1.157) = 0.05$

$n \downarrow \alpha \longrightarrow$	0.20	0.10	0.05	0.02	0.010	0.002
5	0,819	1,058	1,212	1,342	1,396	1,466
6	0,805	1,034	1,238	1,415	1,498	1,642
7	0,787	1,008	1,215	1,432	1,576	1,800
8	0,760	0,991	1,202	1,455	1,601	1,873
9	0,752	0,977	1,189	1,408	1,577	1,866
10	0,722	0,950	1,157	1,397	1,565	1,887
13	0,688	0,902	1,099	1,312	1,441	1,783
15	0,648	0,862	1,048	1,275	1,462	1,778
17	0,629	0,820	1,009	1,188	1,358	1,705
20	0,593	0,777	0,951	1,152	1,303	1,614
23	0,562	0,743	0,900	1,119	1,276	1,555
25	0,543	0,714	0,876	1,073	1,218	1,468
30	0,510	0,664	0,804	0,985	1,114	1,410

Notons $k = 3(n^2 + 27n - 70)(n + 1)(n + 3)/(n - 2)(n + 5)(n + 7)(n + 9)$;
$\nu = (4k - 6)/(k - 3)$

Pour $30 \leqslant n \leqslant 150$ utiliser l'approximation $\gamma_{1,n}\sqrt{\dfrac{(n+1)(n+3)}{6(n-2)}} \times \sqrt{\dfrac{\nu}{\nu-2}} \approx \boldsymbol{t}_\nu$

Lorsque $n > 150$ utiliser l'approximation normale $\gamma_{1,n}\sqrt{\dfrac{(n+1)(n+3)}{6(n-2)}} \approx \mathcal{N}_{0;1}$

© Dunod. La photocopie non autorisée est un délit.

Valeurs critiques du Kurtosis d'une distribution normale

Valeurs de b_α définies par $P_0(B_{2,n} \leqslant b_\alpha) = \alpha$

où $B_{2,n} = \gamma_{2,n} + 3 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4 / \left[\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2$

$n \downarrow \alpha \longrightarrow$	0.01	0.025	0.05	0.1	0.20	0.80	0. 90	0.95	0.975	0.99
7	1.25	1.34	1.41	1.53	1.70	2.78	3.20	3.55	3.85	4.23
8	1.31	1.40	1.46	1.58	1.75	2.84	3.31	3.70	4.09	4.53
9	1.35	1.45	1.53	1.63	1.80	2.98	3.43	3.86	4.28	4.82
10	1.39	1.49	1.56	1.68	1.85	3.01	3.53	3.95	4.40	S.00
12	1.46	1.56	1.64	1.76	1.93	3.06	3.55	4.05	4.56	5.20
15	l.55	1.64	1.72	1.84	2.01	3.13	3.62	4.13	4.66	5.30
20	1.64	1.73	1.83	l.95	2.12	3.20	3.68	4.18	4.68	5.38
25	1.72	1.82	1.92	2.03	2.20	3.24	3.69	4.15	4.63	5.29
30	1.79	1.89	1.98	2.10	2.26	3.26	3.69	4.12	4.57	5.20
40	1.89	1.99	2.07	2.19	2.35	3.29	3.66	4.06	4.46	5.04
45	1.93	2.03	2.11	2.23	2.38	3.29	3.65	4.02	4.41	4.96
50	1.96	2.06	2.15	2.26	2.41	3.29	3.63	4.00	4.36	4.88
60	2.03	2.12	2.21	2.32	2.46	3.29	3.60	3.94	4.28	4.75
70	2.07	2.17	2.25	2.36	2.50	3.28	3.58	3.89	4.20	4.64
100	2.19	2.27	2.35	2.45	2.57	3.26	3.52	3.78	4.03	4.39
200	2.37	2.44	2.51	2.59	2.70	3.22	3.40	3.57	3.75	3.98

Table de la statistique Durbin-Watson pour un seuil de signification 5 %

La table donne les valeurs critiques d_1 et d_2 en fonction de la taille de l'échantillon n et du nombre k de variables explicatives pour un risque d'erreur α de 0,05.
Exemple. Si le nombre de variables explicatives est $k = 1$ et le nombre d'observations est $n = 7$ on déduit que $d_1 = 0{,}70$ et $d_2 = 1{,}36$ et donc $4 - d_1 = 3{,}3$ et $4 - d_2 = 3{,}64$.

