Theorie des tests

Nicolas Chopin
Emmanuelle Crétois
Jean Diebolt

17 février 2004

Sommaire

Introduction **7**

1 Rappels et notations **13**

2 Tests asymptotiques usuels **19**

Introduction 19

2.1 Application à l'estimation par MV 20

2.2 Test de Wald pour tester $H_0[\theta = \theta_0]$ contre $H_1[\theta \neq \theta_0]$ 20

2.3 Utilisation d'un estimateur de $I(\theta_0)$ 22

2.4 Test des scores 23

2.5 Test du rapport des vraisemblances maximales 24

3 Tests simples : optimalité **27**

3.1 Lemme de Neyman-Pearson 28

3.2 Résultats sur le test de Neyman-Pearson 34

4 Tests uniformément plus puissants : $\theta \in \mathbb{R}$ **37**

4.1 Rapport de vraisemblance monotone 38

4.2 Tests UPP de $[\theta_1 \leq \theta \leq \theta_2]$ 41

5 Tests à structure de Neyman : $\theta \in \mathbb{R}^K$ **47**

5.1 Tests à structure de Neyman 48

5.2 Structure de Neyman et structure exponentielle 50

5.3 Application à des problèmes de comparaison d'échantillons 55

6 Tests et régression **59**

6.1 Régression linéaire multiple : un cas particulier 59

6.2 Estimation par moindres carrés de θ 60

6.3 Estimation sans biais de σ^2 61

6.4 Tester l'absence d'effet d'une variable g_k 63

6.5 Tester l'absence d'effet d'un groupe de variables 64

7 Tests utilisant les rangs et les signes **67**

7.1 Rappels sur les statistiques d'ordre 68

7.2 Runs 69

7.3 Comparaison de variables aléatoires 69

7.4 Tests des longueurs 70

7.5 Test de Wilcoxon dans le cas d'observations non couplées 72

7.6 Test des signes 73

7.7 Test de Wilcoxon dans le cas d'observations couplées 74

7.8 Tests d'indépendance de Spearman et Kendall 76

8 Le test du chi-deux **83**

8.1 Tester $[\mu = \mu_0]$ contre $[\mu \neq \mu_0]$(μ, μ_0 mesures de probabilité) 83

8.2 Le chi-deux comme test d'adéquation 86

8.3 Adéquation, suite 89

8.4 Commentaires pratiques 91

8.5 Test d'indépendance 92

9 Introduction aux tests d'adéquation utilisant le processus empirique **99**

9.1 Le processus empirique uniforme 99

9.2 Tester $H_0\,[F = F_0]$ contre $H_1\,[F \neq F_0]$, F_0 continue donnée 105

9.3 Tester l'appartenance de F à un modèle 106

9.4 Applications pratiques 113

9.5 Goodness-of-fit tests based on the empirical distribution function. Tests of Uniformity 114

9.5.1 The Komolgorov-Smirnov statistics 115

9.5.2 The Cramér-von Mises statistic 115

9.6 Tester l'adéquation d'une loi normale $\mathcal{N}(\mu, \sigma^2)$ 115

9.7 Tester l'adéquation d'une loi exponentielle 117

10 Tests bayésiens **121**

10.1 Fondements de la statistique bayésienne 121

10.1.1 Le paradigme bayésien 121

10.1.2 L'approche décisionnelle : estimateurs de Bayes 124

10.1.3 Modélisation a priori 125

10.2 Tests bayésiens : facteurs de Bayes 128

10.2.1 Test d'une hypothèse ponctuelle 129

10.2.2 Tests et lois a priori impropres 131

10.3 Critiques et extensions 132

10.3.1 Tests fréquentistes et dissymétrie des hypothèses 132

10.3.2 Justification asymptotique contre raisonnement conditionnel 133

10.3.3 Le principe de vraisemblance 133

10.3.4 Conclusion 134

Annexes **137**

A Caractérisation de certaines lois **137**

A.1 Lois discrètes 137

A.2 Lois continues 137

B Tables **139**

B.1 Table de la loi normale $\mathcal{N}(0, 1)$ 139

B.2 Table de quantiles de χ^2_ν $(1 \leq \nu \leq 10)$ 139

B.3 Table de quantiles des lois de Student $t_{\nu,\,p}$ $(1 \leq \nu \leq 30)$ 139

C Décomposition spectrale des vecteurs gaussiens 143

D Convergence en distribution et processus 147

D.1 Convergence en distribution abstraite 147

D.2 Représentation de Skorohod 148

D.3 Convergence en distribution des suites de processus continus[1] 149

E Inverse généralisée d'une fonction de répartition 151

E.1 Résultats sur $F^{\leftarrow}$ 151

E.2 Représentation de Skorohod, suite 152

La macro TEX utilisée pour mettre en page ce document est celle de l'éditeur Springer Verlag.

[1] ou continus par morceaux.

Introduction

Devant la multiplication des modèles de plus en plus complexes, rendue possible par l'accroissement de la puissance et de la vitesse des moyens de calcul, il devient de plus en plus nécessaire de contrôler leur bien-fondé. C'est l'objectif des tests dits d'ADÉQUATION (*goodness-of-fit*) qui permettent de valider ou invalider, au vu des observations, l'usage de tel ou tel modèle, ainsi que de valider ou invalider des modèles simplifiés (dans lesquels on fait des hypothèses simplificatrices, et on élimine des paramètres ou bien des variables explicatives peu significatives).

Nous nous limiterons ici à une présentation très simplifiée des tests d'adéquation, en suivant le fil conducteur que constitue la théorie des processus empiriques et en ne traitant que le cas des modèles de lois de probabilité.

Nous laissons donc de côté l'étude des tests d'adéquation destinés à évaluer des modèles paramétriques de taux de défaillance, de fonctions de régression – classique ou linéaire généralisée (GLM) – de fonctions – linéaires ou non – d'autorégression dans le cadre des processus stochastiques à temps discret, etc.

Le schéma des tests d'adéquation de modèles paramétrique est toujours le même. On cherche à déterminer si tel ou tel écart entre une représentation non paramétrique des observations et la représentation analogue du modèle, dont on a estimé les paramètres à partir des mêmes observations (sous l'hypothèse nulle que les observations sont effectivement issues du modèle en question), est (ou n'est pas) assez petit pour que l'on ne puisse pas (ou que l'on puisse) rejeter cette hypothèse nulle, ceci à un niveau de signification donné.

C'est bien ce qui se passe déjà dans le cas du test du chi-deux (voir chapitre 8) avec paramètres estimés. Dans ce cas, le modèle paramétrique est une famille paramétrée de mesures de probabilité sur un espace mesurable donné. Etant donnée une partition finie fixée en cellules (ou boîtes) disjointes, on compte le nombre d'observations contenues dans chaque cellule : c'est la représentation non paramétrique, dans ce cadre, de la mesure de probabilité d'où sont issues les observations – un simple histogramme ! Ensuite, on estime les paramètres de la mesure de probabilité appartenant au modèle paramétrique à valider, en

supposant que les observations sont issues d'une des mesures de ce modèle. Plusieurs estimateurs sont envisageables. Il est commode, par exemple, d'utiliser l'estimateur du maximum de vraisemblance basé sur les données groupées dans les cellules. Ensuite, on calcule la distance entre ces deux représentations, qui consistent ici chacune en un vecteur de taille finie (le nombre de cellules) : le vecteur des fréquences par cellule (lié aux effectifs observés), et le vecteur des mesures de chaque cellule pour la loi appartenant au modèle paramétrique dont on a estimé les paramètres (c'est l'idée d' effectif théorique par cellule). Pour mesurer cette distance (ou écart) entre ces deux vecteurs, on utilise une norme euclidienne choisie de manière à ce que la loi limite, sous l'hypothèse nulle, de cette distance au carré lorsque le nombre d'observations tend vers l'infini soit indépendante de la vraie valeur (inconnue) du ou des paramètres. Cette loi est directement reliée à une loi du chi-deux (tout dépend du choix de la procédure d'estimation).

Nous avons pris le parti de présenter ce test fondamental en termes de mesure et de processus empiriques, afin d'unifier les chapitres 8 et 9.

Dans le chapitre 9, on cherche à tester essentiellement des familles paramétrées de lois sur $\mathbb{R}$, ou sur un intervalle de $\mathbb{R}$, dont les fonctions de répartition sont continues. On montre que tout revient alors à tester l'hypothèse nulle qu'une suite d'observations sont uniformes sur $[0, 1]$. On choisit comme représentation non paramétrique la fonction de répartition empirique. Cela nous conduit au processus empirique uniforme, défini sur $[0, 1]$, puis à un processus gaussien sur $[0, 1]$ étroitement relié au mouvement brownien, et à sa structure dite décomposition de Karhunen-Loève, qui est une extension de l'ACP.

Bien entendu, il existe de nombreuses autres procédures pour tester l'adéquation de tel ou tel modèle particulier (en particulier, la normalité et l'exponentialité). Mais il n'était pas possible de les mentionner ici.

Nous avons introduit les tests appelés non paramétriques au chapitre **??**. Ils s'appuient beaucoup sur les statistiques ordonnées, et sont essentiellement destinés à éprouver, sans hypothèse de modèle, que deux échantillons distincts sont issus d'une même loi (ou non) : tests de permutation, tests de rangs, tests de signe. Dans la mesure du possible, le principe mathématique de ces tests sera étudié en séance d'exercices.

Les premiers chapitres sont classiques. Ils sont consacrés à l'étude des tests paramétriques usuels, selon le schéma de Neyman et Pearson, après un premier chapitre consacré principalement à des rappels succincts sur l'estimation par maximum de vraisemblance et à l'information de Fisher, et un chapitre 2 consacré aux tests asymptotiques usuels (Wald, scores, rapport des vraisemblances maximales) et à leurs liens avec la détermination de régions de confiance, le tout sous forme de rappels accompagnés d'exemples et d'exercices simples.

Les notions fondamentales apparaissent au chapitre 3, consacré aux tests simples : niveau de signification, erreurs de type I et II, puissance, optimalité selon le schéma de Neyman et Pearson. Il est remarquable que le résultat fondamental, qui sera constamment exploité dans les chapitres 4 et 5, le lemme de Neyman-Pearson, se déduise d'une propriété élémentaire du cours d'Intégration ! C'est grâce à ce résultat, étudié ici de façon détaillée, que l'on peut construire à partir des fonctions de vraisemblance, les régions d'acceptation (et donc aussi de rejet) correspondant aux tests de niveau donné les plus puissants contre une alternative simple.

Le chapitre 10 est consacré aux tests bayésiens. Ce chapitre constitue aussi une brève introduction à la démarche bayésienne en Statistique, qui prend chaque jour davantage d'importance. Il apporte un point de vue critique sur les principes développés dans les chapitres précédents.

Je remercie Philippe Choné (Insee) et Nicolas Chopin (Crest-Ensae) pour leur aide et leurs conseils. Nicolas Chopin a rédigé le chapitre consacré aux tests bayésiens. Je remercie Emmanuelle Crétois (Université de Grenoble) pour le chapitre consacré aux tests non paramétriques utilisant les statistiques ordonnées. Je remercie enfin Julien-Elie Taieb (Ensae 00) pour le beau travail de frappe et de mise en page qu'il a effectué.

Premiers éléments bibliographiques

Alcalá, J. T., Cristóbal, J. A. & González Manteiga, W. (1999) Goodness-of-fit test for linear models based on local polynomials. *Statist. Probab. Lett.* **42**, 39–46.

An, H. Z. & Cheng, B. (1991) A Kolmogorov-Smirnov type statistic with application to test for nonlinearity in time series. *Internat. Statist. Rev.* **56**, 287–307.

Antoniadis, A., Gijbels, I. & Grégoire, G. (2000) Model selection using wavelet decomposition and applications. *Biometrika*.

Association pour la Statistique et ses Utilisations (1996) *Inférence non paramétrique. Les statistiques de rangs*, Jean-Jacques Droesbeke et Jeanne Fine éditeurs, Collection "Ellipses", Editions de l'Université Libre de Bruxelles.

Azzalini, A., Bowman, A. W. & Härdle, W. (1989) On the use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

Berger, J. O. (1980) *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag : New York, Berlin, Heidelberg.

Billingsley, P. (1968) *Convergence of Probability Measures*. Wiley : New York.

Bowman, A. W. & Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis : The Kernel Approach with* `S-PLUS` *Illustrations*. Oxford Statistical Science Series : Oxford.

Capéraà, P. & Van Cutsem, B. (1988) *Méthodes et modèles en Statistique non paramétrique*. Dunod : Paris.

Conover, W. J. (1971) *Practical Nonparametric Statistics*. Wiley : New York.

Cox, D. & Koh, E. (1989) A smoothing spline based test of model adequacy in polynomial regression. *Ann. Inst. Statist. Math.* **41**, 383–400.

Cox, D., Koh, E., Wahba, G. & Yandell, B.S. (1988) Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.* **16**, 113–119.

Dacunha-Castelle, D. & Duflo, M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Dacunha-Castelle, D. & Duflo, M. (1994) *Probabilités et Statistiques*, Tome 2. Masson : Paris, 2e édition.

Dagnelie P. (1975) *Théorie et méthodes statistiques*, Tome 2. Vander-Oyez : Bruxelles, 2e édition.

D'Agostino, R. B. & Stephens, M. A. (1986) *Goodness-of-Fit Techniques*, Statistics, textbooks and monographs **68**. Marcel Dekker : New York and Basel.

Davison, A. C. & Tsai, C. L. (1992) Regression model diagnostics. *Internat. Statist. Rev.* **60**, 337–353.

Dette, H. (2000) On a nonparametric test for linear relationships. *Statist. Probab. Lett.* **46**, 307–316.

Dette, H. & Munk, A. (1998) Validation of linear regression models. *Ann. Statist.* **26**, 778–800.

Diebolt, J. & Zuber, J. (1999) Goodness-of-fit tests for nonlinear heteroscedastic regression models. *Statist. Probab. Lett.* **42**, 53–60.

Diebolt, J. & Zuber, J. (2001) On testing the goodness-of-fit of a nonlinear heteroscedastic regression model. *Commun. Statist.- Simulation and Computation* **30**, 195–216.

Dudley, R. M. (1989) *Real Analysis and Probability*, Mathematics Series. Chapman and Hall : New York and London.

Erkel-Rousse, H. (2001) *Introduction à l'économétrie du modèle linéaire*. Polycopié de l'Ensae. Insee–Ensae : Montrouge[2].

Eubank, R. L. & Hart, J. D. Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.* **20**, 1412–1425.

Eubank, R.L. ; Hart, J.D. (1993) Commonality of cusum, von Neumann and smoothing-based goodness-of-fit tests. *Biometrika* **80**, 89–98.

Eubank, R. L., Hart, J. D. & LaRiccia, V. N. (1993) Testing goodness-of-fit via nonparametric function estimation techniques. *Commun. Statist.- Theory Meth.* **22**, 3327–3354.

Eubank, R.L. & Spiegelman, C.H. (1990) Testing the goodness-of-fit of a linear model via nonparametric regression techniques. *J. Am. Statist. Assoc.* **85**, 387–392.

Ferguson, T. S. (1967) *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press : New York and London.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Härdle, W. & Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926–1947.

Hart, J. D. (1997) *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag : New York, Berlin, Heidelberg.

Koul, H. L. & Stute, W. (1999) Nonparametric model checks for times series *Ann. Statist.* **27**, 204–236.

Lecoutre, J.-P. & Tassi, P. (1987) *Statistique non paramétrique et robustesse*. Collection "Economie et statistiques avancées", Economica : Paris.

Malinvaud, E. (1978) *Méthodes statistiques de l'économétrie*. Dunod : Paris, 3e édition.

[2]Cet ouvrage très riche contient une bibliographie très complète sur l'histoire de la Statistique et de l'Econométrie (pages 53–54 et 56–66), et bien sûr sur le modèle linéaire (pages 54–56).

Monfort, A. (1982) *Cours de statistique mathématique*. Collection "Economie et statistiques avancées", Economica : Paris.

Müller, H. G. (1992) Goodness-of-fit diagnostics for regression models. *Scand. J. Statist.* **19**, 157–172.

Rayner, J. C. W. & Best, D. J. (1989) *Smooth Tests of Goodness-of-Fit*. Oxford University Press : Oxford.

Robert, C. P. (1992) *L'analyse statistique bayésienne*. Economica : Paris.

Robert, C. P. (2000) *The Bayesian Choice*. Springer-Verlag : New York, Berlin, Heidelberg.

Schervisch, M. J. (1995) *Theory of Statistics*. Springer-Verlag : New York, Berlin, Heidelberg.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley : New York.

Stephens, M.A. (1978) On the half-sample method for goodness-of-fit. *J. R. Statist. Soc.* B **40**, 64–70.

Stute, W. (1997) Nonparametric model checks for regression. *Ann. Statist.* **25**, 613–641.

Stute, W. & González Manteiga, W. (1996) NN goodness-of-fit tests for linear models. *J. Statist. Plann. Inference* **53**, 75–92.

Stute, W., González Manteiga, W. & Presedo Quindimil, M. (1998) Checks for regression. *J. Am. Statist. Assoc.* **93**, 141–149.

Stute, W., Thies, S. & Zhu, L.X. (1998) Model checks for regression : an innovation process approach. *Ann. Statist.* **26**, 1916–1934.

Tassi, P. (1985) *Méthodes statistiques*. Collection "Economie et statistiques avancées", Economica : Paris.

Ulmo, J. & Bernier, J. (1973) *Eléments de décision statistique*. Presses Universitaires de France : Paris.

Zuber, J. (1997) *Un test chi-carré d'adéquation de modèles paramétriques en régression*. Thèse, Ecole Polytechnique Fédérale de Lausanne, Suisse.

1

Rappels et notations

1. Nous considérerons des variables aléatoires (v.a.) indépendantes et identiquement distribuées (i.i.d.) à valeurs dans un espace mesurable séparable $(E, \mathcal{E})$: E peut être discret, par exemple être un ensemble fini, l'ensemble $\mathbb{N}$, $\mathbb{N}^*$ ou $\mathbb{Z}$, ou un produit cartésien formé à partir de ces divers ensembles ; E peut être continu, par exemple être un intervalle de $\mathbb{R}$ (y compris le cas $\mathbb{R}$), ou une partie de $\mathbb{R}^K$ ($K \geq 1$), etc.

   On supposera que les lois de probabilité μ_θ, $\theta \in \Theta$, dont la famille constitue tel ou tel modèle paramétré, sont dominées par une mesure $\sigma-$finie ν sur $(E, \mathcal{E})$, ou *mesure de référence* : ν peut être la mesure de comptage sur un espace discret, ou la mesure de Lebesgue (convenablement restreinte si nécessaire) sur $\mathbb{R}^K$, ou encore un produit direct des deux.

   Les lois μ_θ admettent donc des densités $f_\theta : E \to \mathbb{R}^+$ relativement à ν, définies pour chaque $\theta \in \Theta$, à un ensemble de ν-mesure nulle près. Cependant, il n'y aura en général pas d'ambiguïté :
   - Si E est discret et ν la mesure de comptage ;
   - Si E est continu (avec pour ν, par exemple, la mesure de Lebesgue), et que l'on exige que chaque f_θ soit continue (ou sinon qu'on adopte, si nécessaire, une convention appropriée, par exemple la continuité à gauche ou à droite, en chaque point de discontinuité de première espèce dans le cas où E est un intervalle de $\mathbb{R}$).

   Lorsque Θ est un intervalle de $\mathbb{R}$ ou un ouvert (en général connexe) de $\mathbb{R}^K$, $K \geq 1$, nous supposerons que pour tout $x \in E$ tel que $f_\theta(x) > 0$ (ensemble en général indépendant de θ), la fonction $\theta \in \Theta \mapsto \ln f_\theta(x)$ admet des dérivées partielles continues d'ordre 2 au moins, et on notera :

$$s_\theta(x) = \begin{cases} \frac{\partial}{\partial\theta} \ln f_\theta(x) & \text{si } \Theta \text{ intervalle de } \mathbb{R} \\ \nabla_\theta \ln f_\theta(x) = \left(\frac{\partial}{\partial\theta_1} \ln f_\theta(x), \ldots, \frac{\partial}{\partial\theta_K} \ln f_\theta(x) \right)^T & \text{si } \Theta \in \mathbb{R}^K \end{cases} \tag{1.1}$$

   la fonction de score du modèle.

   La fonction de score du modèle, $s_\theta(x)$, est donc un vecteur colonne de $\mathbb{R}^K$ pour tout couple (θ, x). On supposera en outre que pour chaque θ, les fonctions composantes de

cette fonction de score sont de carré intégrable par rapport à la mesure $d\mu_\theta = f_\theta\, d\nu$, et aussi que la matrice d'information de Fisher (matrice symétrique non négative $K \times K$)

$$\begin{aligned} I(\theta) &= \left[-\int_E \frac{\partial^2}{\partial\theta_j \partial\theta_{j'}} (\ln f_\theta(x)) f_\theta(x)\ \nu(dx) \right]_{1\leq j,\, j'\leq K} \\ &= \left[\int_E s_\theta(x) s_\theta^T(x) f_\theta(x)\, \nu(dx) \right]_{1\leq j,\, j'\leq K} \end{aligned} \tag{1.2}$$

est bien définie, et définie positive (donc inversible). L'égalité ci-dessus est vérifiée dès que, par le théorème de convergence dominée de Lebesgue, on peut assurer que pour tous $1 \leq j,\ j' \leq K$:

$$\begin{aligned} \frac{\partial}{\partial\theta_j} \int_E \frac{\partial}{\partial\theta_{j'}} (\ln f_\theta(x))\, f_\theta(x)\, \nu(dx) &= \int_E \frac{\partial^2}{\partial\theta_j\ \partial\theta_{j'}} (\ln f_\theta(x))\, f_\theta(x)\, \nu(dx) \\ &\quad + \int_E \frac{\partial}{\partial\theta_j} (\ln f_\theta(x)) \frac{\partial f_\theta(x)}{\partial\theta_{j'}}\, \nu(dx) \end{aligned}$$

Enfin, la régularité de la fonction $(\theta,\, x) \mapsto f_\theta(x)$ doit elle aussi être précisée. L'existence de dérivées partielles mixtes du type

$$\frac{\partial^2}{\partial\theta_j \partial x} f_\theta(x)$$

et leur intégrabilité par rapport à la mesure $d\mu_\theta = f_\theta\, d\nu$ peuvent s'avérer nécessaires. L'ensemble des conditions de ce type (avec un support de μ_θ indépendant de $\theta \in \Theta$) seront appelées les conditions usuelles de régularité.

2. Soit $(X_1, \ldots, X_n)$ un n-échantillon de v.a. i.i.d. à valeurs dans E et suivant chacune la loi $d\mu_{\theta_0} = f_{\theta_0}\, d\nu$ (on notera $f_0 = f_{\theta_0}$, $s_0 = s_{\theta_0,}$ etc., pour simplifier). La fonction de vraisemblance associée est

$$\theta \mapsto \mathcal{L}_n(\theta) = \prod_{i=1}^n f_\theta\,(X_i) = \prod_{i=1}^n f_\theta\,(x_i) \tag{1.3}$$

On passera librement de la notation X_i (qui désigne les v.a. dont les observations x_i sont des réalisations) à la notation x_i (qui désigne les observations). La fonction de log-vraisemblance associée est

$$\theta \mapsto L_n(\theta) = \ln \mathcal{L}_n(\theta) = \sum_{i=1}^n \ln f_\theta\,(X_i) = \sum_{i=1}^n \ln f_\theta\,(x_i) \tag{1.4}$$

On notera

$$\ell_n(\theta) = \frac{1}{n} L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ln f_\theta\,(X_i) = \frac{1}{n} \sum_{i=1}^n \ln f_\theta\,(x_i) \tag{1.5}$$

Si $\ln f_\theta\,(\cdot)$ est μ_{θ_0}-intégrable, alors quand $n \to \infty$:

$$\ell_n(\theta) \xrightarrow{\text{p.s.}} \mathbb{E}_\theta\left[\ln f_\theta\,(X)\right] = \int_E (\ln f_\theta\,(x))\, f_0\,(x)\ \nu(dx)\,, \tag{1.6}$$

où $\mathbb{E}_\theta[\phi(X)]$ désigne (un peu abusivement) l'espérance mathématique de $\phi(X)$ lorsque X est une v.a. de loi μ_θ. On sait que l'entropie

$$\int_E (\ln f_\theta)\ f_0\, d\nu$$

est maximale pour $\theta = \theta_0$, et que son maximum (global) est atteint en ce seul point si la famille des densités f_θ est identifiable.
Puisque, pour n assez grand, $\ell_n(\theta) \approx \int_E (\ln f_\theta)\, f_0 \; d\nu$, [1] il est naturel de chercher à estimer θ_0 sur la base de $(x_1, \ldots, x_n)$, en calculant la valeur $\widehat{\theta}_n$ du paramètre θ où la fonction $\theta \mapsto \ell_n(\theta)$ atteint son maximum global. Comme cette fonction est maximale si et seulement si la vraisemblance $\mathcal{L}_n(\theta)$ est maximale, on appelle cet estimateur l'estimateur du maximum de vraisemblance (MV).

Théorème 1.1 *Sous les conditions usuelles de régularité (éventuellement complétées)*

1. $\widehat{\theta}_n \xrightarrow{\text{p.s.}} \theta_0$ *quand* $n \to +\infty$ *;*
2. *On a :*
$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_0^{-1} s_0(X_i) \; + \; o_P(1),$$
où $I_0 = I(\theta_0)$, $s_0 = s_{\theta_0}$, *et* $o_P(1)$ *désigne un terme qui tend vers* 0 *en probabilité quand* $n \to +\infty$.
3. *On a :*
$$\mathbb{E}_{\theta_0}\left[s_0(X)\right] = 0$$
4. *Enfin,*
$$\sqrt{n}\left[\widehat{\theta}_n - \theta_0\right] \xrightarrow{d} \mathcal{N}(0,\, I^{-1}(\theta_0)) \quad \text{quand} \;\; n \to +\infty\,,$$

Les fonctions $I(\theta)^{-1}\, s_\theta(\cdot)$ s'appellent des *fonctions d'influence.*

Idée de la démonstration. Admettons (1). Pour n assez grand, $\widehat{\theta}_n$ est p.s. proche de θ_0. Les *équations de la vraisemblance* s'écrivent :
$$\dot{\ell}_n\left(\widehat{\theta}_n\right) = 0,$$
la notation $\dot{\ell}_n$ désignant ici la dérivée (si θ réel) ou la différentielle de la fonction ℓ_n (et convention analogue pour $\ddot{\ell}_n$). D'autre part,
$$\ddot{\ell}_n(\widehat{\theta}_n) \xrightarrow{\text{p.s.}} -I_0 \text{ quand } n \to +\infty\,,$$
donc $\ddot{\ell}_n(\widehat{\theta}_n)$ est p.s. définie négative pour n assez grand.

Formons le développement de Taylor à l'ordre 1 de $\dot{\ell}_n(\theta)$ au voisinage de $\theta = \widehat{\theta}_n$, puis choisissons $\theta = \theta_0$:
$$\dot{\ell}_n(\theta_0) = \dot{\ell}_n(\widehat{\theta}_n) + \ddot{\ell}_n(\widehat{\theta}_n)\left[\theta_0 - \widehat{\theta}_n\right] + o_P\left(\left\|\widehat{\theta}_n - \theta_0\right\|\right) \approx I_0\left(\widehat{\theta}_n - \theta_0\right)$$
$$\begin{cases} I_0 \;\text{ matrice carrée} \\ \widehat{\theta}_n \;-\; \theta_0 \;\text{ matrice colonne} \end{cases}$$
avec
$$\dot{\ell}_n(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} s_0(X_i) \tag{1.7}$$

[1] Nous utiliserons tantôt la notation $\int f(x)\, \nu(dx)$, tantôt $\int f\, d\nu$.

Le fait que $\mathbb{E}_{\theta_0}(s_0(X)) = 0$ découle de la définition de la fonction score, de l'identité $\forall\theta$, $\int f_\theta \, d\nu = 1$ et de la possibilité (conséquence des conditions usuelles de régularité) de dériver ou différentier en $\theta \mapsto \int f_\theta \, d\nu$ sous le signe intégral.

On déduit directement le dernier point des deux précédents, à l'aide du théorème central de la limite normale. ■

Remarque 1 (*distance* de Küllback-Leibler, ou *entropique*) *Quand elle est bien définie, il s'agit de la "distance"*[2]

$$d_{KL}(f,\, g) = \int_E f/g \, \ln f/g \; g \, d\nu = \int_E f \, \ln f/g \, d\nu \,,$$

où f et $g \in L^1(E, \nu)$ vérifient $f : E \to \mathbb{R}^+$, $g : E \to \mathbb{R}^+$ et $\int_E f \, d\nu = \int_E g \, d\nu = 1$. L'approximation empirique de $d_{KL}(f_0,\, f_\theta)$ est

$$d_n(f_0,\, f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln \left(\frac{f_0(X_i)}{f_\theta(X_i)} \right)$$

Supposons θ réel, pour simplifier, et considérons $\theta = \theta_0 + h$, h petit : sous les conditions usuelles de régularité, on trouve que

$$d_{KL}(f_{\theta_0},\, f_{\theta_0+h}) = \frac{h^2}{2} \, I(\theta_0) \; + \; o(h^2), \quad \textit{quand } \; h \to 0$$

→ Voir en Annexe B une table de la loi $\mathcal{N}(0,\, 1)$.

Exercices du chapitre 1

Calculer la log-vraisemblance, la fonction score, l'information de Fisher, l'estimateur du maximum de vraisemblance, puis vérifier que les résultats du théorème 1.1 s'appliquent dans les cas suivants (voir Annexe A pour la définition des familles de lois) :

1. *Bernoulli de paramètre* $\theta = p \in [0,\, 1]$;
2. *Poisson* de paramètre $\theta = \lambda > 0$;
3. *Géométrique de paramètre* $\theta = p \in [0, 1[$;
4. *Normale de paramètre*
$$\theta = \begin{bmatrix} \mu \\ \sigma \end{bmatrix} \in \mathbb{R}^2, \mu \in \mathbb{R}, \sigma > 0 \,,$$
passage de σ à σ^2 ;
5. *Exponentielle de paramètre* $\theta = \lambda > 0$;
6. *Gamma de paramètre* $\theta = \lambda > 0$ *inconnu, avec* $r > 0$ *supposé connu et donné.*

[2] Elle n'est pas symétrique !

---

7. *Soit $X_1, \ldots, X_n$ un échantillon i.i.d. issu de la loi uniforme sur $[0, \theta]$, le paramètre $\theta > 0$ étant inconnu. Peut-on définir dans ce cas l'EMV $\widehat{\theta}_n$ de θ ? Peut-on calculer le score, l'information de Fisher ? Le théorème 1.1 s'applique-t-il ?*

2

Tests asymptotiques usuels

Introduction

Ces tests reposent sur les propriétés asymptotiques de suites d'estimateurs lorsque n, la taille de l'échantillon, tend vers $+\infty$. Le plus souvent, il s'agit d'estimateurs du maximum de vraisemblance (MV). Tout repose sur le lemme suivant :

Lemme 2.1 *1. Soit $(Y_n)_{n\geq 1}$ une suite de v.a. à valeurs dans $\mathbb{R}^K$ ($K \geq 1$), telle que*

$$Y_n \xrightarrow{d} Y \text{ quand } n \to +\infty$$
$$Y \sim \mathcal{N}(0, \Sigma)\,,$$

Σ étant définie positive. Alors

$$Y_n^T \, \Sigma^{-1} \, Y_n \xrightarrow{d} \chi^2_K \quad \text{quand } \; n \to +\infty \tag{2.1}$$

2. Ceci reste vrai si on remplace Σ par $\Sigma_n \xrightarrow{\text{p.s.}} \Sigma$ ($n \to +\infty$), avec Σ_n définie positive $\forall n \geq 1$.

Démonstration. Nous ne prouvons que (1), car (2) s'en déduit simplement. Tout repose sur la décomposition spectrale suivante de Y_n (voir Annexe C). Soient $\lambda_1 \geq \ldots \geq \lambda_K$ les valeurs propres (qui sont réelles positives) de Σ, et soit $u_1, \ldots, u_K$ un système orthonormé (base de $\mathbb{R}^K$) de vecteurs propres de Σ, identifiés à des matrices colonnes. On a :

$$Y_n = \sum_{j=1}^{K} \sqrt{\lambda_j}\, \xi_j^{(n)}\, u_j, \quad n \geq 1\,, \tag{2.2}$$

avec $\xi_j^{(n)} = \lambda_j^{-1/2} \langle Y_n, u_j \rangle$ v.a. centrées réduites, telles que

$$\begin{aligned} \xi_j^{(n)} &\xrightarrow{d} \xi_j \\ \xi_j \quad &\sim \quad \mathcal{N}(0,1), \quad \xi_j \text{ i.i.d.} \end{aligned}$$

Alors

$$Y_n^T \Sigma^{-1} Y_n = \sum_{j,j'=1}^{K} \sqrt{\lambda_j \lambda_{j'}}\, \xi_j^{(n)} \xi_{j'}^{(n)} \left(u_j^T \Sigma^{-1} u_{j'}\right)$$

Comme $\Sigma uj = \lambda_j u_j$ ($\lambda_j > 0$), il en résulte que $\Sigma^{-1} u_j = \lambda_j^{-1} u_j$, donc

$$u_j^T \Sigma^{-1} u_{j'} = \lambda_{j'}^{-1} \left(u_j^T u_{j'}\right) = \lambda_{j'}^{-1} \left\langle u_j, u_{j'} \right\rangle = \lambda_j^{-1} \delta_{jj'} \quad \text{(Kronecker)}$$

Finalement,

$$Y_n^T \Sigma^{-1} Y_n = \sum_{j=1}^{K} \lambda_j \left(\xi_j^{(n)}\right)^2 \lambda_j^{-1} = \sum_{j=1}^{K} \left(\xi_j^{(n)}\right)^2 \xrightarrow{d} \sum_{j=1}^{K} \xi_j^2 \sim \chi_K^2 \tag{2.3}$$

■

2.1 Application à l'estimation par MV

Dans ce cas,

$$Y_n = \sqrt{n} \left[\widehat{\theta}_n - \theta_0\right] \xrightarrow{d} \mathcal{N}(0,\, I^{-1}(\theta_0)) \quad \text{quand } n \to +\infty\,,$$

où θ_0 est la vraie valeur du paramètre (à valeurs dans $\mathbb{R}^K$), et $I(\theta_0)$ est la matrice d'information de Fisher du modèle en θ_0, supposée définie positive, tout cela sous les conditions usuelles de régularité.

On prend $\Sigma = I^{-1}(\theta_0)$ dans le lemme ci-dessus. On obtient :

$$n \left[\widehat{\theta}_n - \theta_0\right]^T I(\theta_0) \left[\widehat{\theta}_n - \theta_0\right] \xrightarrow{d} \chi_K^2 \quad \text{quand } n \to +\infty \tag{2.4}$$

Remarquons que la loi limite (une loi du chi-deux à K degrés de liberté) *ne dépend pas* de la vraie valeur, en principe inconnue, θ_0. On parle de *fonction* (ou statistique) *pivotale asymptotique*.

Ce résultat reste vrai si on remplace θ_0 par $\widehat{\theta}_n$, car alors

$$I(\widehat{\theta}_n) \xrightarrow{\text{p.s.}} I(\theta_0)$$

2.2 Test de Wald pour tester $H_0[\theta = \theta_0]$ contre $H_1[\theta \neq \theta_0]$

H_0 est appelée l'*hypothèse nulle*. Elle est *simple* dans ce cas. Soit $0 < \alpha < 1$ le niveau de signification asymptotique du test que l'on veut construire (par exemple, $\alpha = 0,05$).

Cela signifie que l'on veut que sous H_0, la probabilité de rejeter (à tort, puisque sous H_0) l'hypothèse nulle H_0 tende, lorsque $n \to +\infty$, vers α. Dans le cas du test de Wald, la *statistique de test* est

$$T_n^{(1)} = n \left[\widehat{\theta}_n - \theta_0\right]^T I(\theta_0) \left[\widehat{\theta}_n - \theta_0\right] \tag{2.5}$$

Sous H_0, $\mathbb{P}[T_n^{(1)} \leq q] \to \mathbb{P}[\chi^2_K \leq q]$. La notation abusive $\mathbb{P}[\chi^2_K \leq q]$ désigne la probabilité qu'une v.a. de loi chi-deux à K degrés de liberté soit inférieure ou égale à la valeur q. C'est donc la valeur de la fonction de répartition d'une telle v.a. au point q. On choisit $q = z_{1-\alpha}$ tel que $\mathbb{P}[\chi^2_K > z_{1-\alpha}] = \alpha$. Ainsi, $z_{1-\alpha}$ est le $(1-\alpha)$-quantile de χ^2_K.

On ne rejette pas H_0 si $T_n^{(1)} < z_{1-\alpha}$ (on dit alors qu'on accepte H_0).

On rejette H_0 si $T_n^{(1)} > z_{1-\alpha}$.

L'ensemble des réalisations $(x_1, \ldots, x_n)$ de v.a. i.i.d. $(X_1, \ldots, X_n)$ issues du modèle considéré avec $\theta = \theta_0$ telles que

$$T_n^{(1)}(x_1, \ldots, x_n) < z_{1-\alpha} \tag{2.6}$$

s'appelle la *région d'acceptation (asymptotique)* du test. La *région de rejet (asymptotique)* est définie par

$$T_n^{(1)}(x_1, \ldots, x_n) > z_{1-\alpha} \tag{2.7}$$

C'est, à un ensemble de mesure tendant vers 0 quand $n \to +\infty$ près, le complémentaire de la région d'acceptation.

→ Voir en Annexe B une table des lois χ^2_K $(1 \leq K \leq 10)$.

Exemple 1 (Modèle gaussien $\mathcal{N}(\mu, \Sigma)$) *Test de $[\mu = \mu_0]$ contre $[\mu \neq \mu_0]$ avec Σ (matrice variance) connue ou non. Si Σ est inconnue, on la remplace par un estimateur $\widehat{\Sigma}_n$ tel que $\widehat{\Sigma}_n \xrightarrow{\text{p.s.}} \Sigma$ $(n \to +\infty)$ sous H_0. Dans ce cas,*

$$\widehat{\theta}_n = \widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$
$$I(\theta_0) = \Sigma^{-1}$$

En particulier, pour $K = 1$, si on note $\Sigma = \sigma^2$, on accepte H_0 $[\mu = \mu_0]$ si et seulement si

$$(\overline{X} - \mu_0)^2 \leq \frac{\sigma^2 z_{1-\alpha}}{n}, \qquad \mathbb{P}\left[\chi^2_1 \leq z_{1-\alpha}\right] = 1 - \alpha\,,$$

soit encore

$$\overline{X} \in \left[\mu_0 - \sigma\sqrt{\frac{z_{1-\alpha}}{n}},\ \mu_0 + \sigma\sqrt{\frac{z_{1-\alpha}}{n}}\right]$$

Cela revient à écrire que la région d'acceptation est l'ensemble des $(x_1, \ldots, x_n) \in \mathbb{R}^n$ compris entre les deux hyperplans

$$\frac{x_1 + \ldots + x_n}{n} = \mu_0 \pm \sigma\sqrt{\frac{z_{1-\alpha}}{n}}$$

On remarquera que cette région s'exprime à l'aide de la statistique $\overline{X}$. Lorsque toute l'information relative à la famille de lois sous-jacente peut ainsi être résumée, on parle de statistique exhaustive.

Exercice 2.1 (Modèle gaussien unidimensionnel $\mathcal{N}(\mu, \sigma^2)$) *Test de* $\left[\sigma^2 = \sigma_0^2\right]$ *contre* $\left[\sigma^2 \neq \sigma_0^2\right]$ $(\sigma_0^2 > 0)$*, avec* μ *connue ou inconnue.*

1. *Considérer* $\widehat{\theta}_n = \widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \mu)^2$ *: calculer* $I(\sigma_0^2)$.
2. *Construire le test de Wald correspondant si* μ *est connue.*
3. *Comment faire si* μ *est inconnue ?*

Exercice 2.2 (Modèle binomial $\mathrm{Bin}(n, p)$**)** *Test de* $[p = p_0]$ *contre* $[p \neq p_0]$.

Exercice 2.3 (Modèle $\mathrm{Exp}(\theta)$**)** *Ici,* $f_\theta(x) = (1/\theta)\exp(-x/\theta)$ $(\theta > 0\ ;\ x \geq 0)$. *Test de* $[\theta = \theta_0]$ *contre* $[\theta \neq \theta_0]$.

Exercice 2.4 (Modèle de $\mathrm{Poisson}(\lambda)$**)** $\mathbb{P}[X = m] = \frac{\lambda^m}{m!}\,\mathrm{e}^{-\lambda}$ *pour* $\lambda > 0$; $m \in \mathbb{N}$. *Test de* $[\lambda = \lambda_0]$ *contre* $[\lambda \neq \lambda_0]$.

Dans chaque cas, on mettra en évidence la statistique exhaustive.

2.3 Utilisation d'un estimateur de $I(\theta_0)$

Dans les cas un peu plus compliqués, on ne sait plus calculer $I(\theta_0)$ analytiquement. Une possibilité est alors de remplacer $I(\theta_0)$ par son estimateur empirique

$$\widehat{I}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^n s_{\theta_0}(X_i) s_{\theta_0}^T(X_i)\,, \tag{2.8}$$

qui converge p.s. lorsque $n \to +\infty$ vers $I(\theta_0)$, puisque les X_i sont supposées i.i.d. de loi associée au paramètre θ_0.

Exemple 1 (suite) *Supposons, pour simplifier, que* $K = 1$ *et que la variance* $\Sigma = \sigma^2$ *est connue. On a alors, pour* $\theta = \mu$:

$$\ln f_\theta(x) = \mathrm{Cte} - \frac{(x-\mu)^2}{2\sigma^2}\,,$$

d'où

$$s_\theta(x) = \frac{\partial}{\partial \mu}\ \ln f_\mu(x) = s_\mu(x) = \frac{x-\mu}{\sigma^2}$$

et

$$\widehat{I}_n(\mu_0) = \frac{1}{n\sigma^4}\sum_{i=1}^n (X_i - \mu_0)^2$$

Exercices 2.1 à 2.4 (suite) *Déterminer* $\widehat{I}_n(\theta_0)$ *dans chaque cas.*

2.4 Test des scores

On a vu que sous les conditions usuelles de régularité, l'estimateur du MV $\widehat{\theta}_n$ admettait le développement sous forme linéaire suivant

$$\sqrt{n}\left[\widehat{\theta}_n - \theta_0\right] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_0^{-1} s_0(X_i) \, + \, o_P(1) \qquad (n \to +\infty)\,,$$

avec $I_0 = I(\theta_0)$, $s_0 = s_{\theta_0}$ et $o_P(1)$ tendant vers 0 en probabilité quand $n \to +\infty$. Par conséquent, la statistique de test du test de Wald,

$$T_n^{(1)} = n\left[\widehat{\theta}_n - \theta_0\right]^T I_0 \left[\widehat{\theta}_n - \theta_0\right]\,,$$

a même loi limite que

$$T_n^{(2)} = \frac{1}{n}\left[\sum_{i=1}^{n} s_0^T(X_i)\right] I_0^{-1} \left[\sum_{i=1}^{n} s_0(X_i)\right]\,, \tag{2.9}$$

ou sa variante obtenue en remplaçant I_0 par $\widehat{I}_n(\theta_0)$. Le test obtenu ainsi est le *test des scores*. On choisit $z_{1-\alpha}$ comme pour $T_n^{(1)}$.

Exemple 1 (suite) *Supposons encore que $K = 1$ et que la variance $\Sigma = \sigma^2$ est connue. Alors :*

$$\begin{aligned} T_n^{(2)} &= \frac{1}{n}\left[\sum_{i=1}^{n} \frac{X_i - \mu_0}{\sigma^2}\right]^2 \sigma^2 \\ &= \frac{1}{n\sigma^2}\left[n(\overline{X} - \mu_0)\right]^2 \\ &= n\left(\frac{\overline{X} - \mu_0}{\sigma}\right)^2\,, \end{aligned}$$

ce qui redonne, dans ce cas particulièrement simple, la même région de rejet que le test de Wald.

Exercice 2.1 à 2.4 (suite) *Déterminer de même la forme du test des scores dans ces cas.*

Exercice 2.5 *Du point de vue numérique, quels sont les avantages et inconvénients respectifs du test de Wald et du test des scores si $K > 1$?*

Exercice 2.6 *Comment construire pour $K \geq 1$, à partir des résultats énoncés dans ce chapitre, des régions de confiance asymptotiques pour le paramètre à estimer ?*

Exercice 2.7 *Lorsque $K = 1$, comment construire, à partir des résultats énoncés dans ce chapitre, des tests asymptotiques unilatéraux de la forme H_0 $[\theta \leq \theta_0]$ contre H_1 $[\theta > \theta_0]$?*

2.5 Test du rapport des vraisemblances maximales

On souhaite toujours tester $[\theta = \theta_0]$ contre $[\theta \neq \theta_0]$. On note

$$\lambda_n = \frac{\mathcal{L}_n(\theta_0)}{\sup_\theta \ \mathcal{L}_n(\theta)} = \frac{\mathcal{L}_n(\theta_0)}{\mathcal{L}_n(\widehat{\theta}_n)} \tag{2.10}$$

et

$$\begin{aligned} T_n^{(3)} &= -2\ln\lambda_n = 2\left[\ln\mathcal{L}_n(\widehat{\theta}_n) - \ln\mathcal{L}_n(\theta_0)\right] \\ &= 2\left[L_n(\widehat{\theta}_n) - L_n(\theta_0)\right] \\ &= 2n\left[\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)\right] \end{aligned} \tag{2.11}$$

avec $\ell_n(\theta) = n^{-1}\sum_{i=1}^n \ln f_\theta(X_i)$. Sous les hypothèses usuelles de régularité, on sait que :

$$\left\{\begin{array}{l} \dot{\ell}_n(\widehat{\theta}_n) = 0 \\ \ddot{\ell}_n(\widehat{\theta}_n) \quad \text{est définie négative pour } n \text{ assez grand} \\ \ddot{\ell}_n(\widehat{\theta}_n) \xrightarrow{\text{p.s.}} -I(\theta_0) \quad \text{sous } \ H_0\,[\theta = \theta_0] \end{array}\right.$$

Formons le développement de Taylor à l'ordre 2 de $\ell_n(\theta)$ au voisinage de $\theta = \widehat{\theta}_n$, puis choisissons $\theta = \theta_0$:

$$\ell_n(\theta_0) = \ell_n(\widehat{\theta}_n) + \dot{\ell}_n(\widehat{\theta}_n)\,[\theta_0 - \theta_n] + \frac{1}{2}\ddot{\ell}_n(\widehat{\theta}_n)\left[\theta_0 - \widehat{\theta}_n,\, \theta_0 - \widehat{\theta}_n\right] + o_P\left(\|\widehat{\theta}_n - \theta_0\|^2\right)$$

Ainsi, puisque $\ell_n(\widehat{\theta}_n) = 0$,

$$T_n^{(3)} = T_n^{(1)} + o_P(1) \qquad (n \to +\infty)$$

Par conséquent, $T_n^{(3)}$ a même loi asymptotique que $T_n^{(1)}$ et $T_n^{(2)}$. Le test formé à partir de $T_n^{(3)}$ (avec les mêmes valeurs de $z_{1-\alpha}$) est le test du rapport des vraisemblances maximales.

Exemple 1 (suite) *Supposons encore $K = 1$ et $\Sigma = \sigma^2$ connue. Alors* :

$$\begin{aligned} L_n(\mu_0) &= \text{Cte} - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu_0)^2\, L_n(\widehat{\theta}_n) \\ &= \text{Cte} - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2, \end{aligned}$$

donc

$$\begin{aligned} T_n^{(3)} &= \frac{1}{\sigma^2}\sum_{i=1}^n (\overline{X} - \mu_0)\left[2X_i - (\overline{X} + \mu_0)\right] \\ &= \left(\frac{\overline{X} - \mu_0}{\sigma^2}\right)\left[\sum_{i=1}^n (X_i - \overline{X}) + \sum_{i=1}^n (X_i - \mu_0)\right] \\ &= n\left(\frac{\overline{X} - \mu_0}{\sigma^2}\right)^2 = T_n^{(1)} \end{aligned}$$

car

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

Exercice 2.1 à 2.4 (suite) *Déterminer de même la forme du test du rapport des vraisemblances maximales pour ces cas.*

Exercice 2.5 (suite) *Comparer, du point de vue numérique, le test* $T_n^{(3)}$ *aux tests* $T_n^{(1)}$ *et* $T_n^{(2)}$.

Conclusion

– Caractère simplement asymptotique ($n \to +\infty$).
– Aucune notion ni garantie d'optimalité.
– Retenir l'idée de statistique (asymptotiquement) pivotale.
– Retenir le rôle des statistiques exhaustives quand elles existent.
– Attention à bien vérifier les conditions de régularité : voir Monfort (1982).

Eléments bibliographiques pour ce chapitre

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Capéraà, P. & Van Cutsem, B. (1988) *Méthodes et modèles en Statistique non paramétrique*. Dunod : Paris.

Ferguson, T. S. (1967) *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press : New York and London.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Malinvaud, E. (1978) *Méthodes statistiques de l'économétrie*. Dunod : Paris, 3e édition.

Monfort, A. (1982) *Cours de Statistique mathématique*. Collection "Economie et statistiques avancées", Economica : Paris.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley : New York.

Tassi, P. (1985) *Méthodes statistiques*. Collection "Economie et statistiques avancées", Economica : Paris.

Ulmo, J. & Bernier, J. (1973) *Eléments de décision statistique*. Presses Universitaires de France : Paris.

3

Tests simples : optimalité

Introduction

On suppose que les lois μ_θ des v.a. qui nous intéressent, à valeurs dans un espace E, admettent chacune une densité $f_\theta : E \to \mathbb{R}$ relativement à une mesure de référence ν (σ-finie sur l'espace mesurable séparable $(E, \mathcal{E})$). Selon les cas, ν peut être une mesure de Lebesgue, la mesure de comptage sur un ensemble dénombrable, etc.

On souhaite tester l'hypothèse nulle simple H_0 $[\theta = \theta_0]$ contre l'alternative simple H_1 $[\theta = \theta_1]$, $\theta_1 \neq \theta_0$, au vu d'un n-échantillon $(x_1, \ldots, x_n) \in E^n = E \times \ldots \times E$ (n fois). Cela revient en fait à tester H_0 $[f = f_0]$ contre H_1 $[f = f_1]$. Par identifiabilité du modèle $\{\mu_\theta : \theta \in \Theta\}$, $\nu\{x \in E : f_0(x) \neq f_1(x)\} > 0$ puisque $\theta_1 \neq \theta_0$.

Pour simplifier, on supposera que f_0 et f_1 ont pour support E.

On ne considèrera que des tests déterministes. Cela signifie que l'on rejettera ou acceptera H_0 selon que $(x_1, \ldots, x_n)$ appartient à W_n ou à son complémentaire. On appelle W_n la *région de rejet* et W_n^c la *région d'acceptation*. On représente la prise de décision par une fonction $\phi_n : E^n \to [0, 1]$, en fait par $\phi_n = \mathbb{1}_{W_n}$, la fonction indicatrice de l'ensemble W_n.

Définition 3.1 *1.* ***Le risque de première espèce*** *est la probabilité de rejeter H_0 si $\theta = \theta_0$: c'est donc $\mathbb{E}_{\theta_0}[\phi_n(X_1, \ldots, X_n)]$.*

2. ***Le risque de seconde espèce*** *est la probabilité d'accepter H_0 si $\theta = \theta_1$: c'est donc $1 - \mathbb{E}_{\theta_1}[\phi_n(X_1, \ldots, X_n)]$.*
3. *La* ***puissance*** *du test est la probabilité de rejeter H_0 si $\theta = \theta_1$. C'est donc 1− (le risque de seconde espèce), soit $\mathbb{E}_{\theta_1}[\phi_n(X_1, \ldots, X_n)]$.*
4. *Soit $0 < \alpha < 1$. Le test est au (seuil ou au)* ***niveau de signification*** *α si $\mathbb{E}_{\theta_0}[\phi_n(X_1, \ldots, X_n)] \leq \alpha$.*
5. *Notion de* ***p-valeur****. Soit un test de statistique de test T. On effectue une expérience et on observe la réalisation $T = t$. Supposons que la région de rejet soit de la forme $\{T > c\}$. La p-valeur est la probabilité d'observer pour la loi de T sous H_0 une valeur supérieure à t (donc "encore pire"), c'est donc $p = \mathbb{P}(T > t)$. Il est plus précis de calculer cette p-valeur que de répondre simplement "rejet" ou "non rejet". Cela quantifie le rejet au sens suivant : si t est très grand, cette valeur a vraiment peu de chances d'être obtenue sous H_0, et la p-value est alors très petite (et bien sûr $p < \alpha$). Plus elle est petite, plus le rejet est fort. En ce sens, c'est une sorte d'inverse de "distance" à H_0 : si $p << \alpha$ on rejette ; si $p \approx \alpha$ il y a doute, on peut refaire une exprience ; si $p >> \alpha$ on ne peut pas rejeter H_0, aucun doute.*

Suivons le programme de la théorie des tests selon Neyman et Pearson : on se donne *a priori* un niveau de signification α, $0 < \alpha < 1$. On cherche, parmi tous les tests déterministes possibles de niveau α, celui (ou ceux) qui est optimal (sont optimaux), c'est-à-dire qui a (ont) la *puissance maximale*, ceci à taille d'échantillon n donnée.

On souhaite en outre que les tests obtenus soient *convergents*, c'est-à-dire que, pour α fixé,

$$\lim_{n \to +\infty} \mathbb{E}_{\theta_1}[\phi_n(X_1, \ldots, X_n)] = 1$$

et *sans biais*, c'est-à-dire que, pour tout $n \geq 1$,

$$\mathbb{E}_{\theta_1}[\phi_n(X_1, \ldots, X_n)] \geq \alpha$$

3.1 Lemme de Neyman-Pearson

Nous allons voir que, dans le cadre posé, on peut en général déterminer le test optimal pour chaque valeur de α, $0 < \alpha < 1$, et de $n \geq 1$, que la suite des tests obtenus est convergente et qu'ils sont sans biais.

Cette théorie repose sur un résultat général de maximisation sous contraintes convexes, ce qui fait intervenir des multiplicateurs a_j (penser aux multiplicateurs de Lagrange). il s'agit en effet de maximiser la puissance, c'est-à-dire $\mathbb{E}_{\theta_1}[\phi(X_1, \ldots, X_n)]$, sous la contrainte convexe $\mathbb{E}_{\theta_0}[\phi(X_1, \ldots, X_n)] \leq \alpha$. Autrement dit, on doit maximiser une intégrale de la forme $\int \phi g_{H_1} \, d\pi$ (la puissance) pour toutes les fonctions mesurables ϕ à valeurs dans l'intervalle

$[0, 1]$ (on vise ici les indicatrices des régions de rejet) qui vérifient la contrainte $\int \phi g_{H_0}\, d\pi \leq \alpha$ (test de niveau α). On commence par maximiser $\int \phi h\, d\pi$ sans contrainte (autre que le fait que ϕ doit prendre ses valeurs dans l'intervalle $[0, 1]$), c'est l'objet du lemme. Ensuite, on applique ce lemme à $h = g_{H_1} - a g_{H_0}$ (attention à l'inversion d'indices dans l'énoncé du théorème ci-dessous !), d'où essentiellement $\phi = \mathbb{1}_{\{g_{H_1} > a g_{H_0}\}}$. On détermine a par la condition $\int_{\{g_{H_1} > a g_{H_0}\}} g_{H_0}\, d\pi = \alpha$, quand c'est possible. Enfin, on démontre que cette construction permet effectivement de résoudre le problème de maximisation sous contrainte convexe posé.

Théorème 3.1 *Soient :*
– $(G, \mathcal{G})$ *un espace mesurable*
– π *une mesure positive sur* $(G, \mathcal{G})$
– $g_0, g_1, \ldots, g_p \in L^1(G, \pi)$, $p \geq 1$, *des fonctions à valeurs réelles*
– $a_1, \ldots, a_p \geq 0$ *des réels fixés*
– $\phi_0 : G \to [0, 1]$ *une fonction mesurable de la forme :*

$$\phi_0(x) = \begin{cases} 0 & \text{si} \quad g_0(x) < \sum_{j=1}^p a_j g_j(x) \\ \gamma(x) & \text{si} \quad g_0(x) = \sum_{j=1}^p a_j g_j(x) \\ 1 & \text{si} \quad g_0(x) > \sum_{j=1}^p a_j g_j(x) \end{cases} \tag{3.1}$$

Si $\phi : G \to [0, 1]$ *mesurable vérifie*

$$\int_G \phi g_j\, d\pi \leq \int_G \phi_0 g_j\, d\pi \quad \text{pour } 1 \leq j \leq p\,, \tag{3.2}$$

alors on a

$$\int_G \phi g_0\, d\pi \leq \int_G \phi_0 g_0\, d\pi \tag{3.3}$$

Démonstration. Elle repose sur le lemme élémentaire suivant.

Lemme 3.1 *Soit* $h \in L^1(G, \pi)$ *une fonction à valeurs réelles. La valeur maximale des intégrales de la forme*

$$\int_G \psi h\, d\pi, \qquad \psi : G \to [0, 1] \ \text{ mesurable}\,, \tag{3.4}$$

est atteinte par les fonctions de la forme

$$\psi_0(x) = \mathbb{1}_{\{h>0\}}(x) + \gamma(x)\, \mathbb{1}_{\{h=0\}}(x) \quad \pi - \text{p.p.}\,, \tag{3.5}$$

où $\gamma : G \to [0, 1]$ *mesurable est arbitraire.*

Pour établir ce lemme, il suffit de décomposer $\int_G \psi h\, d\pi$ en la somme :

$$\int_{\{h>0\}} \psi h\, d\pi + \int_{\{h=0\}} \psi h\, d\pi + \int_{\{h<0\}} \psi h\, d\pi$$

et de remarquer que puisque $\psi(x) \in [0,\, 1]$ pour tout x,

$$\int_{\{h<0\}} \psi h \, d\pi \leq 0$$

tandis que l'on a

$$\int_{\{h=0\}} \psi h \, d\pi = 0$$

■

A partir de là, formons $h = g_0 - \sum_{j=1}^{p} a_j g_j \in L^1(G,\, \pi)$. D'après le lemme 3.1,

$$\int_G \phi \left[g_0 - \sum_{j=1}^{p} a_j g_j \right] d\pi \leq \int_G \phi_0 \left[g_0 - \sum_{j=1}^{p} a_j g_j \right] d\pi \tag{3.6}$$

pour toute fonction mesurable $\phi : G \to [0,\, 1]$. Comme, dès lors,

$$\int_G \phi g_0 \, d\pi \leq \int_G \phi_0 g_0 \, d\pi - \sum_{j=1}^{p} a_j \left(\int_G \phi_0 g_j \, d\pi - \int_G \phi g_j \, d\pi \right) \leq \int_G \phi_0 g_0 \, d\pi \tag{3.7}$$

car les a_j sont positifs ou nuls et par l'hypothèse (3.2), la conclusion en résulte. ■

Corollaire 1 (Lemme de Neyman-Pearson) *On reprend le cadre décrit. Pour tout α, $0 < \alpha < 1$, donné et pour tout $n \geq 1$, le test déterministe de région de rejet*

$$W_n = \left\{ (x_1, \ldots, x_n) \in E^n : \prod_{i=1}^{n} f_1(x_i) > c_n \prod_{i=1}^{n} f_0(x_i) \right\},$$

où $c_n \in \mathbb{R}^+$ est une constante calculée pour que

$$\mathbb{E}_{\theta_0} \left[\phi(X_1, \ldots, X_n) \right] \leq \alpha$$

soit maximale parmi les fonctions $\phi = \mathbb{1}_{W_n}$ associées à des régions de rejet de cette forme, est le test déterministe de puissance maximale parmi les tests déterministes de niveau α.

Remarque 2 *Si, pour toute constante $c \in \mathbb{R}^+$, l'ensemble*

$$\left\{ (x_1, \ldots, x_n) : \prod_{i=1}^{n} f_1(x_i) = c \prod_{i=1}^{n} f_0(x_i) \right\}$$

est de ν-mesure nulle, alors c_n doit vérifier

$$\mathbb{E}_{\theta_0} \left[\phi(X_1, \ldots, X_n) \right] = \alpha$$

Démonstration du corollaire. On applique le théorème 3.1 avec $G = E^n$, $\pi = \nu \otimes \ldots \otimes \nu$ (n fois), $p = 1$, $a_1 = c_n$, $g_0 = \prod_{i=1}^{n} f_1(x_i)$ et $g_1 = \prod_{i=1}^{n} f_0(x_i)$: attention à l'interversion des indices 0 et 1 !

Supposons d'abord que pour toute constante $c \geq 0$,

$$\pi\{g_0 = cg_1\} = 0 \tag{3.8}$$

Dans ce cas, si $\phi_0(x_1, \ldots, x_n) = \mathbb{1}_{\{g_0 > c_n g_1\}}$, on a :

$$\int \phi g_0 \, d\pi = \mathbb{E}_{\theta_1}[\phi(X_1, \ldots, X_n)] \leq \mathbb{E}_{\theta_1}[\phi_0(X_1, \ldots, X_n)] = \int \phi_0 g_0 \, d\pi \tag{3.9}$$

pour toute fonction mesurable $\phi : E^n \to [0, 1]$ telle que

$$\int \phi g_1 \, d\pi = \mathbb{E}_{\theta_0}[\phi(X_1, \ldots, X_n)] \leq \mathbb{E}_{\theta_0}[\phi_0(X_1, \ldots, X_n)] = \int \phi_0 g_1 \, d\pi \,, \tag{3.10}$$

avec $\mathbb{E}_{\theta_0}[\phi_0(X_1, \ldots, X_n)] = \alpha$ par choix de la constante c_n : ce choix est possible car la fonction $c \geq 0 \mapsto \int_{\{g_0 > cg_1\}} g_1 \, d\pi$ est décroissante de 1 à 0 (0 correspond à $c \to +\infty$) et, par l'hypothèse $\pi\{g_0 = cg_1\} = 0$ pour tout c, elle est continue.

Dans le cas où $\pi\{g_0 = cg_1\}$ peut n'être pas nul pour certaines valeurs de c, le plus simple est d'imposer une nouvelle valeur de α, inférieure ou égale à la précédente, et à laquelle corresponde une valeur de la constante c. Voir les exemples. ■

Remarque 3 *Le cas $\pi\{g_0 = cg_1\} = 0$ pour tout $c \geq 0$ a lieu par exemple lorsque E est une partie "continue" d'un espace $\mathbb{R}^d$, ν est une mesure absolument continue par rapport à la mesure de Lebesgue, et les fonctions f_0 et f_1 sont assez régulières pour que les ensembles*

$$\prod_{i=1}^{n} f_1(x) = c \prod_{i=1}^{n} f_0(x_i)$$

soient des variétés différentiables (sphères, ellipsoïdes, hyperplans, etc.), à d'éventuels ensembles de mesure de Lebesgue nulle près. C'est le cas général pour les lois usuelles lorsque les variables sont continues. Voir les exemples.

Remarque 4 *Le résultat du théorème 3.1 peut suggérer d'introduire une procédure aléatoire si $\pi\{g_0 = cg_1\} = 0$: on pourrait envisager de choisir à la fois $c = c_n \geq 0$ et $\gamma_n \in [0, 1]$ (constantes) telles que le test non déterministe consistant à :*
- *rejeter H_0 si $\prod_{i=1}^n f_1(x_i) > c_n \prod_{i=1}^n f_0(x_i)$, et à :*
- *décider de rejeter H_0 avec probabilité γ_n et de l'accepter avec probabilité $1 - \gamma_n$ si*

$$\prod_{i=1}^{n} f_1(x_i) = c_n \prod_{i=1}^{n} f_0(x_i) \,,$$

soit de niveau exactement α. Cependant, nous ne chercherons pas à préciser davantage cette possibilité. Du point de vue pratique, elle ne présente en effet guère d'intérêt ! Voir les exemples.

Exemple 2 (Modèle gaussien $\mathcal{N}(\mu, \Sigma)$) *$[\mu = \mu_0]$ contre $[\mu \neq \mu_1]$ avec Σ connue. Pour simplifier, prenons $K = 1$ et $\Sigma = \sigma^2$. D'après le corollaire, on doit considérer le rapport des vraisemblances*

$$\frac{\mathcal{L}_n(\mu_1)}{\mathcal{L}_n(\mu_0)} = \prod_{i=1}^{n} \frac{f_{\mu_1}(x_i)}{f_{\mu_0}(x_i)} = \exp\left[\frac{\mu_1 - \mu_0}{\sigma^2}\left(\bar{x} - \frac{\mu_0 + \mu_1}{2}\right)\right]$$

Ainsi, selon que $\mu_1 > \mu_0$ ou $\mu_1 < \mu_0$, la région de rejet (ou région critique) est le demi-espace

$$\overline{x} > c \quad \text{ou} \quad \overline{x} < c$$

et $\{\overline{x} = c\}$ est de mesure de Lebesgue nulle. Supposons $\mu_1 > \mu_0$. La constante c est choisie pour que

$$\mathbb{P}\left[\overline{X} > c \mid H_0\right] = \alpha$$

Or, puisque les X_i sont i.i.d. $\mathcal{N}(\mu_0, \sigma^2)$ sous H_0, on a $\overline{X} \sim \mathcal{N}(\mu_0, \sigma^2/n)$, donc

$$\sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sigma}\right) \sim \mathcal{N}(0, 1)$$

Par conséquent,

$$n\left(\frac{\overline{X} - \mu_0}{\sigma}\right)^2 \sim \chi^2_1$$

On choisit c telle que :

$$\mathbb{P}\left[\sqrt{n}\left(\frac{\overline{X} - \mu_0}{\sigma}\right) > \sqrt{n}\left(\frac{c - \mu_0}{\sigma}\right) \middle| H_0\right] = \alpha,$$

c'est-à-dire que

$$1 - \Phi\left(\sqrt{n}\left(\frac{c - \mu_0}{\sigma}\right)\right) = \alpha\,,$$

ou encore

$$c = \mu_0 + \sigma\frac{u_\alpha}{\sqrt{n}}, \quad \text{avec} \quad 1 - \Phi(u_\alpha) = \alpha, \quad u_\alpha > 0$$

(On peut aussi exprimer c à l'aide de quantiles de χ^2_1, mais attention, le test considéré est unilatéral : $[\mu_1 > \mu_0]$.)

Il faut remarquer ici que :
- La région critique est déterminée par la statistique exhaustive.
- Pour chaque $n \geq 1$, on connaît la loi de la statistique de test, qui est $\overline{X}$: on a ainsi un premier exemple de test non asymptotique (on dit : "à distance finie").
- On peut calculer explicitement, en fonction de μ_1, la puissance du test : c'est

$$\mathbb{P}\left[\overline{X} > \mu_0 + \frac{\sigma u_\alpha}{\sqrt{n}} \middle| H_1\right],$$

avec $\overline{X} \sim \mathcal{N}\left(\mu_1, \sigma^2/n\right)$ sous H_1. C'est donc

$$\mathbb{P}\left[\mu_1 + \frac{\sigma\xi}{\sqrt{n}} > \mu_0 + \frac{\sigma u_\alpha}{\sqrt{n}}\right], \qquad \xi \sim \mathcal{N}(0, 1)$$

Cela vaut

$$\begin{aligned}\mathbb{P}\left[\frac{\sigma\xi}{\sqrt{n}} > -\Delta\mu + \frac{\sigma u_\alpha}{\sqrt{n}}\right] &= \mathbb{P}\left[\xi > -\frac{\Delta\mu\sqrt{n}}{\sigma} + u_\alpha\right] \\ &= \mathbb{P}\left[\xi < \frac{\Delta\mu\sqrt{n}}{\sigma} - u_\alpha\right] \\ &= \Phi\left[\frac{\Delta\mu\sqrt{n}}{\sigma} - u_\alpha\right]\end{aligned}$$

avec $\Delta\mu = \mu_1 - \mu_0 > 0$ par hypothèse. Cette puissance tend vers 1 quand $n \to +\infty$ (test convergent) et de plus, elle est $> \alpha$ dès que $\Delta\mu > 0$ (test sans biais).

→ Voir en Annexe B une table de Φ, f.r. de la loi normale $\mathcal{N}(0,\,1)$.