	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
n	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
6	0,61	1,40								
7	0,70	1,36	0,47	1,90						
8	0,76	1,33	0,56	1,78	0,37	2,29				
9	0,82	1,32	0,63	1,70	0,46	2,13	0,30	2,59		
10	0,88	1,32	0,70	1,64	0,53	2,02	0,38	2,41	0,24	2,82
11	0,93	1,32	0,66	1,60	0,60	1,93	0,44	2,28	0,32	2,65
12	0,97	1,33	0,81	1,58	0,66	1,86	0,51	2,18	0,38	2,51
13	1,01	1,34	0,86	1,56	0,72	1,82	0,57	2,09	0,45	2,39
14	1,05	1,35	0,91	1,55	0,77	0,78	0,63	2,03	0,51	2,30
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Comparaison de la valeur d^* prise par la statistique DW de Durbin-Watson à ces valeurs critiques.

valeur d^*	$0 < d^* \leqslant d_1$	$d_1 < d^* \leqslant d_2$	$d_2 < d^* \leqslant 4 - d_2$	$4 - d_2 < d^* \leqslant 4 - d_1$	$4 - d_1 < d^* \leqslant 4$
décision	$\mathrm{E}(\varepsilon_i \varepsilon_{i+1}) > 0$	doute	indépendance des ε_i	doute	$\mathrm{E}(\varepsilon_i \varepsilon_{i+1}) < 0$

© Dunod. La photocopie non autorisée est un délit.

BIBLIOGRAPHIE

BARNETT V., LEWIS T., *Outliers in Statiscal data*, JohnWiley § Sons, 1992.

BENZECRI J. P., *La pratique de l'analyse des données* (3 volumes), Dunod, 1981.

BERTHIER P., BOUROCHE J.M., *Analyse des données multidimensionnelles*, Presses Universitaires de France, 1977.

BOUROCHE J.M., SAPORTA G., *L'Analyse des données*, Collection Que sais-je, Presses Universitaires de France, 1980.

CAPERAA Ph., VAN CUTSEN B., *Méthodes et modèles en statistiques non paramétriques,* Presses de l'Université de Laval & Dunod, 1988.

CRAMER H., *Mathematical methods of Statistics*, Princeton University Press, 1974.

GOUTEROUX C., *Théorie des sondages*, Economica, 1981.

KARLIN S., *Initiation aux processus aléatoires*, Dunod, 1968.

KAUFFMANN A., *Cours moderne de calcul des probabilités*, Albin Michel, 1965.

KAUFFMANN P., *Information, estimation, tests*, Dunod, 1994.

KENDALL M., GIBONS J. D., *Rank Correlation methods*, Ed. Arnold, 1985.

LABROUSSE Ch., *Introduction à l'économétrie*, Dunod, 1972.

LECOUTRE J.P., *Statistiques et probabilité*, Dunod, 1998.

LEHMANN E. L., *Nonparametrics,* McGraw-Hill*,* 1977.

LEVY P., *Processus stochastiques et mouvement Brownien*, Gauthier-Villars 1965.

MANOUKIAN Ed. B., *Mathematical non parametric statistics*, Gordon and Breach Science Publishers, 1986.

MANOUKIAN Ed. B., *Modern Concepts and theorems of mathématical statistics*, Springer-Verlay, 1990.

SAPORTA G., *Théories et méthodes statistiques*, Editions Technip, 1982.

TASSI Ph., *Méthodes statistiques*, Economica, 1985.

TENENHAUS M., *Méthodes statistiques en gestion*, Dunod, 1996.

THOMAS A., *Économétrie des variables qualitatives*, Dunod, 2000.