Exercice 2.1 (suite) *Modèle gaussien unidimensionnel* $\mathcal{N}(\mu,\,\sigma^2)$. *Test de* $\left[\sigma^2=\sigma_0^2\right]$ *contre* $\left[\sigma^2=\sigma_1^2\right]$ *, avec* μ *connue. Déterminer le test optimal de niveau* α. *Montrer qu'il est défini à l'aide de la statistique exhaustive* $n^{-1}\sum_{i=1}^n(X_i-\mu)^2$. *Préciser la loi de* $\sum_{i=1}^n(X_i-\mu)^2$ *sous* H_0, *puis sous* H_1. *Comment calculer la puissance de ce test ? Est-il convergent ? Sans biais ? C'est un test à distance finie.*

Exemple 3 *Voir Exercice 2.2, Modèle* Bernoulli (p). *Test de* $[p=p_0]$ *contre* $[p=p_1]$, $p_1\neq p_0$. *Dans ce cas, la fonction de vraisemblance s'écrit (relativement à la mesure de comptage)*

$$\mathcal{L}_n(p)=p^{S_n}(1-p)^{n-S_n}$$

où $S_n=\sum_{i=1}^n x_i$, *avec* $x_i=0$ *ou* 1 *(*$X_i=1$ *avec probabilité* p*). Donc*

$$\ln\left(\frac{\mathcal{L}_n\left(p_1\right)}{\mathcal{L}_n\left(p_0\right)}\right)=L_n(p_1)-L_n(p_0)$$
$$=S_n\ln p_1+(n-S_n)\ln\left(1-p_1\right)-S_n\ln p_0-(n-S_n)\ln\left(1-p_0\right)$$
$$=S_n\left[\ln\left(\frac{p_1}{1-p_1}\right)-\ln\left(\frac{p_0}{1-p_0}\right)\right]+n\ln\left(\frac{1-p_1}{1-p_0}\right)$$

La région critique est de la forme :

$$S_n>c\quad\text{si}\quad p_1>p_0$$
$$S_n<c\quad\text{si}\quad p_1<p_0$$

Supposons $p_1>p_0$ *et calculons* c. *Ici* $\{S_n=c\}$ *n'est pas de mesure nulle en général. On choisit l'entier* $c=k_0$ *le plus grand tel que*

$$\sum_{j=k_0+1}^{n}C_n^j\,p_0^j\,\left(1-p_0\right)^{n-j}\leq\alpha$$

Si le résultat est $<\alpha$, *on teste (à* $n\geq 1$ *donné) au niveau exact*

$$\alpha'=\alpha'(p_0,n)=\sum_{j=k_0+1}^{n}C_n^j\,p_0^j\,\left(1-p_0\right)^{n-j}$$

avec $k_0=k_0(p_0,\,n,\,\alpha)$ *déterminé ci-dessus. C'est aussi un test à distance finie.*

Exemple 4 *Voir Exercice 2.3, modèle* $\mathrm{Exp}(\theta)$. *Test de* $[\theta=\theta_0]$ *contre* $[\theta=\theta_1]$. *La région critique est de la forme*

$$\sum_{i=1}^{n}X_i>c\quad\text{si}\quad\theta_1>\theta_0$$
$$\sum_{i=1}^{n}X_i<c\quad\text{si}\quad\theta_1<\theta_0$$

On peut déterminer c *de manière exacte à partir de* α, *car la loi de* $\sum_{i=1}^n X_i$ *est une loi gamma sous* H_0. *(Voir Annexe A : définition de la loi gamma.) C'est un test à distance finie.*

Exemple 5 *Voir Exercice 2.4, Modèle* Poisson(λ). *Test de* $[\lambda = \lambda_0]$ *contre* $[\lambda = \lambda_1]$. *Dans ce cas,*

$$\frac{\mathcal{L}_n(\lambda_1)}{\mathcal{L}_n(\lambda_0)} = \mathrm{e}^{-n(\lambda_1-\lambda_0)} \left(\frac{\lambda_1}{\lambda_0}\right)^{\sum x_i}$$

La région critique est de la forme

$$\sum_{i=1}^{n} X_i > c \quad \text{si} \quad \lambda_1 > \lambda_0$$
$$\sum_{i=1}^{n} X_i < c \quad \text{si} \quad \lambda_1 < \lambda_0$$

Sous H_0, $\sum_{i=1}^{n} X_i \sim$ Poisson$(n\lambda_0)$. *On peut donc procéder de manière analogue à l'Exemple 3. C'est un test à distance finie.*

3.2 Résultats sur le test de Neyman-Pearson

Théorème 3.2 *Etant donné un test pour lequel existe une statistique exhaustive pour* θ, *pour tout test* ϕ *de niveau* α, *il existe un test* ϕ_0 *de niveau* α *construit sur la statistique exhaustive, dont la puissance est supérieure ou égale à celle de* ϕ.

Corollaire 2 *S'il existe une statistique exhaustive pour* θ, *le test de Neyman-Pearson s'exprime à l'aide de cette statistique exhaustive.*

En conséquence, on peut appliquer le lemme de Neyman-Pearson directement à la loi de la statistique exhaustive. C'est ce qu'illustrent les exemples ci-dessus.

D'autre part, on peut répondre aux questions posées.

Théorème 3.3 *Les tests simples de Neyman-Pearson sont sans biais.*

Quant à la convergence :

Théorème 3.4 *Pour les tests simples, si*

$$\nu\{x \in E : f_0(x) \neq f_1(x)\} > 0,$$

alors les tests de Neyman-Pearson sont convergents.

Conclusion

- Caractère non asymptotique : à distance finie.
- Garantie d'optimalité dans le cadre du schéma de Neyman-Pearson. Les tests obtenus sont sans biais et convergents.
- Retenir la méthode de démonstration du résultat fondamental de Neyman-Pearson.
- Notion de p-valeur.
- Retenir, ici aussi, le rôle des statistiques exhaustives quand elles existent.
- Limitation des résultats étudiés dans ce chapitre aux tests simples.

Eléments bibliographiques pour ce chapitre

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Ferguson, T. S. (1967) *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press : New York and London.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Monfort, A. (1982) *Cours de Statistique mathématique*. Collection "Economie et statistiques avancées", Economica : Paris.

Tassi, P. (1985) *Méthodes statistiques*. Collection "Economie et statistiques avancées", Economica : Paris.

Ulmo, J. & Bernier, J. (1973) *Eléments de décision statistique*. Presses Universitaires de France : Paris.

5. Tests simples : optimalité

4

Tests uniformément plus puissants : $\theta \in \mathbb{R}$

Introduction

Nous allons chercher à appliquer les résultats de Neyman-Pearson aux tests composés H_0 $[\theta \in \Theta_0]$ contre H_1 $[\theta \in \Theta_1]$, où au moins l'un des ensembles Θ_0 ou Θ_1 contient plus d'un élément (on parle alors d'hypothèse multiple).

Définition 4.1 *1. Un test composé de* ***niveau de signification*** α *vérifie*

$$\forall \theta \in \Theta_0, \quad \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right] \leq \alpha$$

Il est de ***niveau exactement*** α *si de plus il existe* $\theta \in \Theta_0$ *tel que*

$$\mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right] = \alpha$$

2. Un test composé ϕ_n *est* ***uniformément plus puissant (UPP)*** *de niveau* α *pour tester* H_0 $[\theta \in \Theta_0]$ *contre* H_1 $[\theta \in \Theta_1]$ *si* ϕ_n *est de niveau* α *et si pour tout autre test* ψ_n *de niveau* α *pour tester* H_0 *contre* H_1*, on a :*

$$\forall \theta \in \Theta_1, \quad \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right] \geq \mathbb{E}_\theta \left[\psi_n(X_1, \ldots, X_n)\right]$$

Exemple 1 (suite) *Modèle gaussien* $\mathcal{N}(\mu, \Sigma)$ *avec* $\Sigma = \sigma^2$ *connue et* $K = 1$*, et avec :*

$$\Theta_0 =]-\infty, \mu_0] \quad \text{et} \quad \Theta_1 =]\mu_0, +\infty[$$

On a vu au chapitre 3 que pour tester $[\mu = \mu_0]$*contre* $[\mu = \mu_1]$*,* $\mu_1 > \mu_0$*, le test optimal de niveau* α *était défini par*

$$\overline{X} > \mu_0 + \frac{\sigma u_\alpha}{\sqrt{n}}, \quad \text{avec} \quad 1 - \Phi(u_\alpha) = \alpha, \quad u_\alpha > 0$$

Observons que $\mu_0 + n^{-1/2}\sigma u_\alpha$ *ne dépend pas de* μ_1*. Par conséquent la même règle de décision fournit directement un test UPP pour tester* $[\mu = \mu_0]$ *contre* $[\mu > \mu_0]$ *au niveau* α*. Enfin, cette même règle fournit aussi un test UPP pour tester* $[\mu \leq \mu_0]$ *contre* $[\mu > \mu_0]$ *au niveau* α *exactement.*

Remarquons de plus que ce test est sans biais et convergent :

Définition 4.2 *Un test composé de niveau* α *est (strictement)* ***sans biais*** *si*

$$\forall \theta \in \Theta_1, \quad \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right] \geq \alpha \quad (>)$$

Remarquons que si la fonction puissance $\theta \mapsto \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right]$ est continue, si l'intersection de l'adhérence de Θ_0 et de l'adhérence de Θ_1 n'est pas vide, et si ϕ_n est sans biais de niveau α, alors il existe $\theta \in \Theta_0$ tel que $\mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right] = \alpha$ (donc le test est de niveau α exactement).

Maintenant, pourquoi a-t-on pu dans l'Exemple **1** passer directement du test de Neyman-Pearson de $[\theta = \theta_0]$ contre $[\theta = \theta_1]$, avec $\theta_1 > \theta_0$, à un test UPP de $[\theta \leq \theta_0]$ contre $[\theta > \theta_0]$? La raison en est double :

– La région de rejet du test de Neyman-Pearson dans ce cas est de la forme unilatérale

$$T_n(x1, \ldots, x_n) > c_n \,, \tag{4.1}$$

où $T_n(X_1, \ldots, X_n) = \overline{X}$ est une statistique exhaustive, et où la constante c_n dépend de θ_0, α, n, mais pas de $\theta_1 > \theta_0$.

– La fonction puissance $\theta \longmapsto \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right]$ est croissante sur $\mathbb{R}$.

Ceci conduit à la question : pour quels modèles paramétrés de lois de probabilité μ_θ, $\theta \in \mathbb{R}$, peut-on retrouver ces propriétés ? On suppose toujours que les lois μ_θ admettent une densité f_θ par rapport à une même mesure de référence ν sur l'espace $(E, \mathcal{E})$.

4.1 Rapport de vraisemblance monotone

Convenons de poser $f_{\theta_2}(x)/f_{\theta_1}(x) = +\infty$ si $f_{\theta_1}(x) = 0$ et $f_{\theta_2}(x) > 0$.

Définition 4.3 *La famille de lois $\mu_\theta(\,dx) = f_\theta(x)\,\nu(dx)$, $x \in E$, $\theta \in \mathbb{R}$, est à* ***rapport de vraisemblance monotone*** *(RVM) si pour tout $n \geq 1$, il existe une statistique*

$$T_n : E^n \to \mathbb{R}$$

telle que, pour tout $\theta_1 < \theta_2$, le rapport de vraisemblance

$$\frac{\mathcal{L}_n(\theta_2)}{\mathcal{L}_n(\theta_1)} = \prod_{i=1}^{n} \frac{f_{\theta_2}(x_i)}{f_{\theta_1}(x_i)}$$

soit une fonction monotone de $T_n(x_1, \ldots, x_n)$ pour les x_i pour lesquels ce rapport est défini.

Proposition 4.1 *Si*

$$f_\theta(x) = Z(\theta)^{-1}\, h(x)\, e^{a(\theta)\, t(x)}, \quad x \in E, \;\; \theta \in \mathbb{R}\,, \tag{4.2}$$

avec

$$Z(\theta) = \int_E h(x)\, \exp\left[a\,(\theta)\; t(x)\right]\, \nu(dx) > 0$$

pour tout θ, et si la fonction $\theta \mapsto a(\theta)$ est monotone, alors la famille des lois $f_\theta\, d\nu$ est à RVM.

Démonstration. Dans ce cas, on a, pour $\theta_1 < \theta_2$:

$$\frac{\mathcal{L}_n(\theta_2)}{\mathcal{L}_n(\theta_1)} = \frac{Z(\theta_1)}{Z(\theta_2)} \exp\left[(a(\theta_2) \,-\, a(\theta_1))\; T_n(x_1, \ldots, x_n)\right],$$

avec $T_n = \sum_{i=1}^n t(x_i)$. ■

Exemple 6 $\mathcal{N}(\mu,\, \sigma^2)$ *à* σ^2 *connue,* $\mathcal{N}(\mu,\, \sigma^2)$ *à* μ *connue,* $\mathrm{Bin}(n,\, p)$, $\mathrm{Exp}(\theta)$, $\mathrm{Poisson}(\lambda)$.

Contre-exemple : *Considérons sur $\mathbb{R}$ les lois de Cauchy de densité*

$$f_\theta(x) = \frac{1}{\pi}\; \frac{1}{1 + (x - \theta)^2}, \qquad \theta \in \mathbb{R}$$

Vérifier que pour $\theta_1 < \theta_2$, le rapport $f_{\theta_2}(x)/f_{\theta_1}(x)$ n'est pas monotone.

Théorème 4.1 *Si la famille de lois $d\mu_\theta = f_\theta \, d\nu$, $\theta \in \mathbb{R}$, est à RVM croissant en $T_n : E^n \to \mathbb{R}$, alors les tests dont les régions critiques sont de la forme*

$$W_n = \{(x_1, \ldots, x_n) \in E^n : T_n(x_1, \ldots, x_n) > c_n\} \tag{4.3}$$

sont UPP pour tester H_0 $[\theta \leq \theta_0]$ contre H_1 $[\theta > \theta_0]$, et la fonction puissance

$$\theta \mapsto \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right] ,$$

où $\phi_n = \mathbb{1}_{W_n}$, est croissante (strictement ou au sens large selon les cas).

Démonstration. Montrons d'abord la croissance de $\theta \longmapsto \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right]$. Soit $\theta_2 > \theta_1$.

On a :

$$\begin{aligned} E_{\theta_2}\left[\phi_n\left(X_1, \ldots, X_n\right)\right] &= \int_{\{T_n > c_n\}} \prod_{i=1}^n f_{\theta_2}\left(x_i\right) \quad \nu\left(dx_i\right) \\ &= \int_{\{T_n > c_n\}} \prod_{i=1}^n \left(\frac{f_{\theta_2}\left(x_i\right)}{f_{\theta_1}\left(x_i\right)}\right) \prod_{i=1}^n f_{\theta_1}\left(x_i\right) \quad \nu\left(dx_i\right) \\ &= \int_{\{T_n > c_n\}} G_n\left[T_n\left(x_1, \ldots, x_n\right)\right] \prod_{i=1}^n f_{\theta_1}\left(x_i\right) \quad \nu\left(dx_i\right) , \end{aligned}$$

où $G_n(T_n)$ est une fonction croissante de T_n. Pour simplifier, nous allons supposer que G_n est strictement croissante. D'après le lemme de Neyman-Pearson, le test le plus puissant de niveau α_1, $0 < \alpha_1 < 1$, de l'hypothèse simple $[\theta = \theta_1]$ contre l'alternative simple $[\theta = \theta_2]$ a une région critique de la forme

$$G_n\left[T_n(x_1, \ldots, x_n)\right] > \delta_n, \quad \text{soit} \quad T_n(x_1, \ldots, x_n) > d_n , \tag{4.4}$$

avec

$$\int_{\{G_n \circ T_n > \delta_n\}} \prod_{i=1}^n f_{\theta_1}(x_i) \quad \nu(dx_i) = \alpha_1 , \tag{4.5}$$

soit

$$\int_{\{T_n > d_n\}} \prod_{i=1}^n f_{\theta_1}(x_i) \quad \nu(dx_i) = \alpha_1 \tag{4.6}$$

On choisit α_1 de manière à ce que $d_n = c_n$, donc

$$\alpha_1 = \mathbb{E}_{\theta_1}\left[\phi_n(X_1, \ldots, X_n)\right] \tag{4.7}$$

Comme les tests de Neyman-Pearson sont sans biais, on a donc :

$$\int_{\{G_n \circ T_n > \delta_n\}} \prod_{i=1}^n f_{\theta_2}(x_i) \quad \nu(dx_i) \geq \alpha_1 , \tag{4.8}$$

ce qui permet de conclure.

Pour achever la démonstration du théorème, il reste à procéder comme dans l'Exemple 6 ci-dessus : les régions critiques W_n sont déterminées par la seule condition $\mathbb{E}_{\theta_0}[\mathbb{1}_{W_n}] = \mathbb{P}_{\theta_0}[T_n > c_n] = \alpha$, donc c_n ne dépend pas de θ_1, si on teste $[\theta = \theta_0]$ contre $[\theta = \theta_1]$, $\theta_1 > \theta_0$. Enfin, $\mathbb{E}_\theta[\phi_n] \leq \mathbb{E}_{\theta_0}[\phi_n] = \alpha$ pour $\theta \leq \theta_0$, d'après la première partie. ■

On peut appliquer ce résultat aux Exercices 2.1 à 2.4 et à l'exercice suivant.

Exercice 4.1 *Attention ! Dans cet exercice, les tests considérés peuvent inclure un tirage aléatoire. Soit μ_θ la loi uniforme sur $[0, \theta]$, $\theta > 0$, et soient $X_1, \ldots, X_n$ des v.a. i.i.d. de loi μ_θ.*

1. *Montrer que $T_n = \max_{1 \leq i \leq n} X_i$ est une statistique exhaustive pour θ.*
2. *Montrer que la famille des μ_θ, $\theta > 0$, est à RVM.*
3. *Quelle est la forme générale des tests de Neyman-Pearson pour tester $[\theta = \theta_0]$ contre $[\theta = \theta_1]$? $(\theta_1 > \theta_0)$.*
4. *Parmi ceux-ci, en existe-t-il qui sont indépendants de θ_1 ?*
5. *Existe-t-il des tests UPP de niveau α, $0 < \alpha < 1$, pour tester $[\theta \leq \theta_0]$ contre $[\theta > \theta_0]$?*

Exercice 4.2 *Adapter la démonstration ci-dessus au cas où la fonction G_n n'est pas strictement croissante (tout en restant croissante au sens large).*

Exercice 4.3 *Sous quelles conditions la fonction puissance est-elle strictement croissante ?*

4.2 Tests UPP de $[\theta_1 \leq \theta \leq \theta_2]$

On se limite ici au cas où μ_θ appartient à une famille exponentielle : $d\mu_\theta = f_\theta \, d\nu$, avec

$$f_\theta(x) = Z(\theta)^{-1} h(x) \mathrm{e}^{a(\theta)t(x)}, \quad x \in E, \quad t(x) \in \mathbb{R}, \quad \theta \in \mathbb{R} \tag{4.9}$$

En fait, $\theta \in \Theta$ partie de $\mathbb{R}$. On suppose la fonction $\theta \mapsto a(\theta)$ monotone. On peut montrer que si $\theta_1 \leq \theta_2$, il n'existe pas de test UPP de niveau α, $0 < \alpha < 1$, pour tester $\theta \in [\theta_1, \theta_2]$ contre $\theta \notin [\theta_1, \theta_2]$.

Par contre, on peut déterminer un test UPP si on se restreint à la classe des tests sans biais de niveau α.

Théorème 4.2 *Si $d\mu_\theta = f_\theta \, d\nu$, $\theta \in \Theta \subset \mathbb{R}$, appartient à une famille exponentielle, c'est-à-dire que f_θ est de la forme (4.9) avec $a(\theta)$ et $t(x)$ réels, si la fonction $\theta \mapsto a(\theta)$ est monotone, alors il existe des tests UPP dans la classe des tests sans biais pour tester H_0 ($\theta \in [\theta_1, \theta_2]$) contre H_1 ($\theta \notin [\theta_1, \theta_2]$), $\theta_1 \leq \theta_2$. Ces tests ont une région de rejet W_n ($n \geq 1$) de la forme :*

$$W_n = \{(x_1, \ldots, x_n) \in E^n : T_n < c_{1,n} \quad \text{ou} \quad T_n > c_{2,n}\}, \qquad c_{1,n} < c_{2,n},$$

où

$$T_n = \sum_{i=1}^{n} t(x_i) \tag{4.10}$$

Les deux constantes $c_{1,n}$ et $c_{2,n}$ sont déterminées par les conditions de niveau et de sans biais : si $\phi_n = \mathbb{1}_{W_n}$,

$$\mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right] = \alpha \quad \text{pour } \theta = \theta_1 \text{ et } \theta_2 \quad \text{si} \quad \theta_1 < \theta_2, \tag{4.11}$$

tandis que si $\theta_1 = \theta_2 = \theta_0$, ces conditions s'écrivent

$$\mathbb{E}_{\theta_0} \left[\phi_n(X_1, \ldots, X_n)\right] = \alpha \tag{4.12}$$

et

$$\frac{d}{d\theta} \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right]\Big|_{\theta=\theta_0} = 0 \tag{4.13}$$

Idée de la démonstration. Il s'agit essentiellement d'une application du théorème 3.1. On considère le cas où $\theta_1 < \theta_2$, le cas $\theta_1 = \theta_2 = \theta_0$ s'en déduisant par passage à la limite. C'est encore de la maximisation sous contraintes affines. Il s'agit de maximiser (on reprend les notations de la section 3), pour chaque $\theta \notin [\theta_1, \theta_2]$, la puissance

$$\int \phi g_0 \, d\pi = \mathbb{E}_\theta \left[\phi(X_1, \ldots, X_n)\right],$$

où $g_0 = \prod_{i=1}^n f_\theta(x_i)$, parmi les fonctions $\phi : E^n \to [0, 1]$ mesurables telles que

$$\begin{aligned} \int \phi g_1 \, d\pi &= \mathbb{E}_{\theta_1} \left[\phi(X_1, \ldots, X_n)\right] = \alpha \\ \int \phi g_2 \, d\pi &= \mathbb{E}_{\theta_2} \left[\phi(X_1, \ldots, X_n)\right] = \alpha, \end{aligned}$$

avec

$$g_1 = \prod_{i=1}^{n} f_{\theta_1}(x_i) \quad \text{et} \quad g_2 = \prod_{i=1}^{n} f_{\theta_2}(x_i)$$

D'après le théorème 3.1, le maximum est atteint pour $\phi = \phi_0$ de la forme

$$\phi_0(\underline{x}) = \begin{cases} 0 & \text{si} \quad g_0(\underline{x}) < \sum_{j=1}^2 b_j g_j(\underline{x}) \\ \gamma(\underline{x}) & \text{si} \quad g_0(\underline{x}) = \sum_{j=1}^2 b_j g_j(\underline{x}), \quad 0 \leq \gamma(\underline{x}) \leq 1 \\ 1 & \text{si} \quad g_0(\underline{x}) > \sum_{j=1}^2 b_j g_j(\underline{x}) \end{cases}$$

avec $\underline{x} = (x_1, \ldots, x_n)$, les multiplicateurs b_j ($j = 1, 2$) étant choisis pour que

$$\int \phi_0 g_j \, d\pi = \mathbb{E}_{\theta_j} \left[\phi_0(X_1, \ldots, X_n)\right] = \alpha$$

Ici, les multiplicateurs b_j ne sont plus nécessairement positifs : il faut alors remplacer les inégalités $\int \phi g_j \, d\pi \leq \int \phi_0 g_j \, d\pi$, $j = 1, \ldots, p$, par des égalités correspondantes dans le théorème 3.1. A nouveau, comme dans 4.1, on observe que les b_j $(j = 1, 2)$ ainsi calculés dépendent de $\theta_1 < \theta_2$, n, α, mais pas du $\theta \notin [\theta_1, \theta_2]$ considéré.

On reprend le même raisonnement pour $\theta_1 < \theta < \theta_2$, mais en remplaçant ϕ par $1 - \phi$: on trouve ainsi que pour tout $\theta_1 < \theta < \theta_2$,

$$\int \phi g_0 \, d\pi = \mathbb{E}_\theta \left[\phi(X_1, \ldots, X_n) \right] ,$$

où $g_0 = \prod_{i=1}^n f_\theta(x_i)$, est minimal pour $\phi = \phi_0$, sous les contraintes

$$\int \phi g_1 \, d\pi = \int \phi g_2 \, d\pi = \alpha \tag{4.14}$$

■

Exemple 1 (suite) : *Loi normale $\mathcal{N}(\mu, \sigma^2)$, σ^2 connue : H_0 $[\mu = \mu_0]$ contre $[\mu \neq \mu_0]$. Dans ce cas, la région d'acceptation est de la forme*

$$c_1 \leq \overline{X} \leq c_2$$

Comme $\overline{X} \sim \mathcal{N}\left(\mu_0, \sigma^2/n\right)$ sous H_0, la fonction puissance s'écrit ici, si on note $\sigma_n = \sigma/\sqrt{n}$:

$$\mu \longmapsto \frac{1}{\sqrt{2\pi}} \int_{(c_1-\mu)/\sigma_n}^{(c_2-\mu)/\sigma_n} e^{-\frac{u^2}{2}} \, du$$

donc c_1 et c_2 doivent d'abord vérifier

$$\frac{1}{\sqrt{2\pi}} \int_{(c_1-\mu_0)/\sigma_n}^{(c_2-\mu_0)/\sigma_n} e^{-\frac{u^2}{2}} \, du \, = \, 1 - \alpha \, ,$$

De plus,

$$\frac{d}{d\mu} \, \mathbb{E}_\mu \left[\phi_n \right] \Big|_{\mu=\mu_0} \, = \, 0$$

s'écrit :

$$e^{-\,(c_2-\mu_0)^2/2\sigma_n^2} \, - \, e^{-\,(c_1-\mu_0)^2/2\sigma_n^2} \, = \, 0$$

Il en résulte que c_1 et c_2 doivent être symétriques par rapport à μ_0, d'où

$$c_1 \, = \, \mu_0 \, - \, \frac{\sigma u_{\alpha/2}}{\sqrt{n}}, \qquad c_2 = \mu_0 \, + \, \frac{\sigma u_{\alpha/2}}{\sqrt{n}}$$

Exercice 2.1 : *Loi normale $\mathcal{N}(\mu, \sigma^2)$, μ connue : H_0 $\left[\sigma^2 = \sigma_0^2\right]$ contre H_1 $\left[\sigma^2 \neq \sigma_0^2\right]$. On se ramène au cas $\mu = 0$.*

1. *Poser $T_n = \sum_{i=1}^n X_i^2$, de sorte que $\sigma^{-2} T_n(X_1, \ldots, X_n)$ suit une loi χ_n^2. Exprimer la fonction puissance. Montrer que la première condition $c_1 \leq T_n \leq c_2$ s'écrit :*

$$\frac{1}{\Gamma\left(n/2\right)} \int_{c_1/\sigma_0^2}^{c_2/\sigma_0^2} e^{-\frac{u}{2}} \left(\frac{u}{2}\right)^{(n-1)/2} \frac{du}{2} \, = \, 1 - \alpha$$

2. *Montrer que la seconde condition* $\left(\frac{d}{d\sigma^2} \mathbb{E}_{\sigma^2} [\phi_n] \big|_{\sigma^2=\sigma_0^2} = 0 \right)$ *s'écrit :*

$$e^{-c_1/2\sigma_0^2} \, c_1^{n/2} \; = \; e^{-c_2/2\sigma_0^2} \, c_2^{n/2}$$

3. *Par le TLC,* $T_n(X_1, \ldots, X_n)$ *suit approximativement une loi normale* $\mathcal{N}(n\sigma^2, 2n\sigma^4)$ *quand* n *est assez grand. En déduire que* c_1, c_2 *doivent vérifier approximativement deux conditions qui peuvent s'écrire en termes des fonctions* Φ *et* $\varphi = \Phi'$. *Montrer que si* σ_0^2 *est assez petit, il est raisonnable de prendre*

$$\begin{cases} c_1/n \approx \sigma_0^2 \, (1 \, - \, 2u_{\alpha/2}\sigma_0^2) \\ c_2/n \approx \sigma_0^2 \, (1 \, + \, 2u_{\alpha/2}\sigma_0^2) \end{cases}$$

On pourra de même déterminer la région de rejet pour tester $p \leq p_0$ contre $p > p_0$ dans le cas du modèle de Bernoulli, pour tester $\theta \leq \theta_0$ contre $\theta > \theta_0$ dans le cas du modèle exponentiel, et pour tester $\lambda \leq \lambda_0$ contre $\lambda > \lambda_0$ dans le cas du modèle de Poisson. On pourra aussi chercher la région de rejet pour des tests bilatéraux.

Dernier point. Il faudrait maintenant réviser la notion d'intervalle de confiance. Se souvenant des relations naturelles de dualité entre régions de rejet des tests et régions de confiance, on pourra déduire des résultats de ce chapitre une définition appropriée des intervalles de confiance, puis vérifier qu'ils ne sont pas nécessairement symétriques ni de longueur minimale.

Exercice 4.4 *Revenons au cas de la loi normale* $\mathcal{N}(\mu, \sigma^2)$, μ *connue :* $H_0 \left[\sigma^2 = \sigma_0^2\right]$ *contre* $H_1 \left[\sigma^2 \neq \sigma_0^2\right]$, *avec* $\mu = 0$. *Pour l'estimateur de* σ^2 *défini par* $\widehat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n X_i^2$, *déterminer l'intervalle de confiance de niveau de confiance* α *construit à partir des résultats de l'exercice précédent. Déterminer ensuite l'intervalle de confiance symétrique, puis celui de longueur minimale. Comparer.*

Exercice 4.5 *Un institut de sondage se pose la question suivante. Il va interroger demain des électrices et électeurs à la sortie des bureaux de vote, et leur demander s'ils ont voté pour le candidat A ou pour le candidat B. Il suppose qu'il n'y aura pas de vote blanc ou nul. Il devra ensuite donner un pronostic en attendant les résultats définitifs. Comme sa crédibilité a été fortement entamée par une suite de résultats qui se sont avérés erronés (et qui ont été montés en épingle par la presse satirique), il souhaite être très précis. Quel est le nombre minimum de personnes qu'il doit interroger pour que, si au moins* 51% *des personnes sondées déclarent avoir voté pour A, par exemple, la probabilité d'en déduire avec raison qu'il est le vainqueur soit supérieure ou égale à* 98% *?*

Exercice 4.6 *Un statisticien audacieux doit effectuer un test bilatéral au niveau* 5% *sur le paramètre de moyenne* μ *d'une loi normale de variance connue* σ_0^2. *Il teste l'hypothèse nulle* $\mu = \mu_0$ *contre l'alternative* $\mu \neq \mu_0$. *Il trouve que* $\mu = \mu_0$ *doit être rejetée au niveau considéré. Comme il dispose de la valeur de l'estimateur* $\widehat{\mu}_n$ *de* μ, *il a l'idée de comparer cette valeur à la valeur qui fait l'objet du test,* μ_0. *Il observe que dans son cas, la valeur de l'estimateur est plus grande que la valeur testée. Il en déduit que non seulement l'hypothèse nulle* $\mu = \mu_0$ *a été rejetée, mais que, de plus, on peut affirmer que* $\mu > \mu_0$. *Que penser de cette conclusion ?*

Conclusion

- Notion de fonction (de) puissance. Cas où cette fonction est monotone. Rapport de vraisemblance monotone.
- Garantie d'optimalité : tests uniformément plus puissants (et sans biais dans le cas bilatéral) ; possibilité de construire des tests à distance finie.
- Bien distinguer les tests unilatéraux des tests bilatéraux.
- Retenir le rôle des statistiques exhaustives et des familles exponentielles.
- Relations entre régions de rejet et intervalles de confiance.

Eléments bibliographiques pour ce chapitre

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Ferguson, T. S. (1967) *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press : New York and London.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Monfort, A. (1982) *Cours de Statistique mathématique*. Collection "Economie et statistiques avancées", Economica : Paris.

Tassi, P. (1985) *Méthodes statistiques*. Collection "Economie et statistiques avancées", Economica : Paris.

Ulmo, J. & Bernier, J. (1973) *Eléments de décision statistique*. Presses Universitaires de France : Paris.

4. Tests uniformément plus puissants : $\theta \in \mathbb{R}$

5

Tests à structure de Neyman : $\theta \in \mathbb{R}^K$

Introduction

Soit $\theta = (\theta_1, \ldots, \theta_K)$, $K \geq 1 : \theta \in \Theta$, avec $\Theta \subset \mathbb{R}^K$.

Nous allons nous intéresser à des tests concernant seulement une des coordonnées θ_j de θ, les autres coordonnées étant inconnues (exemple : test pour $\mathcal{N}(\mu, \sigma^2)$ sur le paramètre μ, la variance σ^2 étant inconnue), ou une égalité du type $\theta_j = \theta_k$ contre $\theta_j >$ (ou $<$) θ_k, ou bien $\theta_j \neq \theta_k$, les autres coordonnées étant inconnues. L'hypothèse nulle a la forme H_0 $[\theta \in \Theta_0]$.

Pour ce type de test, il n'existe pas en général de test UPP. Nous allons donc devoir, comme au chapitre 4, nous restreindre à une classe plus petite (mais contenant les tests d'intérêt majeur !) de tests : les *tests α-semblables.*

Définition 5.1 *Un test ϕ_n est* ***semblable de niveau*** *α (ou α-semblable)* ***sur une partie*** *Θ_B de Θ_0 si le risque de première espèce est constant et égal à α sur Θ_B :*

$$\forall \theta \in \Theta_B, \quad \mathbb{E}_\theta \left[\phi_n(X_1, \ldots, X_n)\right] = \alpha$$

Exemple 7 *Soient $X_1, \ldots, X_n$ des v.a. i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. On souhaite tester H_0 $[\sigma^2 = \sigma_0^2$, μ quelconque$]$ contre H_1 $[\sigma^2 > \sigma_0^2$, μ quelconque$]$. Or, la statistique*

$$S_n^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{5.1}$$

suit une loi χ^2_{n-1}, qui ne dépend pas de μ. Sous H_0, elle ne dépend que de σ_0^2. Ainsi, tout test d'hypothèse nulle H_0 est α-semblable sur $\Theta_B = \left\{\left(\mu, \sigma_0^2\right) : \mu \in \mathbb{R}\right\} = \Theta_0$, s'il s'appuie sur la statistique S_n^2.

Nous allons voir comment exploiter l'existence de statistiques exhaustives $T(X_1, \ldots, X_n)$, en particulier dans le cadre des familles exponentielles, pour former des tests conditionnellement (sachant T) semblables UPP, et passer de là à des tests semblables non conditionnellement UPP.

5.1 Tests à structure de Neyman

Nous allons commencer par présenter cette idée de conditionnement par une statistique exhaustive sur un exemple.

Exemple 8 *Soient $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. On souhaite tester H_0 $[\mu \leq 0]$ contre H_1 $[\mu > 0]$, σ^2 étant inconnue. Comme*

$$f_{\mu,\,\sigma^2}(x) = Z^*(\mu, \sigma^2)^{-1} \ \exp\left[\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right], \tag{5.2}$$

la fonction de vraisemblance vaut

$$\theta \longmapsto \mathcal{L}_n(\theta) = \prod_{i=1}^{n} f_{\mu,\,\sigma^2}(X_i) = Z(\theta)^{-n} \exp\left[\theta_1 T_1 + \theta_2 T_2\right], \tag{5.3}$$

avec $\theta = (\theta_1, \theta_2)$, $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/2\sigma^2$, $T_1 = \sum_{i=1}^n X_i$, $T_2 = \sum_{i=1}^n X_i^2$. Tester $[\mu \leq 0]$ revient à tester $[\theta_1 \leq 0]$. Sous l'hypothèse $\theta_1 = 0$, la loi jointe de $(X_1, \ldots, X_n)$ conditionnellement à $T_2 = \sum_{i=1}^n X_i^2 = t_2 > 0$ est, par symétrie sphérique, uniforme sur la sphère de dimension $n-1$, de rayon $\sqrt{t_2}$ et de centre 0, $S(0, \sqrt{t_2})$: on a alors la représentation

$$X_i = \sqrt{t_2}\, V_i, \quad i = 1, \ldots, n, \tag{5.4}$$

où $(V_1, \ldots, V_n)$ désigne une v.a. uniforme sur la sphère $S(0, 1)$, donc de loi indépendante de θ et de (t_1, t_2).