INDEX

ACP, 310
AFC, 310
Alpha de Cronbach, 92
Analyse de variances, 258
Analyse de variance à deux facteurs, 263
Analyse en composantes principales, 310
Analyse factorielle des correspondances, 321
Arrangements avec répétition, 98
Bienaymé-Tchébycheff, 119
Centiles, 11
Coefficient de corrélation linéaire, 46, 229
Combinaison, 99
Concordance de position, 239
Concordance de Kendall, 267
Convergence en loi, 120, 368
Convergence en probabilité, 120, 369
Convergent, 160
Covariance, 45, 229
Décile, 11, 19
Densité de probabilité, 114
Diagramme en bâtons, 5
Distributions liées, marginales, 48
Durbin-Watson, 280
Échantillon, 121, 154
Échelles, 88
Espérance mathématique, 112, 115
Estimateur sans biais, ponctuels, 160
Estimation ponctuelle, 159
Étendue, 26
Fiabilité, 92
Fisher-Snédécor, 142
Fonction de répartition (de X), 6, 7, 111, 116
Formule de Bayes, 107
Friedman, 266, 272
Histogramme, 13
Indice d'Herfindahl, 33
Indice de Gini, 33
Indice des prix, 62
Indices, 62
Indices de concentration, 33
Indices de volume, 62
Intervalle de confiance, 161
Items, 92
Jonckheere-Terpstra, 261
Kolmogorov-Smirnov, 246
Kruskal-Wallis, 262
Kurtosis, 27, 124
Laspeyres, 64, 66
Lissage exponentiel, 77
Log-normale, 146
Loi bêta, 144
Loi de Bernoulli, 127
Loi de Cauchy, 143
Loi de Poisson, 132
Loi de Student-Fisher, 135
Loi exponentielle, 140
Loi gamma, 138
Loi hypergéométrique, 131
Loi logistique, 145
Loi normale, bivariée, 369
Loi uniforme, 141
Lois multinomiales, 148
Lorentz, 31
Markoff, 119
Maximum de vraisemblance (méthode), 160, 368
Médiale, 32
Médiane, 9, 12, 18, 113, 117
Mode, 22
Modèle log-linéaire, 291
Modèle logit, 298, 300
Moment centré d'ordre h, 26, 116
Mouvement saisonnier, 69

© Dunod. La photocopie non autorisée est un délit.

Moyenne arithmétique, 20
Moyenne géométrique, 22
Moyenne harmonique, 21
Moyenne mobile, 72
Moyenne quadratique, 22
Outlier, 253
p-quantiles, 113, 117
Paasche, 65, 66
Permutations, 97
Polygone des fréquences, 5, 15
Quartile, 10, 12, 19
Skewness, 27, 125
Spearman, 234
Stochastiquement, 213
Table de nombres au hasard, 155
Test d'ajustement (du khi-deux), 213, 240, 247
Test d'homogénéité de proportions, 211
Test d'indépendance, 231
Test de Fisher, 255, 266, 270
Test de Kruskal-Wallis, 271
Test Kolmogorov-Smirnov, 246
Tests de comparaison (d'écarts-types), 205
Tests de comparaison de moyennes, 200
Tests de comparaison de proportions, 209
Tests paramétriques de comparaison, 200
Théorème central limite, 126
Trend, 70, 72
Validité, 95
Van Beeck, 126
Variable aléatoire, 110
Variance V(X), 25, 26, 113
Variance expliquée, résiduelle, 54
Volume, 62
Weibull, 147
Wilcoxon, 134
Wilcoxon-Mann-Whitney, 213

GESTION SUP
Finance • Comptabilité

MANAGEMENT RESSOURCES HUMAINES
MARKETING COMMUNICATION
FINANCE COMPTABILITÉ
MANAGEMENT INDUSTRIEL
STRATÉGIE POLITIQUE D'ENTREPRISE
AIDE À LA DÉCISION

Pierre-Charles Pupion

STATISTIQUES POUR LA GESTION
Applications avec Excel et SPSS

2e édition

Enrichie de nouveaux thèmes tels que la réalisation d'un questionnaire, les outliers, les tests de concordance de positionnement, **cette 2e édition couvre la totalité du programme de statistique** – descriptive, inférentielle et exploratoire – dispensée en **formation initiale et approfondie de gestion**. Elle propose :

- **un support théorique** : définitions et propriétés essentielles, accompagnées d'applications pour appréhender et assimiler les notions abordées ;
- **des exemples de traitements par les logiciels** (avec le détail des menus et des options) ;
- **une méthodologie du questionnaire** : élaboration, administration et évaluation ;
- **l'interprétation des résultats** conduisant à l'élaboration des analyses statistiques (analyse financière comparée, études de marché, prévisions de ventes…) ;
- **des exercices et QCM corrigés**.

Cet ouvrage constitue un outil complet de préparation aux examens (test MESSAGE d'entrée en LSG et LCCA, entrée en MAE…).

Public :
- ▶ Étudiants des universités en Licence et Master d'économie et de gestion (AES, IUT, IUP, …)
- ▶ Étudiants des écoles de commerce
- ▶ Professionnels de l'audit et des études de marché

PIERRE-CHARLES PUPION
Ancien élève de l'ENS Cachan, agrégé d'économie et gestion, il est maître de conférences en gestion à l'université de Poitiers.

ISBN 978-2-10-054426-4
www.dunod.com
DUNOD