D'autre part, conditionnellement à $T_2 = t_2 > 0$, la vraisemblance ne dépend plus du paramètre θ_2, mais seulement de θ_1 :

$$\mathcal{L}_n(\theta \mid T_2 = t_2) = Z_1(\theta_1, t_2)^{-n}\, e^{\theta_1 T_1}, \tag{5.5}$$

voir le lemme 5.1 ci-dessous. C'est un des points cruciaux ! Du coup, d'après le chapitre 4 (on est en effet ramené au cas d'un seul paramètre), le test conditionnel sachant $T_2 = t_2 > 0$, UPP pour $[\theta_1 \leq 0]$ contre $[\theta_1 > 0]$, est de la forme :

$$\phi_n(t_1, t_2) = \begin{cases} 1 & \text{si } t_1 > c(t_2) \quad (\text{rejet de } H_0) \\ 0 & \text{si } t_1 < c(t_2) \end{cases} \tag{5.6}$$

Nous devons, pour tout $t_2 > 0$, choisir $c(t_2)$ pour que

$$\begin{aligned} \forall \theta_2, \quad \mathbb{E}_{(0,\,\theta_2)}\left(\phi_n \mid T_2 = t_2\right) &= \mathbb{P}_{(0,\,\theta_2)}[T_1 > c(t_2) \mid T_2 = t_2] \\ &= \mathbb{P}\left[\sum_{i=1}^{n} V_i > \frac{c(t_2)}{\sqrt{t_2}}\right] = \alpha \end{aligned} \tag{5.7}$$

Comme la loi de $(V_1, \ldots, V_n)$ est indépendante de θ et de t_2 (c'est l'autre point crucial !), on obtient

$$\frac{c(t_2)}{\sqrt{t_2}} = r_\alpha \text{ indépendant de } \theta \text{ et de } t_2 \tag{5.8}$$

Il reste à déconditionner, ce qui redonnera

$$\forall \theta_2, \quad \mathbb{E}_{(0,\,\theta_2)}\left[\phi_n\right] = \mathbb{E}_{(0,\,\theta_2)}\left(\mathbb{E}_{\theta_1=0}\left[\phi_n \mid T_2\right]\right) = \alpha \tag{5.9}$$

On aurait de même $\mathbb{E}_\theta\left[\phi_n\right] \leq \alpha$ pour $\theta_1 \leq 0$. Revenons à la forme du test obtenu :

$$\phi_n(t_1,\, t_2) = \begin{cases} 1 & \text{si} \quad t_1/\sqrt{t_2} > r_\alpha \\ 0 & \text{si} \quad t_1/\sqrt{t_2} < r_\alpha \end{cases} \tag{5.10}$$

Ce test est classiquement écrit en termes des statistiques exhaustives habituelles :

$$\begin{cases} \overline{x} = t_1/n \\ s^2 = \sum_{i=1}^n (x_i - \overline{x})^2/n = (t_2/n) - \overline{x}^2 \end{cases} \tag{5.11}$$

On vérifie que $t_1/\sqrt{t_2} > \text{Cte} \Leftrightarrow \sqrt{n-1}\,(\overline{x}/s) > \text{Cte}$. Sous l'hypothèse $\theta_1 = 0$, $\sqrt{n}(\overline{X}/\sigma) \sim \mathcal{N}(0,\, 1)$ et $nS^2/\sigma^2 \sim \chi^2_{n-1}$ sont des v.a. indépendantes, donc

$$\sqrt{n-1}\,\frac{\overline{X}}{S} = \frac{\sqrt{n}(\overline{X}/\sigma)}{\sqrt{n}(S/\sigma\sqrt{n-1})} \tag{5.12}$$

suit une loi de Student à $n-1$ degrés de liberté. Une loi de Student à ν degrés de liberté est, par définition, la loi de la v.a.

$$T_\nu = \frac{Y}{\sqrt{Z/\nu}}, \quad Y \sim \mathcal{N}(0,\, 1), \quad Z \sim \chi^2_\nu, \quad Y \text{ et } Z \text{ indépendantes}$$

Elle admet la densité

$$f_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \frac{1}{\left(\frac{t^2}{\nu} + 1\right)^{\frac{\nu+1}{2}}},$$

ν entier ≥ 1.

Finalement, le test a la forme

$$\phi_n(t_1,\, t_2) = \begin{cases} 1 & \text{si} \quad \sqrt{n-1}\,\overline{x}/s > c_\alpha \\ 0 & \text{si} \quad \sqrt{n-1}\,\overline{x}/s < c_\alpha \end{cases} \tag{5.13}$$

La constante c_α est calculée à partir d'une table de la loi de Student à $n-1$ degrés de liberté :

$$\int_{c_\alpha}^{\infty} f_{n-1}(t)\, dt = \alpha \tag{5.14}$$

Si l'on voulait tester $[\mu \leq \mu_0]$ contre $[\mu > \mu_0]$, σ^2 quelconque, on remplacerait simplement $\overline{x}$ par $\overline{x} - \mu_0$.

Pour tester H_0 $[\mu = \mu_0]$ contre H_1 $[\mu \neq \mu_0]$, on procède de manière analogue. Cela conduit à rejeter H_0 si

$$\sqrt{n-1}\,\frac{\overline{x} - \mu_0}{s} > c_{\alpha/2} \quad \text{ou} \quad < -c_{\alpha/2} \tag{5.15}$$

Cet exemple nous indique le point important, qui est le conditionnement.

Définition 5.2 *Soit T_n une statistique exhaustive du modèle paramétré considéré. Un test ϕ_n est dit à* ***structure de Neyman*** *si*

$$\mathbb{E}_\theta\left[\phi_n(X_1,\ldots,X_n) \mid T_n = t\right] = \alpha \tag{5.16}$$

pour tout t (sauf peut-être pour t dans un ensemble de μ_θ-mesure nulle), pour tout $\theta \in \Theta_B$, où Θ_B est le bord de Θ_0 relativement à l'alternative Θ_1, c'est-à-dire $\Theta_B = \overline{\Theta}_0 \cap \overline{\Theta}_1$.

Par la formule de déconditionnement, si un tel test existe, il est nécessairement α-semblable. Nous allons maintenant nous intéresser à l'existence de tels tests, en particulier dans le cas de familles exponentielles en paramétrage naturel.

5.2 Structure de Neyman et structure exponentielle

Cela nous conduit à reprendre très brièvement certains aspects de la théorie des familles exponentielles en paramétrage naturel, $d\mu_\theta = f_\theta\, d\nu$ sur E,

$$f_\theta(x) = Z(\theta)^{-1} h(x) \exp\left[\sum_{j=1}^{K} \theta_j t_j(x)\right], \qquad x \in E \tag{5.17}$$

Définition 5.3 *L'espace naturel des paramètres Θ_N est l'ensemble des θ pour lesquels la fonction $x \mapsto h(x)\exp[\sum_{j=1}^K \theta_j t_j(x)]$ est ν-intégrable.*

On a les propriétés suivantes :

- *L'espace naturel des paramètres Θ_N est convexe.*
- *Si l'intégrale*

$$\int \ldots \int |\phi(x_1,\ldots,x_n)| \exp\left[\sum_{j=1}^{K} \theta_j \sum_{i=1}^{n} t_j(x_i)\right] \prod_{i=1}^{n} h(x_i)\ \nu(dx_i) < +\infty$$

  pour tout $\theta \in \Theta_N$, alors cette intégrale est une fonction analytique de θ dans l'intérieur $\overset{o}{\Theta}_N$ de Θ_N, que l'on suppose non vide.
  En particulier, $Z(\theta)$ est analytique en $\theta \in \overset{o}{\Theta}_N$, ainsi que $\mathbb{E}_\theta\left[\phi(X_1,\ldots,X_n)\right]$ pour ϕ mesurable bornée.
- *Pour $\theta \in \overset{o}{\Theta}_N$, on a :*

$$\mathbb{E}_\theta\left(t_j(X)\right) = \frac{\partial}{\partial \theta_j} \ln\ Z(\theta)$$

$$\mathrm{Cov}_\theta\left(t_j(X),\, t_{j'}(X)\right) = \frac{\partial^2}{\partial \theta_j\ \partial \theta_{j'}} \ln Z(\theta)$$

– *Si* $n \geq K$ *et* $\overset{o}{\Theta}_N \neq \emptyset$, *alors*

$$\underline{T} = (T_1, \ldots, T_K)$$

avec

$$T_j = T_j(X_1, \ldots, X_n) = \sum_{i=1}^{n} t_j(X_1, \ldots, X_n) \tag{5.18}$$

est une statistique exhaustive pour θ. *De plus, cette statistique exhaustive est* ***complète****, ou* ***totale****, c'est-à-dire que pour toute fonction* $\underline{g}$, *si la fonction*

$$\theta \in \Theta_N \longmapsto \mathbb{E}_\theta \left(\underline{g}(\underline{T}) \right) \tag{5.19}$$

est nulle dès qu'elle est bien définie, alors

$$\forall \theta \in \Theta_N, \quad \mathbb{P}_\theta \left[\underline{g}(\underline{T}) = 0 \right] = 1 \tag{5.20}$$

Cela tient à ce que si la transformée de Laplace d'une fonction est nulle sur un ouvert non vide, alors la fonction est nulle p.p.

– *Pour chaque* θ_1 *(première coordonnée de* θ*) fixé,* $(T_2, \ldots, T_K)$ *est une statistique exhaustive pour* $(\theta_2, \ldots, \theta_K)$.

– *Considérons la loi de* $\underline{T} = (T_1, \ldots, T_K)$ *: on peut montrer qu'elle admet, relativement à la mesure de Lebesgue sur* $\mathbb{R}^K$, *une densité de la forme*

$$f_{\underline{T}}(\underline{t} \mid \theta) = \left(Z_{\underline{T}}(\theta) \right)^{-1} h_{\underline{T}}(\underline{t}) \exp \left[\sum_{j=1}^{K} \theta_j t_j \right] \tag{5.21}$$

Dans ces conditions, la densité de la loi conditionnelle de T_1 *sachant* $(T_2, \ldots, T_K)$ *est donnée par :*

$$f_{T_1 \mid T_2 = t_2, \ldots, T_K = t_K}(t_1 \mid \theta_1) = \left(\int h_{\underline{T}}(\underline{t}) \, \mathrm{e}^{\theta_1 t_1} \, dt_1 \right)^{-1} h_{\underline{T}}(\underline{t}) \, \mathrm{e}^{\theta_1 t_1} \tag{5.22}$$

Elle n'est donc paramétrée qu'en θ_1.

Pour avoir davantage de détails on peut consulter les ouvrages cités en références bibliographiques ci-dessous, par exemple Monfort (1982).

Nous pouvons maintenant chercher à établir le résultat cherché dans un cadre général.

Théorème 5.1 *Avec les notations et hypothèses ci-dessus, les tests déterministes UPP pour* H_0 $[\theta_1 \leq \theta_{0,1}]$ *contre* H_1 $[\theta_1 > \theta_{0,1}]$, *avec* $(\theta_2, \ldots, \theta_K)$ *quelconques,* α*-semblables sur le bord*

$$\Theta_B = \{\theta = (\theta_{0,1}, \theta_2, \ldots, \theta_K)\}$$

de Θ_0 *relativement à l'alternative* Θ_1, *sont de la forme*

$$\phi(t_1, \ldots, t_K) = \begin{cases} 1 & \text{si} \quad t_1 > c(t_2, \ldots, t_K) \\ 0 & \text{si} \quad t_1 < c(t_2, \ldots, t_K) \end{cases}$$

Pour tout $(t_2, \ldots, t_K)$, *la constante (en* t_1*)* $c\,(t_2, \ldots, t_K)$ *est choisie de sorte que*

$$\int_{c(t_2, \ldots, t_K)}^{\infty} f_{\theta_{0,1}}(t_1 \,|\, t_2, \ldots, t_K) \; \lambda_1(dt_1) = \alpha$$

Pour exposer schématiquement la théorie, nous allons simplifier en supposant que $\Theta = R^K$ ($K \geq 2$), et que l'on teste H_0 $[\theta_1 \leq 0]$ contre H_1 $[\theta_1 > 0]$, les autres paramètres n'étant pas spécifiés. En outre, nous supposerons que la loi jointe des statistiques exhaustives $T_1, \ldots, T_K$ (on notera $\underline{T} = (T_1, \ldots, T_K)$ et $\underline{t} = (t_1, \ldots, t_K)$) admet une densité $f_\theta(t_1, \ldots, t_K)$ relativement à la mesure de référence produit $\underline{\lambda}(d\underline{t}) = \lambda_1(dt_1) \otimes \ldots \otimes \lambda_K(dt_K)$, et que la densité conditionnelle

$$f_\theta(t_1 \,|\, t_2, \ldots, t_K) = \frac{f_\theta(t_1, \ldots, t_K)}{\int f_\theta(s, t_2, \ldots, t_K)\ \lambda_1(ds)} \tag{5.23}$$

est bien définie pour tout $(t_1, \ldots, t_K)$ et ne dépend que de θ_1 : on la notera $f_{\theta_1}(t_1 | t_2, \ldots, t_K)$.

Soit $\phi(t_1, \ldots, t_K) = \mathbb{1}_W(t_1, \ldots, t_K)$, W désignant une région de rejet. On lui associe la famille de *tests conditionnels* suivants. Pour tout $(t_2, \ldots, t_K)$ fixé, soit $W(t_2, \ldots, t_K)$ la partie mesurable de $\mathbb{R}$ définie par

$$t_1 \in W(t_2, \ldots, t_K) \Leftrightarrow (t_1, \ldots, t_K) \in W\,,$$

et soit

$$\phi(t_1 \,|\, t_2, \ldots, t_K) \ = \ \mathbb{1}_{W(t_2, \ldots, t_K)}(t_1) \tag{5.24}$$

Réciproquement, si à tout $(t_2, \ldots, t_K)$ on associe une partie mesurable $W(t_2, \ldots, t_K)$ de $\mathbb{R}$, on définit le test ϕ correspondant par sa région de rejet

$$W = \{(t_1, \ldots, t_K) \,:\, t_1 \in W(t_2, \ldots, t_K)\} \tag{5.25}$$

Par définition, le test conditionnel $t_1 \mapsto \phi(t_1 | t_2, \ldots, t_K)$ pour tester H_0 $[\theta_1 \leq 0]$ contre H_1 $[\theta_1 > 0]$ avec $(t_2, \ldots, t_K)$ fixé, est de niveau de signification α si pour tout $\theta_1 \leq 0$,

$$\mathbb{E}_{\theta_1}[\phi(T_1 | t_2, \ldots, t_K) \,|\, T_2 = t_2, \ldots, T_K = t_K] = \int_{W(t_2, \ldots, t_K)} f_{\theta_1}(s | t_2, \ldots, t_K)\ \lambda_1(ds) \leq \alpha \tag{5.26}$$

Il est sans biais si

$$\forall \theta_1 > 0, \quad \mathbb{E}_{\theta_1}[\phi(T_1 | t_2, \ldots, t_K) \,| T_2 = t_2, \ldots, T_K = t_K] > \alpha \tag{5.27}$$

On sait remonter des propriétés vraies pour $\lambda_2 \otimes \ldots \otimes \lambda_K$-presque tout $(t_2, \ldots, t_K)$ des tests conditionnels aux propriétés analogues des tests non conditionnels.

Lemme 5.1 *1. Si, pour $\lambda_2 \otimes \ldots \otimes \lambda_K$-presque tout $(t_2, \ldots, t_K)$ le test conditionnel $t_1 \mapsto \phi(t_1 | t_2, \ldots, t_K)$ pour tester H_0 $[\theta_1 \leq 0]$ contre H_1 $[\theta_1 > 0]$ est de niveau de signification α, alors il en est de même pour le test non conditionnel associé, quels que soient $\theta_2, \ldots, \theta_K$ non spécifiés.*

2. Si, pour $\lambda_2 \otimes \ldots \otimes \lambda_K$-presque tout $(t_2, \ldots, t_K)$ le test conditionnel $t_1 \mapsto \phi(t_1 | t_2, \ldots, t_K)$ est sans biais, alors il en est de même pour le test non conditionnel associé.

3. Si, pour $\lambda_2 \otimes \ldots \otimes \lambda_K$-presque tout $(t_2, \ldots, t_K)$ le test conditionnel $t_1 \mapsto \phi(t_1 | t_2, \ldots, t_K)$ est de niveau de signification α exactement, atteint pour $\theta_1 = 0$, alors le test non conditionnel associé est α-semblable sur le bord $\theta_1 = 0$, quels que soient $\theta_2, \ldots, \theta_K$ non spécifiés.

4. Si, pour $\lambda_2 \otimes \ldots \otimes \lambda_K$-presque tout $(t_2, \ldots, t_K)$ le test conditionnel $t_1 \mapsto \phi(t_1 | t_2, \ldots, t_K)$ est plus puissant que le test conditionnel $t_1 \mapsto \psi(t_1 | t_2, \ldots, t_K)$, alors il en est de même pour les tests non conditionnels ϕ et ψ associés.

Démonstration. Elle repose directement sur l'hypothèse que la densité conditionnelle $f_\theta(t_1|\,t_2,\ldots,t_K)$ ne dépend que de θ_1, et sur la formule de déconditionnement :

$$\begin{aligned}
\mathbb{E}_\theta\left[\phi(T_1,\ldots,T_K)\right] &= \int \phi(t_1,\ldots,t_K)\, f_\theta(t_1,\ldots,t_K)\,\lambda_1(dt_1)\ldots\lambda_K(dt_K)\\
&= \int \phi(t_1|\,t_2,\ldots,t_K) f_{\theta_1}(t_1|\,t_2,\ldots,t_K) f_\theta(t_2,\ldots,t_K)\lambda_1(dt_1)\ldots\lambda_K(dt_K)\\
&= \int_{t_2,\ldots,t_K}\left(\int \phi(t_1|\,t_2,\ldots,t_K)\, f_{\theta_1}(t_1|\,t_2,\ldots,t_K)\,\lambda_1(dt_1)\right)\\
&\quad f_\theta(t_2,\ldots,t_K)\,\lambda_2(dt_2)\ldots\lambda_K(dt_K)\\
&= \mathbb{E}_\theta\left(\mathbb{E}_{\theta_1}\left[\phi(T_1\,|\,T_2,\ldots,T_K)\,|\,T_2,\ldots,T_K\right]\right) \qquad (5.28)
\end{aligned}$$

■

Faisons maintenant l'hypothèse supplémentaire que pour chaque $(t_2,\ldots,t_K)$ fixé, les densités conditionnelles $t_1 \mapsto f_{\theta_1}(t_1|\,t_2,\ldots,t_K)$ sont à RVM croissantes en t_1 : voir le chapitre 4. Alors, d'après le théorème 4.1, il en résulte que le test conditionnel $t_1 \mapsto \phi(t_1|t_2,\ldots,t_K)$ UPP pour tester $H_0\ [\theta_1 \le 0]$ contre $H_1\ [\theta_1 > 0]$ a une région critique $W(t_2,\ldots,t_K)$ de la forme $t_1 > c(t_2,\ldots,t_K)$ et que pour ce test, la fonction

$$\theta_1 \longmapsto \mathbb{E}_{\theta_1}\left[\phi(T_1|\,t_2,\ldots,t_K)\,|\,T_2 = t_2,\ldots,T_K = t_K\right]$$

est croissante. D'après le lemme, on en déduit que parmi tous les tests ψ de $H_0\ [\theta_1 \le 0]$ contre $H_1\ [\theta_1 > 0]$ dont les tests conditionnels associés sont pour presque tout $(t_2,\ldots,t_K)$ de niveau α, le test UPP est le test ϕ dont les tests conditionnels associés sont de la forme

$$t_1 > c(t_2,\ldots,t_K)\ , \qquad (5.29)$$

les constantes (en t_1) $c(t_2,\ldots,t_K)$ étant choisies pour que

$$\mathbb{E}_0\left(\mathbb{1}_{W(t_2,\ldots,t_K)}\,|\,T_2 = t_2,\ldots,T_K = t_K\right) = \int_{c(t_2,\ldots,t_K)}^{\infty} f_0(t_1|\,t_2,\ldots,t_K)\ \lambda_1(dt_1) \qquad (5.30)$$

soit maximale $\le \alpha$. En particulier, dans le cas où

$$c \longmapsto \int_c^{\infty} f_0(t_1|\,t_2,\ldots,t_K)\ \lambda_1(dt_1)$$

est continue, la quantité (5.30) doit valoir α.

Mais une question se pose alors : soit $\phi = \mathbb{1}_W$ un test non conditionnel pour tester $H_0\ [\theta_1 \le 0]$ contre $H_1\ [\theta_1 > 0]$ au niveau α, les paramètres $\theta_2,\ldots,\theta_K$ n'étant pas spécifiés : est-ce que les tests conditionnels associés $t_1 \mapsto \phi(t_1|\,t_2,\ldots,t_K)$ sont, pour presque tout $(t_2,\ldots,t_K)$, de niveau α ? En général, la réponse est négative. Par conséquent, parmi tous les tests non conditionnels $H_1\ [\theta_1 > 0]$ α-semblables sur le bord avec $\theta_2,\ldots,\theta_K$ non spécifiés, il pourrait en exister un UPP, mais dont les tests conditionnels associés ne seraient pas tous (ni même presque tous) de niveau α. Si tel était le cas, le test conditionnel défini par (5.29) et (5.30) ne conduirait pas nécessairement au test non conditionnel α-semblable sur $(\theta_1 = 0)$ UPP.

Pour obtenir à partir de (5.29)-(5.30) un véritable test α-semblable sur $(\theta_1 = 0)$ UPP dans cette classe, il est donc indispensable d'imposer l'hypothèse suivante : pour tous test $\phi = 1$ I_W non conditionnel α-semblable sur $(\theta_1 = 0)$ et pour tous $\theta_2,\ldots,\theta_K$, les tests conditionnels

associés $t_1 \mapsto \phi(t_1|\, t_2, \ldots, t_K)$ sont de niveau α (pour $\theta_1 \leq 0$ contre $\theta_1 > 0$) pour $\lambda_2 \otimes \ldots \otimes \lambda_K-$presque tout $(t_2, \ldots, t_K)$.

Or, cette propriété est vraie dès que la statistique exhaustive $\underline{T} = (T_1, \ldots, T_K)$ est *complète*, ou *totale* (voir (5.19)-(5.20)). On applique en effet (5.19)–(5.20) à la fonction

$$\underline{g}(\underline{T}) = \mathbb{E}_0(\phi(\underline{T})\,|T_2, \ldots, T_K) \;-\; \alpha \tag{5.31}$$

Puisque, par hypothèse, le test $\phi = \mathbb{1}_W$ est α-semblable sur $\theta_1 = 0$, on a par déconditionnement (voir (5.28)) :

$$\forall \theta_2, \ldots, \theta_K, \quad \mathbb{E}_{(0,\,\theta_2,\ldots,\theta_K)}\left[\underline{g}(\underline{T})\right] = \mathbb{E}_{(0,\,\theta_2,\ldots,\theta_K)}\left[\phi(\underline{T})\right] \;-\; \alpha = 0 \tag{5.32}$$

Par conséquent, pour tout $(\theta_2, \ldots, \theta_K)$, l'ensemble des $(t_2, \ldots, t_K)$ tels que

$$\mathbb{E}_0(\phi(\underline{T}\,|\, T_2 = t_2, \ldots, T_K = t_K)) \neq \alpha$$

est alors de $\lambda_2 \otimes \ldots \otimes \lambda_K$-mesure nulle.

Remarque 5 *Les énoncés ci-dessous ne sont, en toute rigueur, corrects que dans le cas continu, c'est-à-dire lorsque la fonction*

$$c \longmapsto \int_c^{\infty} f_{\theta_{0,\,1}}(t_1|\, t_2, \ldots, t_K)\; \lambda_1(dt_1)$$

est continue, ceci pour presque tout $(t_2, \ldots, t_K)$. Ils doivent être adaptés sinon.

Remarque 6 *Le début du paragraphe 3, consacré à la structure exponentielle, montre que dans ce cadre il est possible de vérifier les hypothèses faites, que nous récapitulons :*

- $\Theta = \mathbb{R}^K$*, mais on pourrait considérer, plus généralement, toute partie suffisamment régulière de* $\mathbb{R}^K$*.*
- *Il existe une statistique* $\underline{T} = (T_1, \ldots, T_K)$ ***exhaustive et totale*** *pour* θ*.*
- *La loi jointe de* $\underline{T}$ *admet une densité* $f_\theta(\underline{t})$ *relativement à une mesure de référence* $\underline{\lambda}(d\underline{t})$ *admettant une décomposition en le produit d'une mesure de référence* $\lambda_1(dt_1)$ *et d'une mesure de référence en* $t_2, \ldots, t_K$*.*
- *La densité conditionnelle (5.23) de* T_1 *sachant* $T_2 = t_2, \ldots, T_K = t_K$ *est bien définie et ne* ***dépend que du paramètre*** θ_1*.*
- *De plus, la densité conditionnelle (5.23)* ***est à RVM*** *(croissante, ici)* ***en*** t_1*.*
- *Enfin, pour obtenir un énoncé correct, on suppose que*

$$c \longmapsto \int_{]c,\,+\infty[} f_{\theta_{0,\,1}}(t_1\,|t_2, \ldots, t_K)\; \lambda_1(dt_1)$$

est continue pour presque tout $(t_2, \ldots, t_K)$, mais c'est là seulement une commodité : comme dans le chapitre 4, les énoncés pourraient être adaptés, au prix de complications supplémentaires, aux autres cas.

Principe de la démonstration du théorème 5.1. Soit ϕ un test non conditionnel pour $H_0\ [\theta_1 \leq \theta_{0,\,1}]$ contre $H_1\ [\theta_1 > \theta_{0,\,1}]$, avec $\theta_2, \ldots, \theta_K$ non spécifiés, qui sont UPP parmi les

tests α-semblables sur le bord d'équation $\theta_1 = \theta_{0,1}$. D'après l'hypothèse de complétion, les tests conditionnels $t_1 \mapsto \phi(t_1 | t_2, \ldots, t_K)$ associés vérifient

$$\mathbb{E}_{\theta_{0,1}} [\phi(T_1 | t_2, \ldots, t_K) | T_2 = t_2, \ldots, T_K = t_K] = \alpha \tag{5.33}$$

à un ensemble de $(t_2, \ldots, t_K)$ de mesure nulle près. D'après le théorème 4.1, les tests UPP sont de la forme (5.29) et (5.30). D'après le lemme 5.1, cela conduit à la forme de ϕ. De plus, puisque la fonction $\mathbb{E}_{\theta_1} [\phi | T_2 = t_2, \ldots, T_K = t_K]$ est croissante et en utilisant à nouveau le lemme 5.1, le test ϕ est de niveau α et il est sans biais, et ceci pour tout $(\theta_2, \ldots, \theta_K)$. ■

Théorème 5.2 *On a un résultat analogue pour les tests bilatéraux.*

5.3 Application à des problèmes de comparaison d'échantillons

Exemple 9 *Comparaison de deux n-échantillons issus de lois de Poisson. Les variables i.i.d.* $X_1, X_2, \ldots$ *sont distribuées selon* Poisson(λ_1)*, et* $Y_1, Y_2, \ldots$ *selon Poisson*(λ_2)*. Les variables* X_i *et* Y_j *sont mutuellement indépendantes. On voudrait savoir si ces deux échantillons* $X_1, \ldots, X_n$ *et* $Y_1, \ldots, Y_n$ *sont issus de la même loi. Autrement dit, on souhaite tester*

$$H_0 [\lambda_1 = \lambda_2 = \lambda], \quad \lambda \text{ non spécifié},$$

contre

$$H_1 [\lambda_1 > \lambda_2] \qquad (\text{respectivement}, \lambda_1 < \lambda_2 \text{ ou } \lambda_1 \neq \lambda_2)$$

Dans ce cas, les deux statistiques $T_1 = \sum_{i=1}^n X_i$ *et* $T_2 = \sum_{i=1}^n Y_i$ *sont indépendantes et exhaustives pour* λ_1 *et* λ_2*.*

La loi du couple (T_1, T_2) *s'écrit :*

$$\mathbb{P}[T_1 = k_1, T_2 = k_2] = \frac{1}{k_1! k_2!} e^{-n(\lambda_1 + \lambda_2)} \exp[(\ln \lambda_1) k_1 + (\ln \lambda_2) k_2] \tag{5.34}$$

Si on choisit pour mesure de référence la mesure de comptage ν *sur* $\mathbb{N}$*, on obtient une famille exponentielle de paramètres naturels* $\theta_1 = \ln \lambda_1$ *et* $\theta_2 = \ln \lambda_2$*.*

Or, on veut tester l'égalité de θ_1 *et* θ_2*. Cela suggère d'effectuer un changement de paramétrage, afin de faire apparaître le paramètre* $\tau_1 = \theta_1 - \theta_2$*. Pour cela, on écrit :*

$$\exp[\theta_1 T_1 + \theta_2 T_2] = \exp[\tau_1 T_1 + \theta_2 (T_1 + T_2)] \tag{5.35}$$

On considère donc les nouveaux paramètres $\tau_1 = \theta_1 - \theta_2$ *et* $\tau_2 = \theta_2$*, associés aux statistiques exhaustives*

$$T = T_1 \quad \text{et} \quad U = T_1 + T_2 \tag{5.36}$$

On est alors ramené à $[\tau_1 = 0]$*,* τ_2 *non spécifié, contre une des alternatives* $[\tau_1 > 0]$*,* $[\tau_1 < 0]$ *ou* $[\tau_1 \neq 0]$*. Les tests à structure de Neyman associés consistent à tester* $[\tau_1 = 0]$ *conditionnellement à* $U = u$*. Or,* $U \sim$ *Poisson*$(n(\lambda_1 + \lambda_2))$ *et*

$$\begin{aligned} \mathbb{P}[T = k \mid U = u] &= \frac{\mathbb{P}[T_1 = k, T_2 = u - k]}{\mathbb{P}[U = u]}, \qquad u \geq k \geq 0 \\ &= \frac{u!}{k!(u-k)!} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{u-k} \end{aligned} \tag{5.37}$$

On reconnaît une loi binomiale $\mathrm{Bin}\,(u,\, p = \lambda_1/(\lambda_1+\lambda_2))$.

Sous H_0, $p = \lambda_1/(\lambda_1+\lambda_2) = 1/2$, *tandis que* $p > 1/2$ *sous* H_1 $[\lambda_1 > \lambda_2]$ *par exemple. Il existe donc un test UPP semblable, défini conditionnellement par*

$$\phi_n = \begin{cases} 1 & \text{si} \quad T_1 > c(u) \\ 0 & \text{si} \quad T_1 < c(u) \end{cases} \tag{5.38}$$

Pour déterminer $c(u)$, *on écrit que*

$$\mathbb{P}\,[T_1 > c(u) \mid U = u] \approx \alpha, \qquad 0 < \alpha < 1, \tag{5.39}$$

soit

$$2^{-u} \sum_{k=c(u)+1}^{u} C_u^k \approx \alpha \tag{5.40}$$

(en fait, $\leq \alpha$ *et maximal parmi les termes de cette forme* $\leq \alpha$*).*

Pratiquement, ce test revient à situer k_1 par rapport à un seuil $c(k_1+k_2)$, calculé d'après (5.38). Bien entendu, ce test doit respecter la symétrie entre k_1 et k_2.

Exemple 10 *On considère deux échantillons indépendants, l'un de taille* n_1, $X_1, \ldots, X_{n_1}$ *i.i.d.* $\sim \mathcal{N}(\mu_1,\, \sigma_1^2)$, *l'autre de taille* n_2, $Y_1, \ldots, Y_{n_2}$ *i.i.d.* $\sim \mathcal{N}(\mu_2, \sigma_2^2)$. *On souhaite tester*

$$H_0\,[\mu_1 = \mu_2 = \mu]\,, \quad \mu \text{ non spécifié},$$

contre

$$H_1\,[\mu_1 \neq \mu_2]\,,$$

en supposant que $\sigma_1^2 = \sigma_2^2 = \sigma^2$ *(non spécifié). On introduit les statistiques exhaustives indépendantes* $\overline{X}$ *et* $\overline{Y}$. *Il est naturel de former* $\overline{X} - \overline{Y}$, *qui suit une loi* $\mathcal{N}(\mu_1 - \mu_2,\, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$ *pour tester* $[\mu_1 = \mu_2]$. *On est alors ramené au problème déjà traité de tester la nullité de la moyenne d'une loi normale, sa variance étant inconnue. Ici, on estime* σ^2 *à partir de*

$$S^2 = (n_1 + n_2 - 2)^{-1} \left[\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2 \right] \tag{5.41}$$

avec $(n_1 + n_2 - 2)\,/\,\sigma^2 S^2 \sim \chi^2_{n_1+n_2-2}$ *sous* H_0. *Par indépendance, le rapport*

$$\frac{(\overline{X} - \overline{Y})/\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{S/\sigma} \tag{5.42}$$

soit une loi de Student à $(n_1 + n_2 - 2)$ *degrés de liberté. On rejette donc* H_0 *si*

$$|\overline{x} - \overline{y}| > s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, c_{\alpha/2}, \tag{5.43}$$

où

$$\int_{c_{\alpha/2}}^{\infty} f_{n_1+n_2-2}(t)\, dt = \frac{\alpha}{2}\,, \tag{5.44}$$

$f_{n_1+n_2-2}$ *étant la densité de* $t_{n_1+n_2-2}$.

Exercice 5.1 *Avec les mêmes notations mais en supposant σ^2 connue, comment tester $\mu_1 = \mu_2 = 0$? Comment faire si σ^2 est inconnue ? Voir, par exemple, Ferguson (1967), Erkel-Rousse (2001, pages 132–137) et Monfort (1982).*

Exercice 5.2 *Soit $\mathcal{N}_K(\mu, \Sigma)$ une loi normale sur $\mathbb{R}^K$ ($K \geq 1$). Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. $\mathcal{N}_K(\mu, \Sigma)$. Comment tester $\mu_1 = 0$ contre $\mu_1 \neq 0$ (si $\mu = (\mu_1, \ldots, \mu_K)$) si Σ est connue, et $\mu_2, \ldots, \mu_K$ non spécifiés ? Voir Ferguson (1967), Erkel-Rousse (2001, pages 132–137) et Monfort (1982).*

Conclusion

- Notion de "test conditionnel".
- Garantie d'optimalité : tests uniformément plus puissants parmi les tests α-semblables) ; possibilité de construire des tests à distance finie.
- Retenir le rôle des structures exponentielles et/ou des statistiques exhaustives totales.
- Problèmes de comparaison d'échantillons ; homogénéité d'échantillons ; point de départ de l'analyse de la variance.

Eléments bibliographiques pour ce chapitre

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Erkel-Rousse, H. (2001) *Introduction à l'économétrie du modèle linéaire*. Polycopié de l'Ensae. Insee–Ensae : Montrouge.

Ferguson, T. S. (1967) *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press : New York and London.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Malinvaud, E. (1978) *Méthodes statistiques de l'économétrie*. Dunod : Paris, 3e édition.

Monfort, A. (1982) *Cours de Statistique mathématique*. Collection "Economie et statistiques avancées", Economica : Paris.

Tassi, P. (1985) *Méthodes statistiques*. Collection "Economie et statistiques avancées", Economica : Paris.

Ulmo, J. & Bernier, J. (1973) *Eléments de décision statistique*. Presses Universitaires de France : Paris.

6

Tests et régression

6.1 Régression linéaire multiple : un cas particulier

On considère le modèle de régression linéaire multiple particulier suivant :

Pour $n > 1$ fixé, les points $x_i = x_{i,n}$ $(1 \leq i \leq n)$ sont des "points de design" fixés, déterminés par l'expérimentateur. On les supposera réels ici.

Après transformation linéaire, on suppose que les observations $Y_i \in \mathbb{R}$, $1 \leq i \leq n$, sont reliées aux points de design $x_i = x_{i,n}$, $1 \leq i \leq n$, par la relation

$$Y_i = \theta_0 + \sum_{j=1}^{J-1} \theta_j g_j(x_i) + \sigma\varepsilon_i, i = 1, \ldots, n \tag{6.1}$$

où :

- $\theta_0, \theta_1, \ldots, \theta_{J-1}$ sont des paramètres réels inconnus ;
- les fonctions $g_0 \equiv 1$, g_j vérifient les relations d'orthonormalité relativement au design suivantes :

$$\forall j, k, \ \langle g_j, g_k \rangle_n = \frac{1}{n} \sum_{i=1}^{n} g_j(x_i) g_k(x_i) = \delta_{jk} \tag{6.2}$$

- σ^2 est la variance inconnue du bruit ;
- ε_1, $\varepsilon_2, \ldots$ sont des v.a. i.i.d. $\mathcal{N}(0, 1)$ représentant le bruit.

On souhaite d'abord, bien sûr, estimer θ_j $(0 \leq j \leq J-1)$ ainsi que σ^2.

Ensuite, on souhaite tester H_0 $[\theta_j = 0]$ contre H_1 $[\theta_j \neq 0]$ pour une ou plusieurs valeurs de j, afin de savoir si la (les) variable (s) a (ont) un rôle explicatif ou non.

6.2 Estimation par moindres carrés de θ

Pour θ, on cherche $\widehat{\theta}_n$ minimisant

$$\left\| \underline{Y} - \sum_{j=0}^{J-1} \theta_j \underline{g}_j \right\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{J-1} \theta_j g_j(x_i) \right)^2, \tag{6.3}$$

où on a noté $\underline{Y} = (Y_1, \ldots, Y_n)^T$, et confondu, par abus de notation, $\underline{g}_j$ avec

$$(g_j(x_1), \ldots, g_j(x_n))$$

Comme on a choisi (quitte à avoir effectué au préalable une transformation linéaire sur les fonctions de régression initiales) les g_j de manière à ce qu'elles constituent un système orthonormé relativement au design, on a :

$$\left\| \underline{Y} - \sum_{j=0}^{J-1} \theta_j \underline{g}_j \right\|_n^2 = \|\underline{Y}\|_n^2 - 2 \sum_{j=0}^{J-1} \theta_j \left\langle \underline{Y}, \underline{g}_j \right\rangle_n + \sum_{j=0}^{J-1} \theta_j^2 \tag{6.4}$$

Minimiser cette expression, c'est chercher la combinaison linéaire

$$\underline{\widehat{Y}} = \sum_{j=0}^{J-1} \widehat{\theta}_j \underline{g}_j \in G = \text{Vect } (\underline{g}_0, \ldots, \underline{g}_{J-1}),$$

qui représente la projection orthogonale du vecteur $\underline{Y} \in \mathbb{R}^n$ sur le sev G, supposé de dimension $< n$, donc $J < n$. On trouve :

$$\forall j = 0, \ldots, J-1, \quad \widehat{\theta}_j = \left\langle \underline{Y}, \underline{g}_j \right\rangle_n = \frac{1}{n} \sum_{i=1}^{n} Y_i g_j(x) \tag{6.5}$$

En particulier, $\widehat{\theta}_0 = \overline{Y}$. Revenant à (6.5),

$$\begin{aligned} \widehat{\theta}_j &= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j'=0}^{J-1} \theta_j g_{j'}(x_i) + \sigma \varepsilon_i \right) g_j(x_i) \\ &= \sum_{j'=0}^{J-1} \theta_{j'} \left\langle \underline{g}_j, \underline{g}_{j'} \right\rangle_n + \sigma \left\langle \underline{\varepsilon}, \underline{g}_j \right\rangle_n \\ &= \theta_j + \frac{\sigma \xi_j}{\sqrt{n}}, \end{aligned} \tag{6.6}$$

avec $\underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ et $\xi_j \sim \mathcal{N}(0,1)$ i.i.d. pour $j = 0, \ldots, J$ car $\left\langle \underline{g}_j, \underline{g}_k \right\rangle = \delta_{jk}$.

Par conséquent, $\widehat{\theta}_j$ est un estimateur sans biais de θ_j, et

$$\sqrt{n}(\widehat{\theta}_j - \theta_j) = \frac{\sigma \xi_j}{\sqrt{n}} \sim \mathcal{N}\left(0, \frac{\sigma}{\sqrt{n}}\right) \tag{6.7}$$

6.3 Estimation sans biais de σ^2

Puisque $\underline{\widehat{Y}}$ est la projection orthogonale de $\underline{Y}$ sur $G = \text{Vect}\ (\underline{g}_0, \ldots, \underline{g}_j)$ il est naturel de s'intéresser au vecteur des résidus

$$\underline{e} = \underline{Y} - \underline{\widehat{Y}} \tag{6.8}$$

On estime σ^2 au moyen de la somme des carrés résiduels $\sum_{i=1}^n e_i^2$, convenablement normalisée.

Théorème 6.1 *Pour $J < n$:*

1. $$\widehat{\sigma}^2_{SB} = \frac{1}{n-J}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n-J}\,\|\underline{e}\|_n^2 \tag{6.9}$$

   est un estimateur sans biais de σ^2.

2. *Les résidus (ou écarts résiduels) $e_i = Y_i - \widehat{Y}_i$ sont non corrélés avec les estimateurs $\widehat{\theta}_j$ des θ_j, donc avec toute fonction linéaire des $\widehat{\theta}_j$.*

Remarque 7 *La démonstration de ce théorème (voir ci-dessous) montre que ses résultats restent vrais sous les seules hypothèses que $\mathbb{E}(\underline{\varepsilon}) = 0$ et $\text{Var}(\underline{\varepsilon}) = \sigma^2 I_n$.*

Esquisse de la démonstration.

1. Remarquons d'abord que, par Pythagore,

   $$\|\underline{Y}\|_n^2 = \left\|\underline{\widehat{Y}}\right\|_n^2 + \|\underline{e}\|_n^2\,, \tag{6.10}$$

   avec

   $$\left\|\underline{\widehat{Y}}\right\|_n^2 = \sum_{j=0}^{J-1}\widehat{\theta}_j^2 = \sum_{j=0}^{J-1}\theta_j^2 + \frac{\sigma^2}{n}\sum_{j=0}^{J-1}\xi_j^2 + \frac{2\sigma}{\sqrt{n}}\sum_{j=0}^{J-1}\theta_j\xi_j, \tag{6.11}$$

   tandis que

   $$\begin{aligned}\underline{Y}_n^2 &= \sum_{j=0}^{J-1}\theta_j^2 + \sigma^2\,\|\varepsilon\|_n^2 + 2\sigma\sum_{j=0}^{J-1}\theta_j\left\langle \underline{g}_j, \underline{\varepsilon}\right\rangle_n \\ &= \sum_{j=0}^{J-1}\theta_j^2 + \sigma^2\,\|\varepsilon\|_n^2 + \frac{2\sigma}{\sqrt{n}}\sum_{j=0}^{J-1}\theta_j\xi_j \end{aligned} \tag{6.12}$$

   Ainsi,

   $$\|\underline{e}\|_n^2 = \sigma^2\left[\|\varepsilon\|_n^2 - \frac{1}{n}\sum_{j=0}^{J-1}\xi_j^2\right] \tag{6.13}$$

   Complétons le système orthonormé $(\underline{g}_0, \ldots, \underline{g}_{j-1})$ en une base orthonormée $(\underline{g}_0, \ldots, \underline{g}_{n-1})$ de $\mathbb{R}^n$, et posons

   $$\xi_j = \left\langle \underline{g}_j, \varepsilon\right\rangle_n \qquad \text{pour} \qquad j = 0, \ldots, n-1 \tag{6.14}$$

Les ξ_j obtenus sont encore tous $\mathcal{N}(0,1)$ i.i.d. Par suite,

$$\|\underline{e}\|_n^2 = \sigma^2 \frac{1}{n} \sum_{j=J}^{n-1} \xi_j^2 \tag{6.15}$$

$\Big($puisque $\sum_i \varepsilon_i^2 = \sum_j \xi_j^2\Big)$, et

$$\mathbb{E}(\|\underline{e}\|_n^2) = \frac{\sigma^2}{n} \sum_{j=J}^{n-1} \mathbb{E}(\xi_j^2) = \frac{(n-J)\,\sigma^2}{n} \tag{6.16}$$

Remarquons que, par la loi forte des grands nombres, on a donc

$$\frac{1}{n-J} \sum_{i=1}^{n} e_i^2 = \sigma^2 \frac{\sum_{j=J}^{n-1} \xi_j^2}{n-J} \xrightarrow{\text{p.s.}} \sigma^2 \text{ quand } n \to \infty \tag{6.17}$$

2. On a

$$e_i = Y_i - \widehat{Y}_i = \sigma\varepsilon_i - \frac{\sigma}{\sqrt{n}} \sum_{j=0}^{J-1} \xi_j g_j(x_i) \tag{6.18}$$

et, par conséquent,

$$\begin{aligned} \underline{e} &= \sigma \left[\underline{\varepsilon} - \frac{1}{\sqrt{n}} \sum_{j=0}^{J-1} \xi_j \underline{g}_j \right] \\ &= \sigma \left[\sum_{j=0}^{n-1} \Big\langle \underline{g}_j, \underline{\varepsilon} \Big\rangle \underline{g}_j - \sum_{j=0}^{J-1} \Big\langle \underline{g}_j, \underline{\varepsilon} \Big\rangle \underline{g}_j \right] \\ &= \sigma \sum_{j=J}^{n-1} \Big\langle \underline{g}_j, \underline{\varepsilon} \Big\rangle \underline{g}_j \end{aligned} \tag{6.19}$$

qui ne dépend pas de $\xi_0, \xi_1, \ldots, \xi_J$. ■

Corollaire 3 *Sous les hypothèses faites, la v.a.*

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = \frac{(n-J)\widehat{\sigma}_{SB}^2}{\sigma^2} = \sum_{j=J}^{n-1} \xi_j^2 \tag{6.20}$$

suit une loi χ^2_{n-J}.

On peut montrer, en outre, que sous les hypothèses faites – (x_i points de design, ε_i i.i.d. $\mathcal{N}(0,1)$, on a :

Théorème 6.2 *Les statistiques exhaustives pour* (θ, σ^2) *:*

$$T_j(\underline{Y}) = \left\langle \underline{Y}, \underline{g}_j \right\rangle_n , \quad j = 0, \ldots, J-1 \tag{6.21}$$

et

$$S^2(\underline{Y}) = \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \tag{6.22}$$

forment une famille exponentielle d'ordre $J+1$*. De plus les estimateurs du MV de* θ_j *et de* σ^2 *sont :*

$$\widehat{\theta}_j = \left\langle \underline{Y}, \underline{g}_j \right\rangle_n = T_j(\underline{Y}), \quad j = 0, \ldots, J-1, \tag{6.23}$$

et

$$\widehat{\sigma}^2_{MV} = \frac{1}{n} \|\underline{e}\|_n^2 = S^2(Y) - \sum_{j=0}^{J-1} T_j^2(\underline{Y}) \tag{6.24}$$

Enfin, $\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$ *est indépendante de* $\widehat{\theta}$*.*

6.4 Tester l'absence d'effet d'une variable g_k

Cela revient à tester H_0 $[\theta_k = 0]$ contre H_1 $[\theta_k \neq 0]$, les autres θ_j et σ^2 n'étant pas spécifiés. On applique les méthodes du chapitre 4. Par indépendance des $T_j(\underline{Y}) = \widehat{\theta}_j$ entre eux et avec $\sum_{i=1}^{n} e_i^2$ (voir théorème 6.1) et en conditionnant $T_k(\underline{Y})$ par $T_j(\underline{Y})$, $j \neq k$, et $\sum_{i=1}^{n} e_i^2$, on se ramène à un test conditionnel de la forme : acceptation si

$$c_1 \leq \frac{\sqrt{n}\widehat{\theta}_k}{\widehat{\sigma}} \leq c_2 \,, \tag{6.25}$$

où

$$\widehat{\sigma}^2 = \widehat{\sigma}^2_{SB} = (n-J)^{-1} \sum_{i=1}^{n} e_i^2$$

D'après le corollaire 3, $\sigma^{-2} \sum_{i=1}^{n} e_i^2$ suit une loi χ^2_{n-J} et est indépendante de $\widehat{\theta}_k$. D'autre part, on sait que $n^{1/2}\ \sigma^{-1}\ \widehat{\theta}_k \sim \mathcal{N}(0,\, 1)$ sous H_0. Par suite, le rapport

$$\frac{\sqrt{n}\widehat{\theta}_k/\sigma}{(1/\sigma)\sqrt{\sum_{i=1}^{n} e_i^2/(n-J)}} = \frac{\sqrt{n}\widehat{\theta}_k}{\widehat{\sigma}_{SB}} \tag{6.26}$$

suit sous H_0 une loi de Student t_{n-J}. Pour obtenir un test UPP sans biais, il faut donc prendre $c_1 = -c_2 = c_{\alpha/2}$, où

$$\int_{c_{\alpha/2}}^{\infty} f_{n-J}(t)\, dt = \frac{\alpha}{2} \,, \tag{6.27}$$

f_{n-J} désignant la densité de la loi t_{n-J}.

6.5 Tester l'absence d'effet d'un groupe de variables

Le principe est le même. Soit g_λ, $\lambda \in \Lambda \subset \{0, \ldots, J-1\}$, le groupe de variables dont on souhaite tester l'absence d'effet, ce qui revient à tester H_0 $[\theta_\lambda = 0,\ \lambda \in \Lambda]$, les autres θ_j et σ^2 n'étant pas spécifiés.

On remarque que si l'on veut tester H_0 $[\mu_1 = \mu_2 = 0]$ pour $\mathcal{N}(\mu_1,\, 1) \otimes \mathcal{N}(\mu_2,\, 1)$, une statistique exhaustive est la somme des carrés $\overline{X}^2 + \overline{Y}^2$. On est donc amené à former la statistique de test

$$\frac{\left(n \sum_{\lambda \in \Lambda} \widehat{\theta}_\lambda^2\right) \Big/ \sigma^2}{\widehat{\sigma}^2 / \sigma^2}\,, \tag{6.28}$$

où $\widehat{\sigma}^2 = \widehat{\sigma}^2_{SB}$. Le numérateur suit sous H_0 une loi χ^2_L, avec $L = \mathrm{card}\,(\Lambda)$. Le dénominateur, indépendant du numérateur, suit une loi χ^2_{n-J}. Le rapport

$$F = \frac{n \sum_{\lambda \in \Lambda} \widehat{\theta}_\lambda^2}{L \widehat{\sigma}_{SB}} \tag{6.29}$$

suit sous H_0 une loi de Fisher $\mathcal{F}_{L,\, n-J}$. Sinon, F suit une loi de Fisher $\mathcal{F}_{L,\, n-J}(\gamma^2)$ décentrée de paramètre de décentrage $\gamma^2 = \sigma^{-2} \sum_{\lambda \in \Lambda} \theta_\lambda^2$. On peut montrer que les lois $\mathcal{F}_{L,\, n-J}(\gamma^2)$ sont à RVM en γ^2, donc les tests de région critique $F > c$ sont UPP parmi les tests invariants : c'est la méthode du test de Fisher. Voir Ferguson (1967), Erkel-Rousse (2001, pages 132–137) et Monfort (1982).

Eléments bibliographiques pour ce chapitre

Alcalá, J. T., Cristóbal, J. A. & González Manteiga, W. (1999) Goodness-of-fit test for linear models based on local polynomials. *Statist. Probab. Lett.* **42**, 39–46.

Azzalini, A., Bowman, A. W. & Härdle, W. (1989) On the use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Dagnelie P. (1975) *Théorie et méthodes statistiques*, Tome 2. Vander-Oyez : Bruxelles, 2e édition.

Davison, A. C. & Tsai, C. L. (1992) Regression model diagnostics. *Internat. Statist. Rev.* **60**, 337–353.

Dette, H. (2000) On a nonparametric test for linear relationships. *Statist. Probab. Lett.* **46**, 307–316.

Dette, H. & Munk, A. (1998) Validation of linear regression models. *Ann. Statist.* **26**, 778–800.

Erkel-Rousse, H. (2001) *Introduction à l'économétrie du modèle linéaire*. Polycopié de l'Ensae. Insee–Ensae : Montrouge[1].

[1]Rappelons que cet ouvrage contient une bibliographie très complète sur l'histoire de la Statistique et de l'Econométrie (pages 53–54 et 56–66), et sur le modèle linéaire (pages 54–56 et fins de chapitres).

Eubank, R.L. & Spiegelman, C.H. (1990) Testing the goodness-of-fit of a linear model via nonparametric regression techniques. *J. Am. Statist. Assoc.* **85**, 387–392.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Härdle, W. & Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926–1947.

Malinvaud, E. (1978) *Méthodes statistiques de l'économétrie*. Dunod : Paris, 3e édition.

Monfort, A. (1982) *Cours de Statistique mathématique*. Collection "Economie et statistiques avancées", Economica : Paris.

Stute, W. (1997) Nonparametric model checks for regression. *Ann. Statist.* **25**, 613–641.

Stute, W. & González Manteiga, W. (1996) NN goodness-of-fit tests for linear models. *J. Statist. Plann. Inference* **53**, 75–92.

Stute, W., González Manteiga, W. & Presedo Quindimil, M. (1998) Checks for regression. *J. Am. Statist. Assoc.* **93**, 141–149.

Stute, W., Thies, S. & Zhu, L.X. (1998) Model checks for regression : an innovation process approach. *Ann. Statist.* **26**, 1916–1934.

Tassi, P. (1985) *Méthodes statistiques*. Collection "Economie et statistiques avancées", Economica : Paris.

Ulmo, J. & Bernier, J. (1973) *Eléments de décision statistique*. Presses Universitaires de France : Paris.

7

Tests utilisant les rangs et les signes

Dans ce chapitre, nous étudierons quelques tests statistiques basés sur les rangs, c'est-à-dire uniquement sur les rangs des observations et pas sur leurs valeurs précises. Nous utilisons simplement, pour chacune des observations, le fait qu'il s'agit de la k-ème observation et qu'il y a n observations au total. Ainsi, les tests de rang peuvent être utilisés dès lors qu'il y existe un ordre sur les observations : l'ensemble des observations est muni d'une relation d'ordre, donc d'une d'une échelle ordinale. C'est évidemment le cas d'observations à valeurs réelles. Ces tests ne supposent a priori aucun modèle précis pour représenter le mécanisme probabiliste qui est censé idéalement avoir généré les observations. On peut utiliser ces tests lorsqu'on n'a que peu d'idées sur la façon dont les observations sont distribuées.

Simples et intuitifs, ces tests s'avèrent avoir de bonnes propriétés, ce qui les rend très intéressants. La première et la plus importante de ces propriétés est l'invariance des statistiques de test (et donc de la conclusion des tests) par une transformation monotone croissante des observations. En effet, même si la valeur des observations change sous l'action d'une telle transformation, leur ordre reste le même.

Il faut signaler aussi que ces tests sont peu sensibles aux valeurs aberrantes et aux erreurs de modèle. On dit qu'ils sont *robustes*. Imaginons par exemple que des mesures relatives à un certain phénomène, rangées dans l'ordre croissant, soient : 25, 30, 43, 52, 64, 100, 128, 134, 157, 171, 15 000. La moyenne empirique, statistique estimant la moyenne d'une loi, vaut $\overline{x} = 1445,818$ pour cette série d'observations, tandis que la médiane, statistique de rang puisqu'elle est définie comme la quantité med telle qu'il y ait autant d'observations de valeur inférieure à med que d'observations de valeur supérieure à med, vaut $\text{med} = 100$. La valeur aberrante 15 000 (sans doute due à une erreur de transcription des résultats des mesures) a conduit à une valeur démesuré de la moyenne, mais pas de la médiane. D'autre part, considérons deux échantillons indépendants, l'un provenant de variables aléatoires X_i, l'autre de variables aléatoires Y_i, et intéressons-nous au test de l'hypothèse nulle H_0 [X a la même loi que Y] contre l'alternative H_1 [la loi de Y est "supérieure à" celle de X][1]. Le test des

[1] Cette notion sera définie dans ce chapitre.

rangs approprié est ici le test de Wilcoxon pour des variables non couplées[2]. Maintenant, si nous faisons l'hypothèse supplémentaire que les deux échantillons proviennent de lois normales de moyennes respectives m_1 et m_2 et de même variance approximativement, il est facile de montrer que l'ordre sur les moyennes induit l'ordre stochastique[3]. Aussi, le problème revient-il à tester H_0 $[m_1 = m_2]$ contre H_1 $[m_1 < m_2]$. Il est bien connu (voir le Chapitre 4) que le test le plus puissant dans ce cas est le test de Student. Cependant, si les lois des X_i et des Y_i ne sont pas exactement des lois normales (la normalité ne constitue qu'une approximation souvent très grossière destinée à construire des modèles *commodes*), alors le test des rangs est le seul que l'on puisse utiliser de manière appropriée.

Enfin, les tests basés sur les rangs sont particulièrement recommandés pour les petits échantillons.

Les tests de rang ont été étudiés dès 1904 dans le cadre des problèmes de corrélation, avec les travaux de Spearman, poursuivis en 1938 par par ceux de Kendall. Les tests de rang ont été ensuite étudiés par Wilcoxon autour de 1945 : Wilcoxon cherchait à comparer des échantillons différents au sens de l'ordre stochastique. Puis, les travaux de Huber au cours des années soixante et soixante-dix leur ont donné un nouvel essor en formalisant et étudiant la notion de robustesse. Ils font encore aujourd'hui l'objet de nombreux articles.

7.1 Rappels sur les statistiques d'ordre

Soit un n-échantillon $X_1, \ldots, X_n$ d'une loi de fonction de répartition continue F sur $\mathbb{R}$. La statistique d'ordre est la variable aléatoire S qui fait correspondre à cet échantillon, l'échantillon ordonné dans le sens croissant $X_{1,n} \leq \ldots \leq X_{n,n}$:

$$S(X_1, \ldots, X_n) = (X_{1,n}, \ldots, X_{n,n})$$

Rappel : Si F est continue et si on pose $U_1 = F(X_1), \ldots, U_n = F(X_n)$, alors les U_i sont des variables aléatoires indépendantes de loi uniforme sur [0, 1]. D'autre part, on peut simuler selon F à partir de la fonction réciproque généralisée $F^{\leftarrow}$, en formant $F^{\leftarrow}(U_i)$.

Interprétation des $X_{i,n}$:

- Les variables aléatoires extrêmes $X_{1,n}$ et $X_{n,n}$ interviennent notamment dans les problèmes de sécurité et fiabilité.
- Si n est pair $X_{n/2,n}$ et, si n est impair, $\left[X_{[n/2],n},\ X_{[n/2]+1,n}\right]$ et $(X_{1,n} + X_{n,n})/2$ sont des mesures de tendance centrale.

Le rappels ci-dessus permettent, connaissant la loi de la statistique d'ordre pour la loi uniforme sur [0, 1], de déterminer la loi de la statistique d'ordre dans le cas général, lorsque la loi des X_i admet une densité. Ainsi, nous obtenons :

[2] *Idem.*
[3] *Idem.*

– La densité de la loi de $(X_{1,n}, \ldots, X_{n,n})$ est

$$\tilde{f}(y_1, \ldots, y_n) = n\,! \prod_{i=1}^{n} f(y_i)\ \mathbb{1}_{y_1 \leq \ldots \leq y_n},$$

où f est la densité de la loi des X_i (rappelons que f est la dérivée de F).

– La densité de la loi de $X_{r,n}$ est

$$f_r(y_r) = \frac{n!}{(r-1)!\,(n-r)!} F^{r-1}(y_r)\,(1 - F(y_r))^{n-r} f(y_r)$$

7.2 Runs

Soit un échantillon de variables aléatoires à deux caractères (les variables aléatoires peuvent prendre deux valeurs). On représente par un mot la réalisation de cet échantillon de taille n. Par exemple,

$$X_1, \ldots, X_n = AABABBB \ldots ABA$$

Un run est une suite de mêmes symboles intervenant dans le mot et de longueur maximale parmi les suites ayant cette propriété. Par exemple :

$$(AA)\,(BBB)\,(A)\,(B)\,(A)$$

La longueur des runs et le nombre des runs sont des statistiques exploitables, par exemple pour tester le caractère aléatoire d'un échantillon donné.

7.3 Comparaison de variables aléatoires

On peut avoir à comparer les lois de deux variables aléatoires X et Y, par exemple les tensions de rupture de deux fils de qualités différentes, les actions de deux somnifères, etc. Lorsque les variables sont définies sur la même catégorie d'épreuves, elle sont dites *couplées*.

Exemple : *L'action de deux somnifères qu'on fait agir tour à tour sur le même individu. Par contre, on ne peut pas coupler les tensions de rupture de deux fils.*

— *Cas de variables non couplées* (exemple : tensions de rupture de deux fils). On peut chercher à comparer $\mathbb{E}(X)$ et $\mathbb{E}(Y)$, et déclarer que Y est meilleur que X si $\mathbb{E}(Y) > \mathbb{E}(X)$. Cependant, ce point de vue est un peu simpliste. En effet, cela intéresse davantage un utilisateur de savoir que dans certains cas, le fil X peut être extrêmement résistant, beaucoup plus que le fil Y. On dira que Y est **stochastiquement supérieure ou égale à** X si

$$\forall x \in \mathbb{R}, \quad \mathbb{P}(Y \leq x) \leq \mathbb{P}(X \leq x), \ \text{ soit } \ F_Y(x) \leq F_X(x)\ ;$$

que Y est **stochastiquement supérieure** à X si

$$\forall x \in \mathbb{R}, \quad \mathbb{P}(Y \leq x) < \mathbb{P}(X \leq x), \ \text{ soit } \ F_Y(x) < F_X(x)\ ;$$

que Y est **stochastiquement égale** à X si

$$\forall x \in \mathbb{R}, \quad \mathbb{P}(Y \leq x) = \mathbb{P}(X \leq x), \text{ soit } F_Y(x) = F_X(x)$$

— *Cas des variables couplées.* Si X et Y sont définies sur le même espace d'épreuves, on peut définir $Z = Y - X$. On est alors amené à considérer le signe de la médiane med de Z. On décide que Y est plus grande que X lorsque med > 0. Or,

$$\begin{aligned} \text{med} > 0 &\Leftrightarrow \mathbb{P}(Z > 0) > 1/2 \\ &\Rightarrow \mathbb{P}(Y > X) > \mathbb{P}(Y \leq X) \\ &\Rightarrow \mathbb{P}(Z \leq 0) > \mathbb{P}(-Z \leq 0) \end{aligned}$$

Il semble donc raisonnable de considérer que Y est supérieure à X si $Z = Y - X$ est stochastiquement supérieure à $-Z = X - Y$.

Remarque 8 *Lorsqu'on a le choix, il vaut mieux utiliser des variables couplées. En effet, la variance de $Z = Y - X$ est souvent inférieure aux variances de X et de Y.*

7.4 Tests des longueurs

On considère des variables aléatoires X et Y supposées continues, et on teste

$$\begin{aligned} H_0 &: X \text{ et } Y \text{ ont la même loi} \\ &\quad \text{contre} \\ H_1 &: X \text{ et } Y \text{ n'ont pas la même loi} \end{aligned}$$

Soient un m-échantillon $(X_1, \ldots, X_m)$ de X et un n-échantillon $(Y_1, \ldots, Y_n)$ de Y. On choisit $m \leq n$, quitte à échanger si nécessaire les échantillons. On ordonne les observations des deux échantillons regroupés, et on note X et Y selon que l'observation provient de l'échantillon X ou Y. On obtient des mots de la forme

$$XXYXYYYXY$$

L'ensemble des mots possibles a C_{m+n}^n éléments. L'hypothèse H_0 signifie que les $m + n$ variables aléatoires indépendantes $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ suivent la même loi (au moins approximativement). Sous H_0, la loi sur l'ensemble des mots possibles est uniforme. Au contraire, sous H_1 il existe des intervalles A tels que $\mathbb{P}(X \in A) > \mathbb{P}(Y \in A)$. Les lettres X ont alors tendance à s'agglomérer entre elles, ainsi que les lettres Y. Le *test des longueurs* a donc pour région de rejet $\{R \leq r\}$:

$$\phi = \mathbb{1}_{\{R \leq r\}}$$

Pour déterminer r tel que le test soit de niveau α approximativement, il faut déterminer la loi de R sous H_0.

Théorème 7.1 *Si H_0 est vérifiée et en supposant $m \leq n$, on a*

$$\begin{aligned}
\mathbb{P}(R = 2s) &= \frac{2\, C_{m-1}^{s-1}\, C_{n-1}^{s-1}}{C_{m+n}^{n}}, \\
\mathbb{P}(R = 2s+1) &= \frac{C_{m-1}^{s} C_{n-1}^{s-1} + C_{n-1}^{s} C_{m-1}^{s-1}}{C_{m+n}^{n}} \\
\mathbb{P}(R = 2m+1) &= \frac{C_{n-1}^{m}}{C_{m+n}^{n}}, \quad m < n
\end{aligned}$$

Dans les autres cas, on a $\mathbb{P}(R = r) = 0$. On a toujours $R \geq 2$. Si $m < n$, la valeur maximale de R est $2m+1$. Si $m = n$, la valeur maximale de R est $2m$.

Remarque 9 *Pour de grands échantillons, les calculs deviennent longs et pénibles. On utilise alors l'approximation normale décrite ci-dessous.*

Exemple : *Nous prenons*

$$\alpha = 0,05, \quad m = 5, \quad n = 6$$

Alors $C_{m+n}^{n} = C_{11}^{6} = 462$. Posons $N(s) = \mathbb{P}(R = s) \times 462$. Comme $0,05 \times 462 = 23,1$, il faut chercher r tel que

$$\begin{cases} N(2) + \ldots + N(r) \leq 23,1 \\ N(2) + \ldots + N(r+1) > 23,1 \end{cases}$$

Le calcul de $N(2), N(3), \ldots$ s'effectue en remplissant le tableau

s	0	1	2
$C_{m-1}^{s} = C_{4}^{s}$	1	4	6
$C_{n-1}^{s} = C_{5}^{s}$	1	5	10
$N(2s)$	0	2	40
$N(2s+1)$	0	9	70

Nous avons ici

$$\begin{cases} N(2) + N(3) = 2 + 9 = 11 \leq 23,1 \\ N(2) + N(3) + N(4) \;=\; 51 \;>\; 23,1 \end{cases}$$

Donc il faut prendre $r = 3$ pour construire le test $\left(\phi = \mathbb{1}_{\{R \leq 3\}}\right)$. Dans le cas de grands échantillons, on peut montrer que sous H_0, la loi de R est proche d'une loi normale et que

$$\mathbb{E}(R) = 1 \;+\; \frac{2mn}{m+n}, \qquad \mathrm{Var}(R) = \frac{2mn\,(2mn - m - n)}{(m+n)^2\,(m+n-1)}$$

7.5 Test de Wilcoxon dans le cas d'observations non couplées

On considère deux variables aléatoires continues X et Y, et on se propose de tester H_0 : X et Y ont la même loi contre H_1 : Y est stochastiquement supérieure à X. Pour cela, on considère un m-échantillon $X_1, \ldots, X_m$ de X et un n-échantillon $Y_1, \ldots, Y_n$ de Y. En rangeant les valeurs dans l'ordre croissant, on obtient une suite,

$$XYXXYYXYY \tag{7.1}$$

par exemple.

La statistique T de Wilcoxon est la somme des rangs occupés par les lettres X. Dans (7.1), $T = 1 + 3 + 4 + 7 = 15$. Sous H_1, les valeurs prises par X ont tendance à être inférieures à celles prises par Y. On est conduit au test de Wilcoxon de région de rejet $W = \{T \leq t\}$:

$$\phi = \mathbb{1}_{\{T \leq t\}}$$

Il faut déterminer t pour que le test soit de niveau α, au moins appoximativement. Lorsque m et n ne sont pas trop grands, il faut pour cela déterminer le nombre de cas où $T \leq t$, en rangeant les suites possibles dans l'ordre lexicographique.

Exemple 11 *Nous prenons*

$$\alpha = 0,05, \quad m = 6, \quad n = 5$$

Les possibilités sont

$$(1,2,3,4,5,6)\ ,\ (1,2,3,4,5,7)\ ,\ (1,2,3,4,5,8)\ ,\ (1,2,3,4,5,9)\ ,$$

$$(1,2,3,4,5,10)\ ,\ (1,2,3,4,5,11)\ ,\ (1,2,3,4,6,7)\ ,\ (1,2,3,4,6,8)\ ,$$

$$(1,2,3,4,6,9)\ ,\ (1,2,3,4,6,10)\ ,\ (1,2,3,4,7,8)\ ,\ (1,2,3,4,7,9)\ ,$$

$$(1,2,3,5,6,7)\ ,\ (1,2,3,5,6,8)\ ,(1,2,3,5,6,9)\ ,\ (1,2,3,5,7,8)\ ,$$

$$(1,2,4,5,6,7)\ ,\ (1,2,4,5,6,8)\ ,\ (1,3,4,5,6,7)$$

Il y a 19 *possibilités. Par exemple, si on veut comparer l'effet de deux anasthésiques A et B qui procurent des durées de sommeil aléatoires X et Y sachant que B a un effet qui n'est pas inférieur à celui de A, on va tester*

$$H_0 : B \text{ n'est pas meilleur que } A$$
$$H_1 : B \text{ est stochastiquement supérieur à } A$$

Avec les valeurs

A	$8,00$	$8,97$	$8,07$	$8,32$	$7,86$	$8,37$
B	$9,36$	$8,47$	$9,04$	$8,10$	$8,71$	

on obtient la suite

$$7,86,\ 8,00,\ 8,07,\ 8,10,\ 8,32,\ 8,37,\ 8,47,\ 8,97,\ 9,04,\ 9,36\,,$$

ou encore

$$AAABAABBABB$$

Ceci donne pour T, somme des rangs des A :

$$T = 1 + 2 + 3 + 5 + 6 + 9 = 26$$

Comme

$$\mathbb{P}\,(T \leq 26) = \frac{19}{C_{11}^5} = \frac{19}{462} = 0,041 \leq 0,05\,,$$

l'hypothèse H_0 : [B n'est pas meilleur que A] est à rejeter au niveau 5% (mais ce n'est pas net). Dans le cas des grands échantillons, on peut montrer que, sous H_0, T suit approximativement une loi normale et que

$$\mathbb{E}\,(T) = \frac{m\,(m+n+1)}{2}, \qquad Var\,(R) = \frac{mn\,(m+n+1)}{12}$$

7.6 Test des signes

Soient X et Y deux variables aléatoires couplées (définies sur le même espace d'épreuves). On suppose que $\mathbb{P}\,(X = Y) = 0$. On veut tester l'hypothèse

H_0 : $\mathbb{P}\,(Y < X) = \mathbb{P}\,(Y > X) = 1/2$, ou encore $Z = Y - X$ a pour médiane 0

contre

H_1 : $\mathbb{P}\,(Y > X) > 1/2$, ou encore $Z = Y - X$ a pour médiane $\mu > 0$

Pour cela, on considère n observations indépendantes du couple (X, Y), soit (X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n). Les n couples sont de deux sortes :

– ceux pour lesquels $Y_i > X_i$, en nombre N^+ ;
– ceux pour lesquels $Y_i < X_i$, en nombre N^-.

Nous avons $N^- + N^+ = n$. Sous H_0, pour tout i, $\mathbb{P}\,(Y_i > X_i) = 1/2$ et les deux cas ont autant de chances de se produire. Sous H_1, $\mathbb{P}\,(Y_i > X_i) > 1/2$ pour tout i. La variable N^+ a tendance à être plus grande que dans le cas de H_0. Il est donc naturel de rejeter H_0 lorsque N^+ est trop grand ou que N^- est trop faible. On est conduit au test des signes dont la région de rejet est

$$W = \left\{N^- \leq v\right\}$$

Au seuil α, on prend pour v le plus grand entier x tel que

$$\mathbb{P}\left(N^- \leq x\right) \leq \alpha$$

Sous H_0, N^- suit une loi binomiale Bin$(n, 1/2)$, donc

$$\mathbb{P}\left(N^- \leq x\right) = \left(\frac{1}{2}\right)^n \sum_{0 \leq i \leq x} C_n^i$$

Lorsque n n'est pas trop grand, ces nombres se calculent facilement.

Exemple (Rapidité comparée de deux percolateurs A et B) : *On fait opérer ces deux percolateurs* 16 *jours de suite. On note les résultats. On a ici avantage à les coupler, car d'un jour à l'autre les résultats peuvent être influencés pas la température de l'eau, la qualité du café, etc. Les résultats sont donnés dans le tableau suivant :*

Jours	1	2	3	4	5	6	7	8
A	10,0	11,4	7,6	12,3	7,8	10,8	10,1	7,5
B	11,8	13,7	10,4	10,4	9,6	11,9	10,8	8,3
Jours	9	10	11	12	13	14	15	16
A	7,1	8,4	9,0	11,2	11,5	11,3	9,4	10,2
B	8,5	8,2	10,2	9,7	12,1	12,2	9,1	14,6

Exemple 12 *La question posée est la suivante : ces résultats contredisent-ils l'hypothèse nulle que les percolateurs A et B sont également rapides ? On peut appliquer le test des signes au seuil $\alpha = 0,05$. Ici, A a été plus rapide que B quatre jours : les jours* 4, 10, 12 *et* 15*. Donc $N^- = 4$. On peut calculer :*

$$\mathbf{P}\left(N^- \leq 4\right) = 0,038 < 0,05$$

Il faut donc rejeter l'hypothèse nulle et considérer que A est moins rapide que B.

7.7 Test de Wilcoxon dans le cas d'observations couplées

Soient X et Y deux variables aléatoires couplées (définies sur le même espace d'épreuves). On suppose que $\mathbf{P}\left(X = Y\right) = 0$. On veut tester l'hypothèse

$$\begin{aligned} H_0 \quad &: \quad \mathbf{P}\left(Y < X\right) = \mathbf{P}\left(Y > X\right) = \frac{1}{2} \text{ ou encore } Z = Y - X \text{ a pour médiane } 0 \\ &\quad \text{contre} \\ H_1 \quad &: \quad \mathbf{P}\left(Y > X\right) > \frac{1}{2} \text{ ou encore } Z = Y - X \text{ a pour médiane } \mu > 0 \end{aligned}$$

Si l'on suppose $Z = Y - X$ continue, ceci revient à tester

$$\begin{aligned} H_0 \quad &: \quad Z \text{ est symétrique} \\ &\quad \text{contre} \\ H_1 \quad &: \quad Z \text{ est stochastiquement supérieure à } -Z \end{aligned}$$

Pour cela, on considère n observations indépendantes du couple (X, Y), soit (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n). On range ensuite les valeurs $Z_i = Y_i - X_i$ dans l'ordre des valeurs absolues croissantes (donc sans se préoccuper du signe). En ne retenant ensuite dans ce classement que les signes des Z_i, on obtient une suite de signes + et −, par exemple :

$$- \; - \; - \; + \; - \; + \; + \; + \; + \tag{7.2}$$

Le nombre de suites possibles telles que (7.2) est 2^n. Pour chaque échantillon de taille n, on note W^+ la somme des rangs des signes + et W^- la somme des rangs de signe −. Ainsi, pour (7.2) nous avons

$$\begin{aligned} W^+ &= 4 + 6 + 7 + 8 + 9 = 34 \\ W^- &= 1 + 2 + 3 + 5 = 11 \end{aligned}$$

Remarquons que $W^+ + W^- = n(n+1)/2$. Sous H_0, les suites de n signes telles que (7.2) sont uniformément distribuées. En particulier, $\mathbb{P}(W^+ > W^-) = \mathbb{P}(W^+ < W^-)$. Sous H_1, les signes + ont tendance à être plus nombreux que les signes −, et les Z_i négatifs ont tendance à avoir de plus petites valeurs absolues que les Z_i positifs. Donc les signes − sont alors plutôt situés au début de la suite, et ils sont en général moins nombreux que les signes +. On est ainsi conduit à la région de rejet

$$\{W^- \leq w\}$$

Pour de petites valeurs de n, $\mathbb{P}\{W^- \leq w\}$ se calcule facilement sous H_0, en dénombrant les cas favorables.

Exemple : *On essaie un traitement pour des grains de blé avant de les semer. On fait n fois l'expérience : un terrain est partagé en deux parcelles de natures aussi voisines que possible. Sur l'une, on sème des grains de blé ordinaires ; sur l'autre, la même quantité de grains de blé ayant subi le traitement. On obtient les récoltes respectives X et Y. Les résultats obtenus sont les suivants :*

Parcelles	X	Y	$Z = X - Y$
1	147	167	20
2	169	163	−6
3	161	166	5
4	150	201	51
5	166	198	32
6	190	199	9
7	168	160	−8
8	174	185	11

Ces résultats conduisent-ils à rejeter l'hypothèse nulle que le traitement n'amène aucune amélioration ? Pour répondre à cette question avec le test de Wilcoxon, il faut ranger les Z_i par ordre de valeurs absolues croissantes. On obtient :

$$5, \; -6, \; -8, \; 9, \; 11, \; 20, \; 32, \; 51$$

On a donc $W^- = 2 + 3 = 5$. En supposant que le traitement n'amène pas d'amélioration, on calcule $\mathbb{P}(W^- \leq 5)$ en comptant les cas favorables. On construit pour cela le tableau

suivant :

Evénement	Rangs des signes −	Nombre
$W^- = 0$	Pas de signes − (une suite de signes +)	1
$W^- = 1$	(1)	1
$W^- = 2$	(2)	1
$W^- = 3$	(3) ; (1, 2)	2
$W^- = 4$	(4) ; (1, 3)	2
$W^- = 5$	(5) ; (1, 4) ; (2, 3)	3

Il y a donc 10 *suites favorables et* $2^8 = 256$ *suites possibles. On a donc*

$$\mathbb{P}\left(W^- \leq 5\right) = \frac{10}{256} = 0,099$$

Ce nombre étant nettement inférieur à $0,05$*, on est amené à rejeter au seuil* $0,05$ *l'hypothèse que le traitement n'apporte pas d'amélioration.*

Lorsque n *est grand, on montre que* W^- *suit approximativement une loi normale sous* H_0*, avec :*

$$\mathbb{E}\left(W^-\right) = \frac{n\,(n+1)}{4} \quad \textit{et} \quad \mathrm{Var}\left(W^-\right) = \frac{n\,(n+1)\,(2n+1)}{24}$$

7.8 Tests d'indépendance de Spearman et Kendall

Lorsqu'on étudie un phénomène quantitatif, on peut se demander si les variations sont déterminées, prévisibles ou tout simplement dues au hasard. Nous devons résoudre le problème de décider, au vu d'un échantillon $X_1, \ldots, X_n$, s'il s'agit d'un échantillon aléatoire, c'est- à -dire de variables aléatoires indépendantes et de même loi. Le test a alors pour hypothèse nulle H_0 : Les variables aléatoires X_i sont indépendantes et de même loi de fonction de répartition continue F. Le but d'un tel test, permettant de définir l'alternative, est de déceler, par exemple, une tendance monotone, une tendance périodique, etc.

Sous H_0, l'ordre dans lequel on écrit les X_i n'a pas d'importance tandis que sous l'alternative H_1, il est essentiel. Les statistiques d'ordre vont nous être utiles pour définir ici une procédure de test. On associe à chaque réalisation de X_i son rang i dans l'échantillon tel qu'il se présente, ainsi que son rang R_i dans l'échantillon ordonné. C'est-à-dire que $X_{R_i,n} = X_i$.

Rappelons enfin qu'une manière simple de mesurer (grossièrement) l'indépendance de deux variables aléatoires X et Y consiste à calculer le coefficient de corrélation linéaire empirique entre un échantillon $X_1, \ldots, X_n$ de X et un échantillon $Y_1, \ldots, Y_n$ de Y.

Test du coefficient de corrélation de Spearman

Nous allons appliquer cette méthode pour savoir s'il y a indépendance par rapport à l'ordre. Pour cela, on considère les rangs $1, 2, \ldots, n$ de l'échantillon tel qu'il se présente, et les rangs relatifs à l'échantillon ordonné, $R_1, R_2, \ldots, R_n$. Le coefficient de corrélation de Spearman est défini par

$$r_s \;=\; \frac{\sum_{i=1}^{n}(R_i \,-\, \bar{R})(i \,-\, \bar{i})}{\sqrt{\sum_{i=1}^{n}(R_i \,-\, \bar{R})^2 \,\sum_{i=1}^{n}(i \,-\, \bar{i})^2}}\,,$$

où

$$\bar{R} = \bar{i} = \frac{1}{n}\sum_{i=1}^{n} i \;=\; \frac{n+1}{2}$$

En remarquant que

$$\sum_{i=1}^{n}(R_i \,-\, \bar{R})^2 \;=\; \sum_{i=1}^{n}(i \,-\, \bar{i})^2 \;=\; \sum_{i=1}^{n} i^2 \,-\, n\bar{i}^2 \;=\; \frac{n(n^2\,-\,1)}{12}\,,$$

nous obtenons une expression plus simple pour r_s :

$$r_s \;=\; 1 \;-\; \frac{6}{n(n^2\,-\,1)}\sum_{i=1}^{n}(R_i \,-\, i)^2$$

Exemple : *On a relevé toutes les heures, le 31 juillet 2000 (24 heure = 00 heure le 1er août 2000), le taux d'ozone en* $\mu g/m^3$ *e*n zone urbaine.

Heures	1	2	3	4	5	6	7	8	9	10	11	12
Taux	36	37	38	14	19	11	14	27	52	65	85	91
i	1	2	3	4	5	6	7	8	9	10	11	12
Valeurs ordonnées	11	14	14	19	27	36	37	38	52	53	55	58
R_i	6	7	4	5	8	1	2	3	9	22	24	21
Heures	13	14	15	16	17	18	19	20	21	22	23	24
Taux	98	100	104	111	110	104	89	90	58	53	61	55
i	13	14	15	16	17	18	19	20	21	22	23	24
Valeurs ordonnées	61	65	85	89	90	91	98	100	104	104	110	111
R_i	23	10	11	19	20	12	13	14	15	18	17	16

On peut alors calculer

$$r_s = 1 \,-\, \frac{6 \times 940}{24 \times 575} = 0,5913$$

Cette nouvelle expression nous permet de bien voir ce qui se passe dans le cas d'une tendance monotone. En effet, dans le cas d'une tendance monotone croissante, les rangs dans l'échantillon ordonné sont à peu près les mêmes que dans l'échantillon et $r_s \approx 1$. Dans le cas d'une tendance monotone décroissante, les rangs sont inversés : schématiquement, $R_1 = n$, $\ldots$, $R_n = 1$. Cela se traduit par $R_i = n + 1 - i$ et conduit à $r_s = -1$. Nous pouvons donc construire la région de rejet du test de Spearman selon l'alternative considérée (selon que le test est unilatéral ou bilatéral) :

$\{r_s > c\}$	pour une tendance monotone croissante
$\{\lvert r_s\rvert > c\}$	pour une tendance quelconque
$\{\lvert r_s\rvert > c\}$	pour une tendance monotone décroissante

La constante c est déterminée par le niveau α et le caractère, unilatéral ou bilatéral, du test. Il nous faut donc déterminer sous H_0 la loi de r_s.

Théorème 7.2 *Si* H_0 *est vérifiée,*

$$\sqrt{n-1}\, r_s \;\xrightarrow{\;d\;}\; \mathcal{N}(0,\,1) \quad \textit{quand } n \to \infty$$

Pour $11 \leq n \leq 30$,

$$r_s \sqrt{\frac{n-2}{1-r_s^2}} \approx T_{n-2}\,,$$

où T_{n-2} désigne la loi de Student à $n-2$ degrés de liberté.

Pour $n \leq 10$, Kendall a tabulé entièrement la loi de r_s, ou plutôt la loi de $S = \sum_{i=1}^{n}(R_i - i)^2$.

Eléments de démonstration. On montre d'abord que la loi de r_s est symétrique,

$$\mathbb{E}\left(r_s^{2k+1}\right) = 0 \ \text{ pour tout entier naturel } k$$

On calcule ensuite la variance de r_s. Le point clé de la démonstration est que sous H_0, nous avons si $i \neq j$,

$$\begin{aligned}
\mathrm{Cov}(R_i,\, R_j) &= \frac{\sum_{i\neq j} r_i r_j}{n(n-1)} - \mathbb{E}\,(R_i)\,\mathbb{E}\,(R_j) \\
&= \frac{1}{n(n-1)}\left(\frac{n^2(n+1)^2}{4} - \frac{n(n+1)(2n+1)}{6}\right) - \frac{(n+1)^2}{4} \\
\mathrm{Cov}(R_i,\, R_j) &= -\frac{(n+1)}{12}
\end{aligned}$$

Puis, plus généralement on montre que pour tout $k \in \mathbb{N}$,

$$\lim_{n\to+\infty} (n-1)^k\, \mathbb{E}\left(r_s^{2k}\right) = \frac{(2k)!}{2^k k!}\,,$$

qui sont les moments de la loi normale centrée réduite. On utilise donc les quantiles de la loi normale centrée réduite lorsque $n > 30$. Lorsque $11 \leq n \leq 30$, la valeur critique est donnée par

$$\frac{t}{\sqrt{n-2+t}}, \qquad t \text{ étant un quantile de la loi de Student } T_{n-2}$$

Exemple : *pour les* 24 *données précédentes, au niveau* 5%,

$$t = 1,717 \ \text{ pour le test unilat'eral et } \ n-2 = 22, \ \ r_s = 0,5913 > 0,3526$$

Donc H_0 est rejetée au niveau 5%.

Test du coefficient de corrélation de Kendall

Pour tester le caractère aléatoire d'un échantillon, une autre méthode s'inspire de la remarque suivante : s'il existe une tendance monotone croissante, l'échantillon et l'échantillon ordonné auront tendance à être semblables. Dans l'échantillon, cela se traduira par un petit nombre d'inversions, c'est-à-dire de cas où pour $i < j$ on a $X_i > X_j$. Nous allons donc considérer le nombre d'inversions

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}_{\mathbb{R}_+^*}(X_i - X_j)$$

On remarque que la statistique Q mesure aussi le nombre d'inversions de la statistique de rang $R = (R_1, \ldots, R_n)$. Dans l'exemple, $Q = 5+5+5+1+2+0+0+0+0+4+4+6+6$

$+ \ldots = 82$. Regardons maintenant les valeurs que peut prendre la statistique Q. Le nombre maximum d'inversions est atteint lorsque pour tout couple $i < j$ nous avons $X_i > X_j$. Dans ce cas, il y a toujours inversion et les X_i sont en ordre décroissant : on est en présence d'une tendance monotone décroissante. Le nombre maximum d'inversions est donc

$$N_{\max} = C_n^2 = \frac{n(n-1)}{2}$$

Lorsque au contraire, les X_i sont en ordre croissant, il n'y a pas d'inversion et $Q = 0$. La statistique Q peut varier de 0 à $n(n-1)/2$. Sa loi peut être déterminée par sa fonction génératrice (transformée de Laplace)

$$G(u) = \mathbb{E}\left(e^{uQ}\right)$$

On peut montrer que si la taille n de l'échantillon est assez grande,

$$\mathbb{E}(Q) \approx \frac{n(n-1)}{4}, \qquad \mathrm{Var}(Q) = \frac{n(n-1)(2n+5)}{72}$$

Le coefficient de corrélation de Kendall est défini à partir de Q par

$$\tau = 1 \; - \; \frac{4Q}{n(n-1)},$$

où τ varie de -1 à 1. Sous H_0,

$$\mathbb{E}(\tau) = 0, \qquad \mathrm{Var}(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

On montre que τ est asymptotiquement distribué selon une loi normale.

Exemple : *Pour les* 24 *données précédentes, au niveau* 5%, $\tau = 28/69 = 0,406$, *et pour un test bilatéral on rejette* H_0 *si* $|\tau| > 0,286$.

Exercice 7.1 *On relève les poids en grammes de* 10 *pommes de qualités A et B.*

A	192	197	207	182	191
B	212	201	209	214	203

Les résultats permettent-ils de rejeter l'hypothèse que le poids d'une pomme a même loi pour les qualités A et B ? On fera un test des longueurs.

Exercice 7.2 *On veut essayer un produit chimique comme insecticide. Pour cela, on compte les insectes dans un lopin de terre de même aire. On obtient les résultats suivants pour* 13 *lopins, dont* 5 *ont été traités.*

Lopins non traités	45	88	16	6	28	122	62	13
Lopins traités	23	104	2	9	30			

Utiliser la procédure de Wilcoxon pour déterminer au seuil 5% *si l'insecticide a une action efficace.*

Exercice 7.3 *Un fabricant de crème antisolaire veut savoir s'il améliore sa crème en ajoutant un nouvel ingrédient. Il choisit* 7 *volontaires et leur enduit la moitié du dos avec une crème, l'autre moitié avec l'autre crème. Il mesure la noirceur de la peau après exposition.*

Volontaire	1	2	3	4	5	6	7
Ancienne crème	42	51	31	61	44	55	48
Nouvelle crème	38	53	36	52	33	49	36

Doit-on rejeter l'hypothèse selon laquelle le nouvel ingrédient n'apporte aucune amélioration au niveau $0,1$ *? Faire un test des signes et un test de Wilcoxon. Que conseiller au fabricant ?*

Exercice 7.4 *On mesure la pression sanguine systolique de* 11 *patients avant et après administration d'un médicament dont on sait qu'il peut la faire baisser, mais pas l'augmenter. Pour chaque patient, la baisse de la pression sanguine (pression avant moins pression après) est :*

$$7,\ 5,\ 12,\ -3,\ -5,\ 2,\ 14,\ 18,\ 19,\ 21,\ -1$$

Utiliser un test non paramétrique pour voir si ces observations contredisent l'hypothèse H_0 *: pas de modification systématique de la pression systolique.*

Exercice 7.5 *Dans un laboratoire, on note le nombre de jours où le système informatique a fonctionné correctement jusquà une défaillance :*

$$4,\ 2,\ 12,\ 22,\ 12,\ 34,\ 15,\ 8,\ 32,\ 3,\ 1,\ 10,\ 12$$

Peut-on dire que ces données sont des réalisations indépendantes et de même loi de la variable "Nombre de bons fonctionnements avant une défaillance" ?

Conclusion

- Retenir l'idée d'ordre stochastique.
- Attention à bien distinguer les variables non couplées des variables couplées.
- Utiliser la détermination exacte de la distribution statistique sous l'hypothèse nulle pour les petits échantillons et l'approximation gaussienne pour les grands échantillons.
- Retenir le coefficient de corrélation entre les rangs de l'échantillon et les rangs de l'échantillon ordonné, ainsi que le nombre d'interversions dans l'échantillon, comme des statistiques permettant de tester le caractère aléatoire d'un échantillon.

Emmanuelle CRÉTOIS

Eléments bibliographiques pour ce chapitre

Association pour la Statistique et ses Utilisations (1996) *Inférence non paramétrique. Les statistiques de rangs*, Jean-Jacques Droesbeke et Jeanne Fine éditeurs, Collection "Ellipses", Editions de l'Université Libre de Bruxelles : Bruxelles.

Capéraà, P. & Van Cutsem, B. (1988) *Méthodes et modèles en Statistique non paramétrique*. Dunod : Paris.

Conover, W. J. (1971) *Practical Nonparametric Statistics*. Wiley : New York.

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Dagnelie P. (1975) *Théorie et méthodes statistiques*, Tome 2. Vander-Oyez : Bruxelles, 2e édition.

Ferguson, T. S. (1967) *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press : New York and London.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Lecoutre, J.-P. & Tassi, P. (1987) *Statistique non paramétrique et robustesse*. Collection "Economie et statistiques avancées", Economica : Paris.

Monfort, A. (1982) *Cours de statistique mathématique*. Collection "Economie et statistiques avancées", Economica : Paris.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley : New York.

Tassi, P. (1985) *Méthodes statistiques*. Collection "Economie et statistiques avancées", Economica : Paris.

Rayner, J. C. W. & Best, D. J. (1989) *Smooth Tests of Goodness-of-Fit*. Oxford University Press : Oxford.

8

Le test du chi-deux

8.1 Tester $[\mu = \mu_0]$ contre $[\mu \neq \mu_0]$(μ, μ_0 mesures de probabilité)

Principe : Soient $X_1, \ldots, X_n$ des v.a. i.i.d. de loi μ sur l'espace mesurable $(E, \mathcal{E})$. Soit μ_0 une loi de probabilité donnée sur $(E, \mathcal{E})$. On souhaite tester $[\mu = \mu_0]$ contre $[\mu \neq \mu_0]$ sans faire aucune hypothèse de modélisation paramétrique.

On propose de former une partition finie de E en K parties $J_1, \ldots, J_K$.

Pour chaque k, $1 \leq k \leq K$, on compte le nombre de points X_i de l'échantillon appartenant à la partie J_k. Soient n_k ce nombre et N_k la v.a. correspondante.

Le nombre théorique attendu de points dans J_k est

$$\mathbb{E}[N_k] = n\mu(J_k) = np_k, \tag{8.1}$$

en notant $p_k = \mu(J_k)$.

Pour tester $H_0\,[\mu = \mu_0]$, on veut mesurer un écart entre le vecteur (colonne) $(n_1, \ldots, n_K)^T$ et le vecteur (colonne) $(np_1, \ldots, np_K)^T$. Pour aboutir à une loi limite ($n \to +\infty$) indépendante des valeurs de μ_0, on considère la forme quadratique

$$\mathcal{T}_n^2 = \sum_{k=1}^{K} \frac{(n_k - np_k)^2}{np_k} \quad (p_k \text{ supposés} > 0,\ \forall k) \tag{8.2}$$

Théorème 8.1 *Pour $\mu = \mu_0$ et $p_k = \mu_0(J_k)$, sous H_0,*

$$\mathcal{T}_n^2 \xrightarrow{d} \chi^2_{K-1} \quad \text{quand } n \to +\infty \tag{8.3}$$

Avant de démontrer ce résultat, nous allons introduire la mesure empirique, le processus empirique associé, et établir un lemme qui nous servira par la suite.

Définition 8.1

1. ***Mesure empirique*** *: C'est la mesure aléatoire sur $(E, \mathcal{E})$:*

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \tag{8.4}$$

Par définition, pour tout $A \in \mathcal{E}$,

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i \in A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A(X_i) \tag{8.5}$$

Pour toute fonction $f : E \to \mathbb{R}$ mesurable et finie :

$$\mu_n(f) = \int f \, d\mu_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{8.6}$$

2. ***Processus empirique associé*** *: C'est $\mathbb{Z}_n = \sqrt{n}\,[\mu_n - \mu]$.*

Lemme 8.1 *1. $\forall A \in \mathcal{E}$, $\mathbb{Z}_n(A) \xrightarrow{d} \mathcal{N}(0, \mu(A)(1-\mu(A)))$. On a un résultat analogue pour les lois fini-dimensionnelles.*

2. $\forall A, B \in \mathcal{E}$, $\mathbb{E}[\mathbb{Z}_n(A)\mathbb{Z}_n(B)] = \mu(A \cap B) - \mu(A)\mu(B)$. On a un résultat analogue pour les lois fini-dimensionnelles.

Démonstration.

1. Clair par le TLC.
2. En effet :

$$\mathbb{E}[\mathbb{Z}_n(A)\mathbb{Z}_n(B)] = \frac{1}{n}\sum_{i,j=1}^{n} \mathbb{E}\left([\mathbb{1}_A(X_i) - \mu(A)]\,[\mathbb{1}_B(X_i) - \mu(B)]\right)$$

Comme la v.a. $\mathbb{1}_A(X_i) - \mu(A)$ est centrée pour tout i et par indépendance, il ne subsiste que les termes diagonaux (c'est-à-dire où $i = j$). ■

Démonstration du théorème. On remarque d'abord (voir le lemme 8.1 (1)) que

$$\mathcal{T}_n^2 = \sum_{k=1}^{K} \frac{\mathbb{Z}_n^2(J_K)}{\mu(J_K)} \xrightarrow{d} \mathcal{T}^2 = \sum_{k=1}^{K} \frac{\mathbb{Z}^2(J_k)}{\mu(J_k)} \quad \text{quand } n \to +\infty \tag{8.7}$$

Posons

$$\underline{Y} = (Y_1, \ldots, Y_k)^T = \left(\frac{\mathbb{Z}(J_1)}{\sqrt{\mu(J_1)}}, \ldots, \frac{\mathbb{Z}(J_K)}{\sqrt{\mu(J_K)}} \right)^T \tag{8.8}$$

(vecteur colonne de $\mathbb{R}^K$). Si on note, respectivement,

$$\langle \underline{x}, \underline{y} \rangle = \sum_{j=1}^{K} x_j y_j \quad \text{et} \quad \|\underline{x}\|^2 = \sum_{j=1}^{K} x_j^2$$

le produit scalaire canonique sur $\mathbb{R}^K$ et le carré de la norme associée, on a :

$$\mathcal{T}^2 = \|\underline{Y}\|^2 = \langle \underline{Y}, \underline{Y} \rangle \tag{8.9}$$

Calculons la matrice de variance du vecteur aléatoire $\underline{Y}$ de $\mathbb{R}^K$. D'après le lemme 8.1, cette matrice est, si l'on note $\lambda_k = \mu(J_k)$,

$$\Gamma = \begin{pmatrix} 1-\lambda_1 & & & & \\ & \ddots & & -\sqrt{\lambda_k \lambda_l} & \\ & & \ddots & & \\ & \sqrt{\lambda_k \lambda_l} & & \ddots & \\ & & & & 1-\lambda_k \end{pmatrix} \tag{8.10}$$

Posons $\alpha_k = \sqrt{\lambda_k}$, $1 \leq k \leq K$, et notons $\underline{a} = (\alpha_1, \ldots, \alpha_K)^T$. Comme

$$\sum_{k=1}^{K} \alpha_k^2 = \sum_{k=1}^{K} \lambda_k = \sum_{k=1}^{K} \mu(J_k) = \mu(E) = 1,$$

on trouve

$$\Gamma = I - \underline{a}\underline{a}^T \tag{8.11}$$

Ainsi,

$$\forall v \in \mathbb{R}^K, \quad \Gamma v = \underline{v} - \langle \underline{a}, \underline{v} \rangle \, \underline{a} \tag{8.12}$$

L'endomorphisme $\Gamma \in \mathcal{L}(\mathbb{R}^K)$ (identifié à sa matrice dans la base canonique de $\mathbb{R}^K$) est donc la projection orthogonale sur l'hyperplan $H = \underline{a}^\perp$ (orthogonal du vecteur $\underline{a}$). Soit $\underline{u}_1, \ldots, \underline{u}_K$ une base orthonormée de $\mathbb{R}^K$ constituée de vecteurs propres de Γ (une telle base existe parce que Γ est un opérateur symétrique) : plus précisément, $\underline{u}_1, \ldots, \underline{u}_{K-1} \in H$ et $\underline{u}_K = \underline{a}$, de sorte que $\Gamma \underline{u}_k = \underline{u}_k$ pour $1 \leq k \leq K-1$, et $\Gamma \underline{u}_K = 0$.

On pourrait alors directement appliquer les résultats présentés en Annexe (ACP). Cependant, nous développons ici le raisonnement.

Décomposons $\underline{Y}$ dans cette base orthonormée :

$$\underline{Y} = \sum_{k=1}^{K} \langle \underline{Y}, \underline{u}_k \rangle \, \underline{u}_k \tag{8.13}$$

Comme, par définition de l'opérateur de variance, on a :

$$\forall \underline{x} \text{ et } \underline{y} \in \mathbb{R}^K, \quad \mathbb{E}\left(\langle \underline{Y}, \underline{x} \rangle \langle \underline{Y}, \underline{y} \rangle \right) = \langle \Gamma \underline{x}, \underline{y} \rangle, \tag{8.14}$$

on obtient :

$$\forall 1 \leq j,k \leq K-1,\ \mathbb{E}\big(\langle \underline{Y}, \underline{u}_j\rangle \langle \underline{Y}, \underline{u}_k\rangle\big) = \delta_{jk}$$
$$\mathbb{E}\,\langle \underline{Y}, \underline{u}_K\rangle = \langle \Gamma \underline{u}_K, \underline{u}_K\rangle = 0 \tag{8.15}$$

En conséquence, les v.a. $\xi_j = \langle \underline{Y}, \underline{u}_j\rangle, 1 \leq j \leq K-1$, sont i.i.d. $\mathcal{N}(0,1)$, tandis que $\langle \underline{Y}, \underline{u}_K\rangle = 0$. D'où la décomposition en système orthonormal

$$\underline{Y} = \sum_{k=1}^{K-1} \xi_k \underline{u}_k, \tag{8.16}$$

donc

$$\mathcal{T}^2 = \|\underline{Y}\|^2 = \sum_{k=1}^{K-1} \xi_k^2 \sim \chi^2_{K-1} \tag{8.17}$$

■

8.2 Le chi-deux comme test d'adéquation

La problématique des tests d'adéquation pour les familles paramétrées de lois probabilités, ou modèles de lois, est la suivante : étant donné un échantillon i.i.d. (ou supposé tel) $X_1, \ldots, X_n$, est-il raisonnable de poser que ces v.a. sont issues de tel modèle de lois de probabilité (par exemple, $\mathcal{N}(\mu, \sigma^2)$, $\mathrm{Exp}(\theta)$, etc.), sans faire d'hypothèse sur la valeur inconnue du paramètre (par exemple, sur (μ, σ^2), sur θ, etc.) ?

Dans le cas du chi-deux, l'idée naturelle est de remplacer, sous l'hypothèse H_0 (l'échantillon est issu d'une loi de la forme $\mu_\theta = f_\theta\, d\nu$, $\theta \in \Theta$, θ inconnu), le vrai paramètre θ_0, inconnu, par un estimateur suffisamment bon, $\widehat{\theta}_n$.

La statistique de test devient

$$\widehat{\mathcal{T}}_n^2 = \sum_{k=1}^{K} \frac{(n_k - n\hat{p}_k)^2}{n\widehat{p}_k} \tag{8.18}$$

Elle correspond au processus modifié $\widehat{\mathbb{Z}}_n = \sqrt{n}\left[\mu_n - \mu_{\widehat{\theta}_n}\right]$, qui se décompose sous la forme

$$\widehat{\mathbb{Z}}_n = \mathbb{Z}_n - \sqrt{n}\left[\mu_{\widehat{\theta}_n} - \mu_0\right], \tag{8.19}$$

où $\mu_0 = \mu_{\theta_0}$. Sous des hypothèses de régularité, puisque $\widehat{\theta}_n \to \theta_0$ p.s. sous H_0, on a approximativement

$$\forall 1 \leq k \leq K,\ \ \widehat{\mathbb{Z}}_n(J_k) \approx \mathbb{Z}_n(J_k) - \sqrt{n}\left[\widehat{\theta}_n - \theta_0\right] \dot{\mu}_0(J_k), \tag{8.20}$$

où on a noté $\dot{\mu}_0 = \dot{\mu}_{\theta_0}$, avec $\dot{\mu}_\theta = (\mathrm{d}/\mathrm{d}\theta)\ \mu_\theta$: on suppose dans toute cette partie, pour simplifier, que θ est réel. On notera, comme ci-dessus :

$$\lambda_k = \mu_0(J_k) = \alpha_k^2,$$

avec

$$\sum_{k=1}^{K} \alpha_k^2 = 1\dot{\lambda}_k,$$

$$\dot{\mu}_0(J_k) = 2\dot{\alpha}_k \alpha_k,$$

$$\dot{\alpha}_k = \dot{\lambda}_k / 2\sqrt{\lambda_k},$$

$$\underline{\widehat{Y}} = (\widehat{Y}_1, \ldots, \widehat{Y}_K) = \left(\frac{\widehat{\mathbb{Z}}(J_1)}{\sqrt{\lambda_1}}, \ldots, \frac{\widehat{\mathbb{Z}}(J_K)}{\sqrt{\lambda_K}} \right),$$

où $\widehat{\mathbb{Z}}_n(A) \xrightarrow{d} \widehat{\mathbb{Z}}(A)$.

Supposons que $\mu_\theta = f_\theta \, d\nu$, θ réel, avec les hypothèses usuelles de régularité. Considérons l'estimateur $\widehat{\theta}_n$ du MV. On sait que :

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{s_0(X_i)}{I_0} \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_0}\right), \tag{8.21}$$

avec :

$$s_0(x) = s_{\theta_0}(x) = \frac{\partial}{\partial \theta} (\ln f_\theta(x)) \Big|_{\theta=\theta_0} \quad \text{(fonction score en} \theta_0) \ ,$$

$$\mathbb{E}_{\theta_0}[s_0(X)] = \int s_0(x) f_0(x) \quad \nu(dx) = \int s_0 \quad d\mu_0 = 0 \quad (f_0 = f_{\theta_0}),$$

et

$$I_0 = \mathbb{E}_{\theta_0}\left[s_0^2(X)\right] = \int s_0^2 \quad d\mu_0 = -\int \frac{\partial^2}{\partial \theta^2} (\ln f_\theta(x)) \Big|_{\theta=\theta_0} \quad \mu_0(dx) > 0$$

Théorème 8.2 *Sous H_0, si $\widehat{\theta}_n$ est l'estimateur du MV et sous les hypothèses usuelles de régularité, on a, quand $n \to +\infty$:*

$$\widehat{\mathcal{T}}_n^2 \xrightarrow{d} \widehat{\mathcal{T}}^2 = \sum_{k=1}^{K-2} \xi_k^2 + \ \gamma_0 \xi_{K-1}^2, \tag{8.22}$$

avec $0 \leq \gamma_0 \leq 1$, ξ_k i.i.d. et $\mathcal{N}(0, 1)$ $(1 \leq k \leq K-1)$. Plus précisément,

$$\gamma_0 = 1 \ - \ I_0^{-1} \sum_{k=1}^{K} \frac{\left(\int_{J_k} s_0 \, \mathrm{d}\mu_0 \right)^2}{\mu_0(J_k)} \tag{8.23}$$

Esquisse de la démonstration. On part de l'approximation (8.20). Un calcul de covariance donne, sous H_0 :

$$\text{Var} \left[\widehat{\mathbb{Z}}_n(J_k) \right] \approx \lambda_k(1 - \lambda_k) \ - \ \frac{\dot{\mu}_0^2(J_k)}{I_0}, \quad 1 \leq k \leq K, \tag{8.24}$$

car

$$\dot{\mu}_0(J_k) = \int_{J_k} \dot{f}_0 \, d\nu = \int_{J_k} s_0 \, d\mu_0$$

De même, pour $j \neq k$,

$$\text{Cov} \left[\widehat{\mathbb{Z}}_n(j), \widehat{\mathbb{Z}}_n(k) \right] = -\lambda_j \lambda_k - \frac{\dot{\mu}_0(J_j)\dot{\mu}_0(J_k)}{I_0}$$

Par conséquent, la matrice de variance $\widehat{\Gamma}$ de $\underline{\widehat{Y}}$ s'exprime ainsi :

$$\widehat{\Gamma} = I - \underline{a}\underline{a}^T - \underline{b}\underline{b}^T, \tag{8.25}$$

avec

$$\underline{b} = \frac{1}{\sqrt{I_0}} \left(\frac{\dot{\lambda}_1}{\sqrt{\lambda_1}}, \ldots, \frac{\dot{\lambda}_K}{\sqrt{\lambda_K}} \right)^T \tag{8.26}$$

Or, puisque l'on a, pour tout θ,

$$\sum_{k=1}^{K} \alpha_k^2(\theta) = \sum_{k=1}^{K} \mu_\theta(J_k) = 1, \tag{8.27}$$

il en résulte par dérivation que

$$\forall \theta, \quad \sum_{k=1}^{K} \alpha_k(\theta)\dot{\alpha}_k(\theta) = 0, \tag{8.28}$$

avec

$$\dot{\alpha}_k(\theta) = \dot{\mu}_\theta(J_k)/2\sqrt{\mu_\theta(J_k)},$$

et donc

$$\dot{\alpha}_k = \dot{\lambda}_k/2\sqrt{\lambda_k}$$

Par suite,

$$\langle \underline{a}, \underline{b} \rangle = \frac{1}{\sqrt{I_0}} \sum_{k=1}^{K} \alpha_k \frac{\dot{\lambda}_k}{\sqrt{\lambda_k}} = 0 \tag{8.29}$$

et

$$0 \leq \|\underline{b}\|^2 = \frac{1}{I_0} \sum_{k=1}^{K} \frac{\dot{\lambda}_k^2}{\lambda_k} = \frac{1}{I_0} \sum_{k=1}^{K} \frac{\left(\int_{J_k} s_0 \, d\mu_0 \right)^2}{\mu_0(J_k)} \leq 1, \tag{8.30}$$

car, d'après l'inégalité de Cauchy-Schwarz, on a

$$\left(\int_{J_k} s_0 \, d\mu_0 \right)^2 \leq \left(\int_{J_k} s_0^2 \, d\mu_0 \right) \mu_0(J_k), \;\; 1 \leq k \leq K, \tag{8.31}$$

et que

$$\sum_{k=1}^{K} \int_{J_k} s_0^2 \, d\mu_0 = \int s_0^2 \, d\mu_0 = I_0 \tag{8.32}$$

De plus, on a inégalité stricte dès que s_0^2 n'est pas constante μ_0-p.p. sur l'un des J_k.

Si on pose

$$\underline{b}_0 = \frac{\underline{b}}{\|\underline{b}\|}, \ \ \|\underline{b}_0\| = 1, \ \ \underline{a} \perp \underline{b}_0, \tag{8.33}$$

on a donc :

$$\widehat{\Gamma} = I - \underline{a}\underline{a}^T - \|\underline{b}\|^2 \ \underline{b}_0 \underline{b}_0^T, \tag{8.34}$$

ou encore, en posant $\gamma_0 = 1 - \|\underline{b}\|^2$, $0 \leq \gamma_0 \leq 1$:

$$\forall x, \quad \widehat{\Gamma}\underline{x} = \underline{x} - \langle \underline{a}, \underline{x} \rangle \, \underline{a} - \langle \underline{b}_0, \underline{x} \rangle \, \underline{b}_0 + \gamma_0 \, \langle \underline{b}_0, \underline{x} \rangle \, \underline{b}_0, \tag{8.35}$$

donc les valeurs propres de $\widehat{\Gamma}$ sont

$$1, \ldots, 1, \ \gamma_0 \quad 1 \text{ répété } K-2 \text{ fois} \qquad \blacksquare \tag{8.36}$$

Remarque 10 *Le nombre γ_0 mesure l'erreur relative que l'on commettrait en considérant que*

$$\sum_{k=1}^{K} \frac{\left(\int_{J_k} s_0 \, d\mu_0 \right)^2}{\mu_0 \left(J_k \right)} = I_0$$

Notons qu'en général, la limite quand $K \to +\infty$ du membre de gauche est bien égale à I_0, pourvu que l'on considère des partitions de plus en plus fines bien ajustées à s_0, f_0 et ν.

Application : Comme on a montré que

$$\sum_{k=1}^{K-2} \xi_k^2 \leq \widehat{\mathcal{T}}^2 \leq \sum_{k=1}^{K-1} \xi_k^2, \tag{8.37}$$

il en résulte que, si on note $q_{1-\alpha}(\nu)$, $\nu \in \mathbb{N}^*$, le $(1-\alpha)$-quantile de la loi χ^2_ν :

- Si $\widehat{\mathcal{T}}_n^2 < q_{1-\alpha}(K-2)$, on ne rejette pas H_0 ;
- Si $\widehat{\mathcal{T}}_n^2 > q_{1-\alpha}(K-1)$, on rejette H_0 ;
- Si $q_{1-\alpha}(K-2) \leq \widehat{\mathcal{T}}_n^2 \leq q_{1-\alpha}(K-1)$, on ne peut pas conclure sur la base du seul encadrement ci-dessus.

Généralisation : On peut montrer de même que, si le paramètre $\theta = (\theta_1, \ldots, \theta_q)$ est q-dimensionnel et si on utilise encore l'estimateur du MV $\widehat{\theta}_n$ pour θ, alors, sous les conditions usuelles de régularité, la v.a. limite $\widehat{\mathcal{T}}^2$ s'écrit (si $K \geq q+2$) :

$$\widehat{\mathcal{T}}^2 = \sum_{k=1}^{K-q-1} \xi_k^2 + \gamma_{K-q} \xi_{K-q}^2 + \cdots + \gamma_{K-1} \xi_{K-1}^2, \tag{8.38}$$

avec

$$0 \leq \gamma_{K-q}, \ldots, \gamma_{K-1} \leq 1$$

8.3 Adéquation, suite

Il serait beaucoup plus commode d'utiliser un estimateur de θ_0 qui, introduit dans $\mathcal{T}_n^2$, aboutirait à une loi limite indépendante de θ_0 et, si possible, facilement tabulable.

L'interprétation de γ_0 en 8.2 suggère de remplacer l'estimateur $\widehat{\theta}_n$ du MV par l'estimateur $\overline{\theta}_n$ du MV basé sur les données groupées dans les J_k, $1 \leq k \leq K$: ainsi, $\overline{\theta}_n$ est solution (on suppose encore θ unidimensionnel, pour simplifier) de :

$$\sum_{k=1}^{K} n_k \, \frac{\dot{p}_k(\theta)}{p_k(\theta)} = 0 \tag{8.39}$$

La fonction de score est alors

$$s_\theta(x) = \sum_{k=1}^{n} \frac{\dot{p}_k(\theta)}{p_k(\theta)} \, \mathbb{1}_{J_k}(x)$$

Une autre possibilité est de remplacer $\widehat{\theta}_n$ par l'estimateur $\overline{\theta}_n$ qui minimise

$$\sum_{k=1}^{K} \frac{(n_k - n\overline{p}_k(\theta))^2}{n\overline{p}_k}$$

Nous allons établir un résultat concernant ce dernier choix (minimisation de la distance du chi-deux), mais on obtiendrait la même chose avec le premier (données groupées).

Théorème 8.3 *Sous les conditions usuelles de régularité convenablement adaptées, la loi limite de la statistique de test*

$$\overline{T}_n^2 = \sum_{k=1}^{K} \frac{(n_k - n\overline{p}_k)^2}{n\overline{p}_k} \tag{8.40}$$

quand $n \to +\infty$, où $\overline{p}_k = p_k(\overline{\theta}_n)$, est

$$\overline{T}^2 = \sum_{k=1}^{K-q-1} \xi_k^2, \quad \xi_k \text{ i.i.d. } \mathcal{N}(0,\, 1), \tag{8.41}$$

où $\dim \Theta = q \geq 1$ est la dimension de l'espace des paramètres. C'est donc une loi χ^2_{K-q-1} (pour $K \geq q+2$).

On dit qu'on enlève un degré de liberté (dans la loi du chi-deux) par coordonnée indépendante du paramètre.

Ainsi, on trouve une loi du χ^2_{K-2} si on teste $\mathrm{Exp}(\theta)$ de cette manière, mais une loi χ^2_{K-3} si on teste $\mathcal{N}(\mu,\, \sigma^2)$, le couple $(\mu,\, \sigma^2)$ étant inconnu.

Principe de la démonstration. On procède comme ci-dessus.

On montre d'abord que, sous H_0, si θ est unidimensionnel :

$$\sqrt{n}(\overline{\theta}_n - \theta_0) \approx \frac{\sum_{k=1}^{K} \frac{\mathbb{Z}_n(J_k)}{\sqrt{\lambda_k}} \frac{\dot{\lambda}_k}{\sqrt{\lambda_k}}}{\sum_{k=1}^{K} \frac{\dot{\lambda}_k^2}{\lambda_k}} \overset{d}{=} \frac{\sum_{k=1}^{K} Y_k \, \frac{\dot{\lambda}_k}{\sqrt{\lambda_k}}}{\sum_{k=1}^{K} \frac{\dot{\lambda}_k^2}{\lambda_k}} \tag{8.42}$$

Notons

$$\underline{b}^* = \left(\frac{\dot{\lambda}_1}{\sqrt{\lambda_1}}, \ldots, \frac{\dot{\lambda}_K}{\sqrt{\lambda_K}} \right)^T \tag{8.43}$$

$$\begin{aligned} \overline{I}_0 &= \|\underline{b}^*\|^2 = \sum_{k=1}^{K} \mu_0(J_k) \left[\frac{\partial}{\partial \theta} \ln \mu_\theta(J_k) \Big|_{\theta=\theta_0} \right]^2 \\ &= -\sum_{k=1}^{K} \mu_0(J_k) \left[\frac{\partial}{\partial \theta^2} \ln \mu_\theta(J_k) \Big|_{\theta=\theta_0} \right] \end{aligned} \tag{8.44}$$

que l'on suppose > 0. On trouve successivement :

$$\forall 1 \leq k \leq K, \quad \overline{\mathbb{Z}}(J_k) = \mathbb{Z}(J_k) - \frac{\langle \underline{Y}, \underline{b}^* \rangle}{\langle \underline{b}^*, \underline{b}^* \rangle} \dot{\lambda}_k, \tag{8.45}$$

d'où

$$\overline{\underline{Y}} = \underline{Y} - \frac{\langle \underline{Y}, \underline{b}^* \rangle}{\langle \underline{b}^*, \underline{b}^* \rangle} \underline{b}^* = \underline{Y} - \langle \underline{Y}, \underline{b}_0 \rangle \, \underline{b}_0 \tag{8.46}$$

avec $\|\underline{b}_0\| = 1$: $\overline{\underline{Y}}$ est donc le projeté orthogonal de $\underline{Y}$ sur l'hyperplan $\underline{b}_0^\perp$ orthogonal à $\underline{b}_0$. Or, on a vu (théorème 8.1) que $\underline{Y}$ pouvait s'écrire sous la forme :

$$\underline{Y} = \sum_{j=1}^{K-1} \langle \underline{Y}, \underline{u}_j \rangle \, \underline{u}_j + \langle \underline{Y}, \underline{a} \rangle \, \underline{a} \tag{8.47}$$

avec $\|\underline{a}\| = 1$, $\underline{a} \perp \underline{b}_0$ (voir théorème 8.2), $(\underline{u}_1, \ldots, \underline{u}_{K-1})$ base orthonormée quelconque de l'hyperplan $\underline{a}^T$, $\langle \underline{Y}, a \rangle = 0$ p.s., et $\xi_j = \langle \underline{Y}, \underline{u}_j \rangle$ i.i.d. $\mathcal{N}(0,1)$, $1 \leq j \leq K-1$. Pour conclure, après avoir supprimé le terme $\langle \underline{Y}, \underline{a} \rangle \, \underline{a}$ (puisqu'il est nul p.s.), il suffit de choisir $\underline{u}_{K-1} = \underline{b}_0$ et de compléter la base orthonormée $(\underline{u}_1, \ldots, \underline{u}_{K-2})$ de $\underline{a}^\perp$ à partir de là. Il reste bien alors

$$\overline{\underline{Y}} = \sum_{j=1}^{K-2} \xi_j \underline{u}_j, \quad \xi_j \text{ i.i.d.} \sim \mathcal{N}(0,\, 1), \tag{8.48}$$

donc

$$\left\| \overline{\underline{Y}} \right\|^2 = \sum_{j=1}^{K-2} \xi_j^2 \tag{8.49}$$

■

8.4 Commentaires pratiques

Le test d'adéquation du chi-deux proposé en 8.3 (Pearson-Fisher) est commode, en général simple à utiliser, à condition de respecter un certain nombre de règles, dont les plus importantes concernent le choix de K (nombre de parties de la partition). Ce choix doit être fait de manière à ce que :

– K soit grand devant la dimension q de Θ ;

– K soit petit devant n.

En particulier, chaque J_k, $1 \leq k \leq K$, devrait toujours contenir au moins 5 points de l'échantillon, sans quoi la loi de $\overline{\mathcal{T}}_n^2$ est trop loin de la loi limite de $\overline{\mathcal{T}}^2(\chi^2_{K-q-1})$, à partir de laquelle la région de rejet est déterminée en fonction de α, $0 < \alpha < 1$.

Ces considérations ont conduit des statisticiens à modifier le test afin d'introduire des parties J_k aléatoires, dépendant de l'échantillon $X_1, \ldots, X_n$, afin d'essayer, par exemple, de choisir une partition dont les parties contiennent un même nombre de points de l'échantillon. Cela modifie en général la loi limite de la statistique de test, et nous ne développerons pas davantage ce point de vue, souvent conseillé en pratique, car conduisant à de meilleures approximations (à n fini ou modéré) et à une puissance accrue.

D'autres améliorations ont été – et sont encore – proposées. Par exemple, des statisticiens ont introduit des formes quadratiques aléatoires (car dépendant en général de l'échantillon) plus générales que $\mathcal{T}_n^2$: notons

$$\underline{Y}_n(\theta) = \left(\frac{n_k - np_k(\theta)}{\sqrt{np_k(\theta)}},\ 1 \leq k \leq K \right)^T \tag{8.50}$$

(c'est un vecteur colonne). Soit $\tilde{\theta}_n$ un estimateur bien choisi de θ, et

$$Q_n = Q_n(X_1, \ldots, X_n)$$

une matrice $K \times K$ symétrique positive. On considère la statistique de test associée

$$\tilde{\mathcal{T}}_n^2(Q) = \underline{Y}_n(\tilde{\theta}_n)^T\, Q_n\, \underline{Y}_n(\tilde{\theta}_n) \tag{8.51}$$

En choisissant astucieusement Q_n (typiquement en termes de matrices d'information de Fisher), il est possible de construire ainsi des statistiques de test dont la loi limite est une loi χ^2 et qui conduisent à des tests plus puissants.

Cependant, les tests de type chi-deux tendent à rester moins puissants que les tests basés sur la fonction de répartition empirique dans le cas de données réelles continues, tests dont nous allons parler au chapitre 9.

Ils restent pourtant intéressants, particulièrement pour les observations discrètes (par exemple, lois de Poisson) ou multidimensionnelles, en raison de leur simplicité, ou encore pour tester l'adéquation de modèles peu courants, contenant d'autres paramètres que les paramètres de position et d'échelle, et pour lesquels les tests du chapitre 9 n'ont pas été tabulés.

8.5 Test d'indépendance

Le test du chi-deux peut être utilisé pour tester l'indépendance de deux variables aléatoires X et Y. Supposons que X et Y prennent toutes deux un nombre fini de valeurs : X appartient à l'ensemble I de cardinal K, d'éléments notés i, tandis que Y appartient à l'ensemble J de cardinal L, d'éléments notés j. Notons $p_i = P\{X = i\}$ et $q_j = P\{Y = j\}$. On veut tester

l'hypothèse nulle H_0 [X et Y sont indépendantes]. Si l'on note $r_{i,j} = P\{X = i, Y = j\}$, cela revient à tester l'hypothèse nulle H_0 [$\forall i,\ j,\ r_{i,j} = p_i\, q_j$].

Exemple *Une partie d'une population a été vaccinée, l'autre non. On veut déterminer si le vaccin est efficace ou non. Pour cela, on teste l'indépendance de la variable X, représentant les modalités Vacciné (1) ou Non Vacciné (2), et de la variable Y, représentant les modalités Contaminé (1) ou Non Contaminé (2). Si l'indépendance de X et de Y ne peut pas être rejetée au niveau de signification asymptotique α retenu (le choix de ce niveau doit être fixé avec soin, avec les expérimentateurs), alors on ne peut pas, au vu de l'expérience effectuée, conclure que le vaccin soit efficace. Supposons que la population totale sur laquelle a porté l'expérience comporte $N = 274$ individus, avec 3 vaccinés contaminés $(1,1)$, 10 non vaccinés contaminés $(2,1)$,, 144 vaccinés non contaminés $(1,2)$ et 117 non vaccinés non contaminés $(2,2)$. Il sera commode de noter cela $n_{1,1} = 3$, $n_{2,1} = 10$, $n_{1,2} = 144$ et $n_{2,2} = 117$. Supposons que le niveau choisi a été $\alpha = 1$ %. Que conclure ?*

Le problème posé s'interprète comme un test d'adéquation. Il s'agit de tester

$$H_0\,[\forall i,\ j,\ \ r_{i,j} = p_i\, q_j],$$

mais sans connaître les p_i ni les q_j. Notons $N_{i,.} = \sum_{j\in J} n_{i,j}$ et $N_{.,j} = \sum_{i\in I} n_{i,j}$, avec $N = \sum_{j\in J} N_{.,j} = \sum_{i\in I} N_{i,.}$. On estime, sous H_0, p_i par $\widehat{p}_i = N_{i,.}/N$ et q_j par $\widehat{q}_j = N_{.,j}/N$. Comme la somme des $\widehat{p}_i$ est égale à 1 de même que celle des $\widehat{q}_j$, on estime en fait $K+L-2$ paramètres indépendants, à partir desquels on obtient, sous H_0, les estimateurs $\widehat{r}_{i,j} = N_{i,.}N_{.,j}/N^2$ des $r_{i,j}$. Pour chaque couple $(i,\ j)$, l'effectif observé est $n_{i,j}$ et l'effectif théorique attendu sous H_0 est $N\widehat{r}_{i,j} = N_{i,.}N_{.,j}/N$.

La statistique de test est alors

$$\sum_{(i,\,j)\in I\times J} \frac{(n_{i,j} \ - \ N_{i,.}N_{.,j}/N)^2}{N_{i,.}N_{.,j}/N}$$

Elle suit approximativement une loi du chi-deux à $KL - (K+L-2) - 1 = (K-1)(L-1)$ degrés de liberté.

Exemple (suite) *On obtient ici $N_{1,.} = 147$, $N_{2,.} = 127$, $N_{.,1} = 13$ et $N_{.,2} = 261$, avec $N = 274 = 147 + 127 = 261 + 13$. La statistique de test vaut approximativement*

$$\frac{(3 \ - \ 7,0)^2}{7,0} + \frac{(144 \ - \ 140,0)^2}{140,0} + \frac{(10 \ - \ 6,0)^2}{6,0} + \frac{(117 \ - \ 121,0)^2}{121,0} \approx 2,3 + 0,1 + 2,7 + 0,1 \approx 5,2$$

Le quantile d'ordre $0,99$ de la loi χ^2_1 vaut $6,63$.

Au niveau retenu, on ne peut pas conclure nettement que le vaccin soit efficace. Dans un tel cas, on ne pourrait que recommander de nouvelles expériences portant sur des effectifs supérieurs, tout en faisant remarquer que de toutes manières l'efficacité du vaccin semble douteuse.

Il est important de bien comprendre que cette utilisation du test du chi-deux comme test d'indépendance peut aussi conduire à un test d'homogénéité, c'est-à-dire à tester l'hypothèse nulle que plusieurs échantillons indépendants formés chacun de réalisations i.i.d. de v.a. pouvant prendre K valeurs distinctes sont en fait de même loi. En effet, à partir du tableau

des $n_{i,\,j}$ on peut construire les fréquences conditionnelles

$$f_{i|j} = n_{i,\,j}/N_{\cdot,\,j} \quad \text{et} \quad g_{j|i} = n_{i,\,j}/N_{i,\,\cdot}$$

On peut considérer qu'on dispose soit de L échantillons indépendants de tailles $N_{\cdot,\,j}$, $j \in J$, de réalisations i.i.d. de variables aléatoires de type X, soit de K échantillons indépendants de tailles $N_{i,\,\cdot}$, $i \in I$, de réalisations i.i.d. de variables aléatoires de type Y. Dans le premier cas on cherche à tester l'hypothèse nulle que les vecteurs de fréquences conditionnelles

$$\mathbf{f}_{|j} = \big(f_{i|j} : i \in I\big)$$

correspondent tous, lorsque j décrit J, à la même loi

$$\mathbf{p} = (p_i : i \in I)\,,$$

qui serait alors estimée par les $\widehat{p}_i = N_{i,\,\cdot}/N$, $i \in I$. Dans le second cas, on cherche à tester l'hypothèse nulle que les vecteurs de fréquences conditionnelles

$$\mathbf{g}_{|i} = \big(g_{j|i} : j \in J\big)$$

correspondent tous, lorsque i décrit I, à la même loi

$$\mathbf{q} = (q_j : j \in J)\,,$$

qui serait alors estimée par les $\widehat{q}_j = N_{\cdot,\,j}/N$, $j \in J$.

Exemple (suite) *Ici,* $n_{1,\,1} = 3$, $n_{2,\,1} = 10$, $n_{1,\,2} = 144$ et $n_{2,\,2} = 117$, avec $N_{1,\,\cdot} = 147$, $N_{2,\,\cdot} = 127$, $N_{\cdot,\,1} = 13$ et $N_{\cdot,\,2} = 261$, *donc, si on ne considère que les* $\mathbf{f}_{|j}$ *(les calculs sont analogues pour les* $\mathbf{g}_{|i}$ *) :*

$$\mathbf{f}_{|1} = (3/13,\, 10/13) \approx (0.23,\, 0.77) \quad \text{et} \quad \mathbf{f}_{|2} = (144/261,\, 117/261) \approx (0.55,\, 0.45)\,,$$

tandis que

$$\widehat{\mathbf{p}} = (147/274,\, 127/274) \approx (0.53,\, 0.47)$$

Le vecteur $\mathbf{f}_{|1}$ *est très différent du vecteur* $\widehat{\mathbf{p}}$, *mais la contribution de leur écart à la statistique de test est petite car l'effectif correspondant,* $N_{\cdot,\,1} = 13$, *est faible. Intuitivement, elle porte donc peu d'information. Cela peut se lire sur la formule ci-dessus de la statistique de test du chi-deux. Comme on l'a vu plus haut, au niveau asymptotique* $\alpha = 1$ %, *on ne peut pas conclure ici que les lois conditionnelles de* X *sachant* $j = 1$ *et* $j = 2$ *soient significativement différentes. Au niveau asymptotique* $\alpha = 2.5$ %, *ces lois conditionnelles ne sont pas nettement significativement différentes. Par contre, au niveau asymptotique* $\alpha = 5$ %, *elles sont significativement différentes. Remarquons que le test du chi-deux utilisé comme test d'homogénéité n'est pas un test de puissance maximale. Il est possible que malgré la petitesse relative de* $N_{\cdot,\,1}$ *ici, la différence entre* $\mathbf{f}_{|1}$ *et* $\mathbf{f}_{|2}$ *soit assez marquée pour que le test d'homogénéité UPP* α*-semblable correspondant à la situation (deux échantillons indépendants de variables indépendantes de Bernoulli, l'un de taille* 13 *et l'autre de taille* 261 *) rejette l'hypothèse nulle d'homogénéité dès le niveau* $\alpha = 1$ %.

Pour finir, remarquons qu'on peut en déduire un test d'homogénéité pour plusieurs échantillons indépendants, considérés comme provenant de la réalisation de variables aléatoires issues du même modèle paramétrique de lois. Il suffit pour cela de fabriquer une partition finie de l'espace des valeurs et de comptabiliser pour chaque échantillon le nombre d'observations qui appartiennent à chacune des boîtes. Cependant, un tel test n'aurait qu'une puissance médiocre. Mieux vaut, lorsque c'est possible, avoir recours à des tests d'homogénéité UPP α-semblables construits spécialement pour la famille paramétrée de lois considérée.

Exercice 8.1 *On reprend la définition de la mesure empirique (voir (8.4) – (8.6)) μ_n et du processus empirique associé*

$$\mathbb{Z}_n = \sqrt{n}\,[\mu_n - \mu]\,. \qquad \left(\text{On note } \mu(\varphi) = \int \varphi\, d\mu.\right)$$

1. *Montrer que si φ et $\psi \in L^2(\mu)$, on a :*
   (a) $\mathbb{E}\,[\mathbb{Z}_n(\varphi)] = \mathbb{E}\,[\mathbb{Z}_n(\psi)] = 0$.
   (b) $\mathbb{E}\,[\mathbb{Z}_n(\varphi)\mathbb{Z}_n(\psi)] = \mu(\varphi\psi) - \mu(\varphi)\mu(\psi)$.
   (c)
   $$\begin{bmatrix} \mathbb{Z}_n(\varphi) \\ \mathbb{Z}_n(\psi) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mu(\varphi^2) - \mu^2(\varphi) & \mu(\varphi\psi) - \mu(\varphi)\mu(\psi) \\ \mu(\varphi\psi) - \mu(\varphi)\mu(\psi) & \mu(\psi^2) - \mu^2(\psi) \end{bmatrix}\right)$$
2. *Montrer que, si $(u_1, \ldots, u_K)$ est un système fini ($K \geq 1$) orthonormé de $L^2_0(\mu)$, c'est-à-dire des éléments de $L^2(\mu)$ dont l'intégrale par rapport à μ est nulle (donc $\mu(u_1) = \cdots = \mu(u_K) = 0$), alors :*
   $$\underline{Y}_n = (\mathbb{Z}_n(u_1), \ldots, \mathbb{Z}_n(u_K))^T \xrightarrow{d} \mathcal{N}(0,\, I_K)$$
3. *On considère un système orthonormé $(u_1, \ldots, u_K)$ fixé de $L^2_0(\mu)$ comme ci-dessus, pour $\mu = \mu_0$ définie comme suit.*

Soit Θ un intervalle de $\mathbb{R}$, $\theta \in \Theta \mapsto \mu_\theta$ une famille à un paramètre de mesures de probabilité sur $(E, \mathcal{E})$. On souhaite tester l'adéquation de ce modèle à un échantillon $X_1, \ldots, X_n$ i.i.d. donné. L'hypothèse nulle est donc qu'il existe θ_0 (unique car on suppose le modèle identifiable) tel que les v.a. X_i $(1 \leq i \leq n)$ soient de loi $\mu_0 = \mu_{\theta_0}$. En pratique, ce paramètre θ_0 serait inconnu, donc à estimer sous l'hypothèse H_0.

On fait les hypothèses suivantes :

(A1) $\forall \varphi \in L^2(\mu_0)$ la fonction $\theta \to \mu_\theta(\varphi)$ est C^2 et bornée, ainsi que ses dérivées première et seconde, sur un intervalle $\overline{B}(\theta_0, r_0) = [\theta_0 - r_0, \theta_0 + r_0]$ supposé contenu dans Θ, avec $r_0 > 0$. On notera :

$$\dot{\mu}_\theta(\varphi) = \frac{\mathrm{d}}{\mathrm{d}\theta}\mu_\theta(\varphi), \quad \dot{\mu}_0(\varphi) = \dot{\mu}_{\theta_0}(\varphi)$$
$$\ddot{\mu}_\theta(\varphi) = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\mu_\theta(\varphi), \ddot{\mu}_0(\varphi) = \ddot{\mu}_{\theta_0}(\varphi)$$

$$m_\theta = (\mu_\theta(u_1), \ldots, \mu_\theta(u_K))^T, \quad m_0 = m_{\theta_0}$$
$$\dot{m}_\theta = \frac{\mathrm{d}m_\theta}{\mathrm{d}\theta} = (\dot{\mu}_\theta(u_1), \ldots, \dot{\mu}_\theta(u_K))^T, \quad \dot{m}_0 = \dot{m}_{\theta_0}$$
$$\ddot{m}_\theta = \frac{\mathrm{d}^2 m_\theta}{\mathrm{d}\theta} = (\ddot{\mu}_\theta(u_1), \ldots, \ddot{\mu}_\theta(u_K))^T, \quad \ddot{m}_0 = \ddot{m}_{\theta_0},$$

et $\langle \underline{a}, \underline{b} \rangle_K = \sum_{j=1}^K a_j b_j$ *le produit scalaire canonique de* $\mathbb{R}^K$.

(A2) $\forall \theta \in \Theta, \quad \dot{m}_\theta \neq 0$.

On définit l'estimateur $\widehat{\theta}_n$ *de* θ_0 *sous* H_0 *comme la valeur de* θ*, supposée bien définie et unique, qui minimise le carré de la distance euclidienne*

$$\|\underline{m}_n - m_\theta\|_K^2 = \sum_{j=1}^K (\mu_n(u_j) - \mu_\theta(u_j))^2, \quad \theta \in \Theta,$$

entre le point $\underline{m}_n = (\mu_n(u_1), \ldots, \mu_n(u_K))^T$ *et la courbe* $\theta \mapsto m_\theta$ *de* $\mathbb{R}^K$*. Rappelons que sous* H_0, $\underline{m}_n \xrightarrow{\text{p.s.}} m_0$.

On admet :

(A3) Sous H_0, $\widehat{\theta}_n \xrightarrow{\text{p.s.}} \theta_0$ *lorsque* $n \to \infty$.

1. *(a) Montrer que l'on a*

$$\langle \underline{m}_n - m_{\widehat{\theta}_n}, \dot{m}_{\widehat{\theta}_n} \rangle_K = 0$$

et

$$\left| \langle \underline{m}_n - m_{\widehat{\theta}_n}, \ddot{m}_{\widehat{\theta}_n} \rangle_K \right| \leq \|\dot{m}_{\widehat{\theta}_n}\|_K^2$$

(b) On note $\widehat{\mathbb{Z}}_n = \sqrt{n}\left[\mu_n - \mu_{\widehat{\theta}_n}\right]$ *et* $\underline{\widehat{Y}}_n = (\widehat{\mathbb{Z}}_n(u_1), \ldots, \widehat{\mathbb{Z}}_n(u_K))^T$.
Montrer que sous H_0*, on a :*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{\langle \underline{Y}_n, \dot{m}_{\widehat{\theta}_n} \rangle_K}{\|\dot{m}_{\widehat{\theta}_n}\|_K^2} + o_P(1)$$

En déduire que sous H_0,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\|\dot{m}_0\|_K^2}\right)$$

(c) On note $\underline{Y} = (\mathbb{Z}(u_1), \ldots, \mathbb{Z}(u_K))^T$ *une v.a. limite en distribution de* $\underline{Y}_n (n \to \infty)$.
Montrer que sous H_0,

$$\underline{\widehat{Y}}_n \xrightarrow{d} \underline{Y} - \frac{\langle \underline{Y}, \dot{m}_0 \rangle_K}{\|\dot{m}_0\|_K^2} \quad \stackrel{def}{=} \underline{\widehat{Y}}$$

En déduire la loi de $\underline{\widehat{Y}}$ *sous* H_0.

2. *On suppose que* $d\mu_\theta = f_\theta \, d\nu$, f_θ *densité de* μ_θ *par rapport à la mesure de référence* ν. *Toujours sous les hypothèses* $(A1)-(A3)$*, plus des hypothèses d'intégrabilité à préciser, montrer que*

$$\|\dot{m}_0\|_K^2 \leq \int \left(\frac{\dot{f}_0}{f_0}\right)^2 f_0 \, d\nu = I_0 \quad \text{(informationdeFisher)}$$

Que se passerait-il si on faisait tendre K *vers l'infini ?*

Conclusion

- Caractère simplement asymptotique ($n \to +\infty$).
- Aucune garantie d'optimalité.
- Notion de test d'adéquation.
- Importance de la prise en compte des estimateurs utilisés pour le modèle sous H_0.
- Mesure empirique et processus empirique.
- Attention aux limitations sur le nombre minimal d'observations par "boîte" si on veut pouvoir utiliser les lois asymptotiques avec une approximation correcte.
- Tester l'indépendance, c'est-à-dire l'homogénéité de plusieurs échantillons répartis dans des "boîtes".

Eléments bibliographiques pour ce chapitre

Association pour la Statistique et ses Utilisations (1996) *Inférence non paramétrique. Les statistiques de rangs*, Jean-Jacques Droesbeke et Jeanne Fine éditeurs, Collection "Ellipses", Editions de l'Université Libre de Bruxelles : Bruxelles.

Capéraà, P. & Van Cutsem, B. (1988) *Méthodes et modèles en Statistique non paramétrique*. Dunod : Paris.

Conover, W. J. (1971) *Practical Nonparametric Statistics*. Wiley : New York.

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Dagnelie P. (1975) *Théorie et méthodes statistiques*, Tome 2. Vander-Oyez : Bruxelles, 2e édition.

D'Agostino, R. B. & Stephens, M. A. (1986) *Goodness-of-Fit Techniques*, Statistics, textbooks and monographs **68**. Marcel Dekker : New York and Basel.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Lecoutre, J.-P. & Tassi, P. (1987) *Statistique non paramétrique et robustesse*. Collection "Economie et statistiques avancées", Economica : Paris.

Monfort, A. (1982) *Cours de statistique mathématique*. Collection "Economie et statistiques avancées", Economica : Paris.

Rayner, J. C. W. & Best, D. J. (1989) *Smooth Tests of Goodness-of-Fit*. Oxford University Press : Oxford.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley : New York.

Tassi, P. (1985) *Méthodes statistiques*. Collection "Economie et statistiques avancées", Economica : Paris.

Ulmo, J. & Bernier, J. (1973) *Eléments de décision statistique*. Presses Universitaires de France : Paris.

Zuber, J. (1997) *Un test chi-carré d'adéquation de modèles paramétriques en régression*. Thèse, Ecole Polytechnique Fédérale de Lausanne, Suisse.

9

Introduction aux tests d'adéquation utilisant le processus empirique

9.1 Le processus empirique uniforme

Les v.a. considérées ici seront à valeurs réelles. A toute v.a. réelle X de loi μ (μ mesure de probabilité sur $\mathbb{R}$) on associe sa fonction de répartition F définie par :

$$\forall x \in \mathbb{R}, \quad F(x) = \mu\left(]-\infty, x]\right) = \mathbb{P}\left\{X \leq x\right\} \tag{9.1}$$

Cette fonction de répartition (f.r.) est :
- Croissante de 0 à 1 ;
- Elle n'admet donc au plus qu'une infinité dénombrable de points de discontinuité, qui sont tous de première espèce (penser par exemple au cas où $X \sim \text{Poisson}(\lambda)$) ;
- En chacun de ces points de discontinuité éventuels elle est continue à droite, c'est-à-dire que si x_0 est un tel point,

$$F(x_0) = \lim_{x \to x_0+} F(x) \tag{9.2}$$

On notera G la f.r. des v.a. uniformes sur $[0, 1]$ ($\mathcal{U}[0, 1]$)

$$G(x) = \begin{cases} 0 \text{ pour } x \leq 0 \\ x \text{ pour } 0 \leq x \leq 1 \\ 1 \text{ pour } x \geq 1 \end{cases} \tag{9.3}$$

Par un abus de notation commode, on notera aussi G la fonction définie sur $[0, 1]$ par

$$\forall t \in [0, 1], \quad G(t) = t \tag{9.4}$$

Dans la suite du chapitre 9, *nous supposerons F continue.*

Lemme 9.1 *Si X est une v.a. réelle de f.r. F continue sur $\mathbb{R}$, alors $F(X)$ suit une loi uniforme sur $[0, 1]$.*

De plus, étant donnée une v.a. $U \sim \mathcal{U}[0, 1]$, on peut reconstruire une v.a. Y de f.r. F à

l'aide de la fonction réciproque généralisée $F^{\leftarrow}$ de F :

$$Y = F^{\leftarrow}(U) \tag{9.5}$$

avec

$$F^{\leftarrow}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\} \tag{9.6}$$

Bien entendu, si F est continue et strictement croissante d'un intervalle J de $\mathbb{R}$ sur $(0, 1)$ (bornes non précisées), alors son inverse généralisée $F^{\leftarrow}$ coïncide sur J avec $F^{-1} : (0,1) \to J$, la bijection continue réciproque. C'est le cas qui va nous intéresser dans la suite, en général.

Ainsi, pour les v.a. réelles dont la f.r. est continue, on peut se ramener aux résultats concernant les v.a. $\mathcal{U}[0, 1]$, puis revenir en arrière. Nous allons exploiter ce principe fondamental.

Définition 9.1 *Si $X_1, \ldots, X_n$ sont des v.a. de f.r. F sur $\mathbb{R}$, leur* ***fonction de répartition empirique*** *$\mathbb{F}_n$ est définie par*

$$\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) = \mu_n([-\infty, x]) \tag{9.7}$$

Dans le cas où $X_1, \ldots, X_n$ sont $\mathcal{U}[0, 1]$, on notera $\mathbb{G}_n$ leur f.r. empirique.
Si F est continue, si $X_1, \ldots, X_n \sim F$, si $U_i = F(X_i)$, $1 \leq i \leq n$, on peut écrire :

$$\forall x, \quad \mathbb{G}_n[F(x)] = \mathbb{F}_n(x)$$

Lemme 9.2 *Si $X_1, \ldots, X_n$ sont i.i.d. $\sim F$, alors :*

1. *Pour tout x, $\mathbb{F}_n(x) \xrightarrow{\text{p.s.}} F(x)$ quand $n \to +\infty$;*
2. *Pour tout x, $\sqrt{n}(\mathbb{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}[0, F(x)(1 - F(x))]$.*

Théorème 9.1 (Glivenko-Cantelli) *Si $X_1, \ldots, X_n$ sont i.i.d. $\sim F$, alors*

$$\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \xrightarrow{\text{p.s.}} 0 \qquad \text{quand } n \to +\infty$$

Exercice 9.1

1. *Pour $0 < t < 1$, montrer que $\{x \in \mathbb{R} : F(x) \geq t\}$ est un intervalle fermé à gauche, donc de la forme*
$$[F^{\leftarrow}(t), +\infty[$$
(Utiliser la croissance de F et sa continuité à droite.)
Que peut-il se passer pour $t = 0$ ou $t = 1$?
2. *En déduire que $F(F^{\leftarrow}(t)) \geq t$ pour tout $t \in [0, 1]$.*
3. *Montrer que $F^{\leftarrow}(F(x)) \leq x$ pour tout x.*
4. *Quand a-t-on égalité dans 2. et 3. ?*

Exercice 9.2 *Soit X une v.a. de type Bernoulli, prenant la valeur $a \in \mathbb{R}$ avec probabilité p $(0 < p < 1)$ et la valeur $b > a$, $b \in \mathbb{R}$, avec probabilité $q = 1 - p$.*

1. *Déterminer et représenter la f.r. F de X.*
2. *Même chose pour $F^{\leftarrow}$.*
3. *Vérifier dans ce cas les résultats de l'exercice 9.1.*
4. *Soit $U \sim \mathcal{U}[0, 1]$: construire, sous la forme d'un algorithme simple, la v.a. $Y = F^{\leftarrow}(U)$. Vérifier que $Y \stackrel{d}{=} X$. Déterminer la loi de $F(Y)$: retrouve-t-on U ?*
5. *Généraliser les résultats obtenus au cas d'une v.a. X prenant les valeurs $a_1 < a_2 < \ldots < a_K$ avec probabilités $p_1 > 0$, $p_2 > 0, \ldots, p_K > 0$, $\sum_1^K p_k = 1$.*

Exercice 9.3 *Démontrer que $Y = F^{\leftarrow}(U)$, $U \sim \mathcal{U}[0, 1]$, admet F pour f.r.*

Exercice 9.4 *Reprenant l'Exercice 9.1, montrer que si F est continue, alors la v.a. transformée $F(X)$ de la v.a. X de f.r. F est une v.a. uniforme $\mathcal{U}[0, 1]$.*

Exercice 9.5

1. *Soit $U_1, \ldots, U_n$ un échantillon de v.a. i.i.d. $\mathcal{U}[0, 1]$, et soit $Y_i = F^{\leftarrow}(U_i)$. En utilisant l'Exercice 9.3, montrer que*

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x) &= \frac{1}{n}\sum_{i=1}^{n} I(U_i \leq F(x)) \\ &= \mathbb{G}_n(F(x)) \qquad \text{pour tout } x \end{aligned}$$

2. *Soit $X_1, \ldots, X_n$ un échantillon de v.a. i.i.d. de f.r. F continue. Soit $U_i = F(X_i)$. D'après l'Exercice 9.4, les U_i sont $\mathcal{U}[0, 1]$. Soit $Y_i = F^{\leftarrow}(U_i)$. D'après 1, on a*

$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x) = \mathbb{G}_n(F(x)) \qquad \text{pour tout } x$$

En déduire que pour tout x, on a l'égalité en loi :

$$\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \stackrel{d}{=} \mathbb{G}_n(F(x))$$

Etendre ce résultat en montrant que pour tout $K \geq 1$ et tous $x_1 < \ldots < x_K$, on a

$$(\mathbb{F}_n(x_1), \ldots, \mathbb{F}_n(x_K)) \stackrel{d}{=} (\mathbb{G}_n(F(x_1)), \ldots, \mathbb{G}_n(F(x_K)))$$

Exercice 9.6

1. *En utilisant l'Exercice 9.5, expliquer en quel sens précis on peut ramener, dans le cas où F est continue, la démonstration du théorème de Glivenko-Cantelli à la démonstration du résultat plus simple suivant :*
"Si $U_1, \ldots, U_n$ sont i.i.d. $\sim \mathcal{U}[0, 1]$, si on note $G(t) = t$ sur $[0, 1]$ et

$$\mathbb{G}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(U_i \leq t) \ ,$$

alors

$$\sup_{0 \leq t \leq 1} |\mathbb{G}_n(t) - G(t)| \xrightarrow{\text{p.s.}} 0 \quad \text{quand } n \to +\infty$$

2. Essayer d'établir ce dernier résultat.

Toujours à partir du TLC, on peut montrer aussi que pour tout entier $K \geq 1$ et toute famille $x_1 < x_2 < \ldots < x_K$ de K nombres réels classés par ordre croissant, le vecteur aléatoire

$$\mathbb{Z}_n = \left(\sqrt{n}\left[\mathbb{F}_n(x_k) - F(x_k)\right] \ : \ 1 \leq k \leq K\right)^T$$

converge en loi quand $n \to +\infty$ vers un vecteur aléatoire gaussien de $\mathbb{R}^K$, centré et de matrice variance Γ définie par :

$$\begin{cases} \Gamma_{kk} = F(x_k)\left[1 - F(x_k)\right] & 1 \leq k \leq K \\ \Gamma_{jk} = F(x_j) \wedge F(x_k) - F(x_j)\,F(x_k) & \text{si } \ j \neq k \end{cases} \tag{9.8}$$

($a \wedge b$ désigne le minimum de a et de b).

En particulier, si $0 \leq u_1 < u_2 < \ldots < u_K \leq 1$,

$$\left(\sqrt{n}\left[\mathbb{G}_n(u_k) - u_k\right] \ : \ 1 \leq k \leq K\right)^T$$

converge en loi vers un vecteur gaussien de $\mathbb{R}^K$, centré et de matrice variance Γ définie par :

$$\Gamma_{jk} = u_j \wedge u_k - u_j u_k, \qquad 1 \leq j, k \leq K\,, \tag{9.9}$$

avec en particulier $\Gamma_{kk} = u_k^2 - u_k = u_k\left[1 - u_k\right]$.

Les objets mathématiques $\mathbb{F}_n$, $\mathbb{G}_n$, $\sqrt{n}\left[\mathbb{F}_n - F\right]$, $\sqrt{n}\left[\mathbb{G}_n - G\right]$ sont des v.a. à valeurs dans un espace de fonctions (sur $\mathbb{R}$ ou $[0,\,1]$) : ce sont des **processus aléatoires**.

Définition 9.2 *1. Si $X_1, \ldots, X_n$ i.i.d. $\sim F$, $\sqrt{n}\left[\mathbb{F}_n - F\right]$ est le **processus empirique** basé sur $X_1, \ldots, X_n$.*

2. Dans le cas de v.a. $\mathcal{U}[0,\,1]$, on note

$$\mathbb{B}_n = \sqrt{n}\left[\mathbb{G}_n - G\right], \quad \textit{qui est défini sur } [0,\,1]\,,$$

*et on parle simplement de **processus empirique uniforme**.*

Si F est continue, si $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$, si $U_i = F(X_i)$, on peut écrire

$$\forall x, \quad \sqrt{n}\left[\mathbb{F}_n(x) - F(x)\right] = \mathbb{B}_n\left(F(x)\right)$$

On se ramènera donc systématiquement au processus empirique uniforme. (Voir exercice 9.5 ci-dessus pour plus de précisions.)

Il est possible de préciser le théorème de Glivenko-Cantelli :

Théorème 9.2 (Dvoretsky-Kiefer-Wolfowitz) *Si les v.a. $U_1, \ldots, U_n$ sont i.i.d. $\sim \mathcal{U}[0,\,1]$, alors pour tout entier $n \geq 1$ et pour tout $a > 0$,*

$$\mathbb{P}\left\{\sup_{t \in [0,\,1]} |\mathbb{B}_n(t)| \ > \ a\right\} \leq 4\sqrt{2}\,e^{-2a^2}$$

On peut démontrer le résultat suivant, qui est suggéré (mais pas prouvé) par (9.9) ci-dessus.

Théorème 9.3 *Soit $U_1, U_2, \ldots$ une suite de v.a. i.i.d. $\mathcal{U}[0, 1]$. La suite $(\mathbb{B}_n)_{n\geq 1}$ des processus empiriques uniformes converge en distribution vers un processus gaussien centré $\mathbb{B}$, de fonction de covariance*

$$(t_1, t_2) \in [0, 1] \times [0, 1] \longmapsto r_{\mathbb{B}}(t_1, t_2) = \mathbb{E}(\mathbb{B}(t_1)\mathbb{B}(t_2))$$

définie par

$$r_{\mathbb{B}}(t_1, t_2) = t_1 \wedge t_2 - t_1 t_2 \tag{9.10}$$

Définition 9.3 *Un tel processus s'appelle un* ***pont brownien*** *sur* $[0, 1]$.

Nous n'avons pas défini la convergence en distribution d'une suite de processus $(\mathbb{B}_n \xrightarrow{d} \mathbb{B})$ quand $n \to +\infty$, et nous ne le ferons pas dans ce chapitre. Voir Annexe D.

Nous dirons seulement que de ces résultats, on peut déduire les énoncés suivants :

Corollaire 4
1. $\sup_{0\leq t\leq 1} \mathbb{B}_n(t) \xrightarrow{d} \sup_{0\leq t\leq 1} \mathbb{B}(t)$
2. $\sup_{0\leq t\leq 1} |\mathbb{B}_n(t)| \xrightarrow{d} \sup_{0\leq t\leq 1} |\mathbb{B}(t)|$
3. $\int_0^1 \mathbb{B}_n^2(t)\, dt \xrightarrow{d} \int_0^1 \mathbb{B}^2(t)\, dt$

Il s'agit ici de la convergence en loi de suites de v.a. réelles. De plus, les limites dans (1)-(3) sont bien définies, car on peut montrer que $\mathbb{B}$ est p.s. une fonction continue sur $[0, 1]$.

Remarque 11 *Le pont brownien tire son nom de ce que :*
- *Il est relié au mouvement brownien, ou processus de Wiener, au sens suivant :*
  - *Si $\mathbb{W}$ est un processus de Wiener standard sur $\mathbb{R}^+$ (donc $\mathbb{W}(0) = 0$ p.s.), alors le processus*
  $$t \in [0, 1] \longmapsto \mathbb{W}(t) - t\mathbb{W}(1)$$
  est un pont brownien sur $[0, 1]$.
  - *Réciproquement, si $\mathbb{B}$ est un pont brownien sur $[0, 1]$ et si ξ est une v.a. $\mathcal{N}(0, 1)$ indépendante de $\mathbb{B}$, alors le processus*
  $$t \in [0, 1] \longmapsto \mathbb{B}(u) + u\xi$$
  est la restriction à $[0, 1]$ d'un processus de Wiener sur $\mathbb{R}^+$.
- $\mathbb{B}(0) = \mathbb{B}(1)$ *p.s., donc $\mathbb{B}$ est un "pont" sur $[0, 1]$!*

Tout cela se démontre par des calculs sur les fonctions de covariance, caractéristiques des processus gaussiens centrés.

Nous allons voir comment exploiter le corollaire du théorème 9.3, pourvu qu'on sache calculer la loi des v.a. limites (1)-(3). Pour cela, on utilisera les résultats suivants.

Exercice 9.7 *On rappelle qu'un processus de Wiener standard sur $\mathbb{R}^+$ est un processus gaussien $\mathbb{X}$ défini sur $\mathbb{R}^+$, tel que*

$$\begin{array}{ll} \mathbb{X}(0) = 0 & p.s. \\ \mathbb{E}(\mathbb{X}(t)) = 0 & \text{pour tout } t \in \mathbb{R} \\ \mathbb{E}\left(\mathbb{X}(s)\,\mathbb{X}(t)\right) = s \wedge t & \text{le minimum de } s \text{ et de } t \end{array}$$

pour tous s et $t \in \mathbb{R}^+$.

1. *Soit $\mathbb{W}$ un tel processus de Wiener, restreint à $[0,\,1]$. Calculer la fonction moyenne et la fonction de covariance du processus $\mathbb{Y}$ défini par*

$$\mathbb{Y}(t) = \mathbb{W}(t) - t\mathbb{W}(1), \quad t \in [0,\,1]$$

2. *Même question pour*

$$\mathbb{Z}(t) = (1-t)\mathbb{W}\left(\frac{t}{t-1}\right), \quad t \in [0,\,1[$$

3. *Soit $\mathbb{B}$ un pont brownien sur $[0,\,1]$. Calculer la fonction moyenne et la fonction de covariance des processus*

$$t \in [0,\,1] \longmapsto \mathbb{B}(t) + t\xi\,,$$

où ξ est une v.a. $\mathcal{N}(0,\,1)$ indépendante de $\mathbb{B}$, et de

$$x \in \mathbb{R}^+ \longmapsto (1+x)\,\mathbb{B}\left(\frac{x}{x+1}\right)$$

En déduire qu'il s'agit dans les deux cas d'un processus de Wiener.

Théorème 9.4 1. *Pour tout $a \geq 0$:*

$$\mathbb{P}\left(\sup_{t\in[0,\,1]} B(t) > a\right) = e^{-2a^2} \tag{9.11}$$

2. *Pour tout $a \geq 0$:*

$$\mathbb{P}\left(\sup_{t\in[0,\,1]} |B(t)| > a\right) = 2\sum_{k=1}^{\infty} (-1)^{k+1}\, e^{-2k^2a^2} \tag{9.12}$$

Remarque 12 *On montre que le membre de droite de (9.12) est équivalent à $2e^{-2a^2}$ quand $a \to +\infty$. On pourra le montrer à titre d'exercice.*

Théorème 9.5 (décomposition de Karhunen-Loève de $\mathbb{B}$)

$$\mathbb{B} \stackrel{d}{=} \sum_{j=1}^{+\infty} \frac{1}{j\pi}\,\xi_j b_j\,, \tag{9.13}$$

où les ξ_j sont i.i.d. $\mathcal{N}(0,\,1)$ $(j \geq 1)$ et

$$\forall t \in [0,\,1]\,, \qquad b_j(t) = \sqrt{2}\sin(j\pi t)\,, \tag{9.14}$$

les b_j formant une base orthonormée de l'espace de Hilbert $L^2_0\,[0,\,1]$ des fonctions de carré intégrable et d'intégrale nulle sur l'intervalle $[0,\,1]$.

Corollaire 5

$$\int_0^1 \mathbb{B}^2(t)\,dt \stackrel{d}{=} \sum_{j=1}^{\infty} \frac{1}{j^2\pi^2}\,\xi_j^2 \tag{9.15}$$

Ainsi, la loi de $\int_0^1 \mathbb{B}^2(t)\,dt$ est-elle une somme pondérée de v.a. indépendantes suivant chacune une loi du chi-deux à 1 degré de liberté.

9.2 Tester $H_0\,[F = F_0]$ contre $H_1\,[F \neq F_0]$, F_0 continue donnée

Soient $X_1, \ldots, X_n$ des v.a. i.i.d. $\sim F$, supposée continue, et soient

$$U_i = F(X_i), \quad 1 \leq i \leq n\,, \tag{9.16}$$

les v.a. i.i.d. $\mathcal{U}\,[0,\,1]$ associées. Sous H_0, $U_i = F_0(X_i)$. Par conséquent, tout revient à tester l'hypothèse nulle H_0' [les U_i, $1 \leq i \leq n$, sont uniformes sur $[0,\,1]$], sachant qu'elles sont i.i.d.

– La statistique de test de *Kolmogorov-Smirnov* est

$$D_n = \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F_0(x)| \stackrel{d}{=} \sup_{t \in [0,\,1]} |\mathbb{G}_n(t) - t| \quad \text{sous } H_0 \tag{9.17}$$

– La statistique de test de *Cramér-von Mises* est

$$W_n^2 = \int_{\mathbb{R}} n\,[\mathbb{F}_n(x) - F_0(x)]^2\,dF_0(x) = \int_0^1 \mathbb{B}_n^2(t)\,dt \quad \text{sous } H_0 \tag{9.18}$$

Des formules pratiques pour calculer D_n et W_n existent : si $U_{1,\,n} \leq \ldots \leq U_{n,\,n}$ représente l'échantillon ordonné associé à $\{U_i : 1 \leq i \leq n\}$, on peut montrer que :

$$\left\{ \begin{array}{l} D_n = \max\,(D_n^+,\,D_n^-) \quad \text{avec} \\ D_n^+ = \max_{1 \leq i \leq n}\,(i/n \,-\, U_{i,\,n}) \text{ et} \\ D_n^- = \max_{1 \leq i \leq n}\,(U_{i,\,n} \,-\, (i-1)/n) \end{array} \right. \tag{9.19}$$

$$W_n^2 \stackrel{d}{=} \sum_{i=1}^{n} \left[U_{i,n} - \frac{2i-1}{2n} \right]^2 + \frac{1}{12n} \tag{9.20}$$

D'après les théorèmes 9.3, 9.4, 9.5 et leurs corollaires, sous H_0,

$$\mathbb{P}\left[\sqrt{n}\, D_n > a\right] \longrightarrow 2 \sum_{k=1}^{\infty} (-1)^{k+1}\, e^{-2k^2a^2} \quad (a \geq 0) \tag{9.21}$$

et

$$\mathbb{P}\left[W_n^2 > a\right] \longrightarrow \mathbb{P}\left[\sum_{j=1}^{\infty} \frac{1}{j^2\pi^2}\, \xi_j^2 > a \right] \quad (a \geq 0)\,, \tag{9.22}$$

où les ξ_j sont i.i.d. $\mathcal{N}(0,\,1)$ $(j \geq 1)$. Les limites ont été tabulées. On peut donc former des tests de niveau asymptotique α, $0 < \alpha < 1$, en cherchant le $(1-\alpha)$-quantile $q_{1-\alpha}^{\mathrm{KS}}$ ou $q_{1-\alpha}^{\mathrm{CM}}$ de (9.21) ou (9.22), puis en rejetant H_0 si :

– Pour le test de Kolmogorov-Smirnov,

$$\mathrm{KS}_n = \sqrt{n} D_n > q_{1-\alpha}^{\mathrm{KS}} \tag{9.23}$$

– Pour le test de Cramér-von Mises,

$$\mathrm{CM}_n = W_n^2 > q_{1-\alpha}^{\mathrm{CM}} \tag{9.24}$$

Remarque 13 *A partir de (9.11), on peut aussi tester* H_0 $[F \leq F_0]$ *contre* H_1 $[F > F_1]$, *en utilisant* D_n^+. *Sous* H_0, $D_n^+ = \sup_{t\in[0,\,1]} (\mathbb{G}_n(t) - t)$.

9.3 Tester l'appartenance de F à un modèle

Soit $\mathcal{M}_0 = \{F_\theta \ : \theta \in \Theta\}$ le modèle à tester. On veut plus précisément tester H_0 $[F \in \mathcal{M}_0]$ contre H_1 $[F \notin \mathcal{M}_0]$. Sous H_0, il existe θ, que nous noterons θ_0, tel que $F = F_{\theta_0}$. Mais ce θ_0 est inconnu, et on ne peut que l'estimer. Nous utiliserons ici l'estimateur $\widehat{\theta}_n$ du MV, mais ce qui va suivre resterait vrai pour tout autre estimateur $\tilde{\theta}_n$ vérifiant sous H_0 :

$$\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_0(X_i) \,+\, o_P(1) \quad \text{quand } n \to +\infty$$

Nous remplaçons le processus empirique $\sqrt{n}\,[\mathbb{F}_n - F_0]$ par le processus $\sqrt{n}\left[\mathbb{F}_n - F_{\widehat{\theta}_n}\right]$. Pour $\widehat{\theta}_n$ proche de θ_0, on a approximativement

$$\begin{aligned} \sqrt{n}\left[\mathbb{F}_n - F_{\widehat{\theta}_n}\right] &= \sqrt{n}\,[\mathbb{F}_n - F_0] - \sqrt{n}\left[\mathbb{F}_{\widehat{\theta}_n} - F_0\right] \\ &\approx (\mathbb{B}_n \circ F_0) - \sqrt{n} \sum_{k=1}^{K} \left.\frac{\partial F_\theta}{\partial \theta_k}\right|_{\theta=\theta_0} (\widehat{\theta}_{n,k} - \widehat{\theta}_{0,k}) \end{aligned} \tag{9.25}$$

si $\theta = (\theta_1, \ldots, \theta_K)$, $\widehat{\theta}_n = (\widehat{\theta}_{n,1}, \ldots, \widehat{\theta}_{n,K})$, etc. On notera $\nabla_\theta F_\theta$ le gradient de $\theta \mapsto F_\theta$ au point θ : $\nabla_\theta F_\theta = \left(\frac{\partial F_\theta}{\partial \theta_1}, \ldots, \frac{\partial F_\theta}{\partial \theta_K}\right)^T$. Supposons que F_θ admette une densité f_θ par rapport à

la mesure de Lebesgue sur $\mathbb{R}$, ce pour chaque $\theta \in \Theta$, et que le support de f_θ ne dépende pas de θ. Supposons aussi vérifiées les conditions usuelles de régularité, plus quelques autres ! Comme $\mathbb{B}_n \xrightarrow{d} \mathbb{B}$ quand $n \to \infty$, on peut déduire de l'approximation (9.25) l'approximation suivante :

$$\sqrt{n}\left[\mathbb{F}_n - F_{\widehat{\theta}_n}\right] \approx \mathbb{B} \circ F_0 \; - \; (\nabla_0 F)^T \sqrt{n}\left[\widehat{\theta}_n - \theta_0\right] , \tag{9.26}$$

où $\nabla_0 F = \nabla_\theta F_\theta|_{\theta=\theta_0}$ est un vecteur colonne, ainsi que $\widehat{\theta}_n - \theta_0$.

D'autre part, sous H_0,

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_n - \theta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_0^{-1} s_0(X_i) \; + \; o_P(1) & (9.27) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_0^{-1} s_0\left[F_0^{-1}(U_i)\right] \; + \; o_P(1) & (9.28) \end{aligned}$$

Posons $h_0 = s_0 \circ F_0^{-1}$: c'est une fonction à valeurs dans $\mathbb{R}^K$ (vecteurs colonnes). D'autre part, notons

$$\sqrt{n}\left[\mathbb{F}_n - F_{\widehat{\theta}_n}\right] = \widehat{\mathbb{B}}_n \circ F_0 \tag{9.29}$$

afin de tout ramener à des v.a. $\mathcal{U}[0,\, 1]$. On peut réécrire (9.25) sous la forme

$$\widehat{\mathbb{B}}_n \approx \mathbb{B} \; - \; g_0^T \, I_0^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_0(U_i) \right) , \tag{9.30}$$

où $g_0 = (\nabla_0 F) \circ F_0^{-1}$ est une fonction à valeurs dans $\mathbb{R}^K$ (vecteurs colonnes), I_0^{-1} est une matrice symétrique $K \times K$ et $n^{-1/2} \sum_{i=1}^{n} h_0(U_i)$ est une v.a. à valeurs dans $\mathbb{R}^K$ (vecteurs colonnes), centrée, de matrice variance

$$I_n = \frac{1}{n} \sum_{i=1}^{n} h_0(U_i) h_0^T(U_i) \xrightarrow{\text{p.s.}} I_0 \quad \text{quand } n \to \infty , \tag{9.31}$$

et qui converge en loi vers une v.a. $\xi = (\xi_1, \ldots, \xi_K)^T$ qui suit une loi $\mathcal{N}(0,\, I_0)$. Ainsi, on peut *conjecturer* que

$$\widehat{\mathbb{B}}_n \xrightarrow{d} \widehat{\mathbb{B}} = \mathbb{B} \; - \; g_0^T I_0^{-1} \xi \quad \text{quand } n \to \infty , \tag{9.32}$$

$\widehat{\mathbb{B}}$ étant un processus gaussien défini sur $[0,\, 1]$. Mais le vecteur gaussien ξ n'est pas indépendant de $\mathbb{B}$. Il nous faut préciser la covariance de $\mathbb{B}(t)$ et de ξ pour chaque $t \in [0,\, 1]$. Pour cela, nous *suggérons* (seulement) le calcul suivant (la notation Id désigne la fonction identité, $\forall t$, $Id(t) = t$) :

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_0(U_i) &= \sqrt{n} \int_0^1 h_0 \, d\mathbb{G}_n \\ &= \sqrt{n} \left(\int_0^1 h_0 \, d\mathbb{G}_n - \int_0^1 h_0 \, dt \right) \\ &\quad \left(\text{car } \int_0^1 h_0 \, dt = \int_0^1 s_0 \, dF_0 = 0 \right) & (9.33) \end{aligned}$$

$$\begin{aligned} &= \int_0^1 h_0 \, d\mathbb{B}_n \\ &\quad \big(\text{car } \mathbb{B}_n = \sqrt{n}\,[\mathbb{G}_n - Id] \text{ sous } H_0\big) \\ &\quad \xrightarrow{d} \int_0^1 h_0 \, d\mathbb{B} \qquad \text{quand } n \text{ tend vers } +\infty\,, \end{aligned}$$

ceci parce que $\mathbb{B}_n \xrightarrow{d} \mathbb{B}$, mais ce passage nécessite une démonstration détaillée qui dépasse le cadre de ce cours.

L'intégrale $\int_0^1 h_0 \, d\mathbb{B}$ est une **intégrale stochastique** par rapport au pont brownien. Elle est bien définie pourvu que $\|h_0\| \in L^2\,[0,\,1]$, c'est-à-dire que l'intégrale

$$\int_0^1 \|h_0\|^2 \, dt = \int \|s_0\|^2 \, dF_0 = \sum_{k=1}^{K} \int s_{0,\,k}^2 \, dF_0$$

soit convergente, ce qui fait partie des hypothèses de régularité. En utilisant les représentations de la remarque, on peut démontrer que pour toutes fonctions $\ell_1 \in L^2\,[0,\,1]$ et $\ell_2 \in L^2\,[0,1]$,

$$\int_0^1 \ell_\alpha \, d\mathbb{B} \sim \mathcal{N}\left(0,\, \int_0^1 \ell_\alpha^2 \, dt - \left(\int_0^1 \ell_\alpha \, dt\right)^2\right) \tag{9.34}$$

pour $\alpha = 1,\, 2$ et que

$$\mathbb{E}\left(\int_0^1 \ell_1 \, d\mathbb{B} \, \int_0^1 \ell_2 \, d\mathbb{B}\right) = \int_0^1 \ell_1 \ell_2 \, dt - \left(\int_0^1 \ell_1 \, dt\right)\left(\int_0^1 \ell_2 \, dt\right) \tag{9.35}$$

Par conséquent, on a, pour tout $t \in [0,1]$:

$$\begin{aligned} \mathbb{E}\,(\mathbb{B}(t)\,\xi_k) &= \mathbb{E}\left(\int_0^1 \mathbb{1}_{[0,\,t]} \, d\mathbb{B} \, \int_0^1 h_{0,\,k} \, d\mathbb{B}\right) \\ &= \mathbb{E}\left(\int_0^1 \mathbb{1}_{[0,\,t]}\,(s) \;\; h_{0,\,k}\,(s) \; ds\right) \\ &= \int_0^t h_{0,\,k}\,(s) \; ds, \qquad 1 \le k \le K, \end{aligned} \tag{9.36}$$

car $\int_0^1 h_{0,\,k}(s)\, ds = 0$. On écrit cela de manière plus ramassée :

$$\mathbb{E}\,(\mathbb{B}(t)\,\xi) = \int_0^t h_0 \, ds \;\; \in \mathbb{R}^K \;\; \text{(vecteurs colonnes)} \tag{9.37}$$

On peut enfin en déduire la forme de la fonction de covariance $r_{\widehat{\mathbb{B}}} = \widehat{r}$ du processus gaussien centré $\widehat{\mathbb{B}}$, en combinant (9.32) et (9.37) :

$$\begin{aligned} \mathbb{E}\left(\widehat{\mathbb{B}}(t_1)\,\widehat{\mathbb{B}}(t_2)\right) &= r_{\mathbb{B}}(t_1,\,t_2) + g_0^T(t_1) I_0^{-1} \mathbb{E}\left(\xi\xi^T\right) I_0^{-1} g_0(t_2) \\ &\quad - \; g_0^T(t_1) I_0^{-1} \left(\int_0^{t_2} h_0 \, ds\right) \\ &\quad - \; g_0^T(t_2) I_0^{-1} \left(\int_0^{t_1} h_0 \, ds\right) \end{aligned} \tag{9.38}$$

avec

$$\mathbb{E}\left(\xi_j \xi_k\right) = \int_0^1 h_{0,\,j}(s) h_{0,\,k}(s)\, ds, \quad 1 \leq j, k \leq K, \tag{9.39}$$

c'est-à-dire :

$$\mathbb{E}(\xi \xi^T) = \int_0^1 h_0 h_0^T\, dt = \int_{\mathbb{R}} s_0 s_0^T\, dF_0 = I_0 \tag{9.40}$$

Finalement,

$$\begin{aligned} r_{\widehat{\mathbb{B}}}(t_1,\, t_2) &= r_{\mathbb{B}}(t_1,\, t_2) + g_0^T(t_1) I_0^{-1} g_0(t_2) \\ &- g_0^T(t_1) I_0^{-1} \left(\int_0^{t_2} h_0\, ds \right) \\ &- g_0^T(t_2) I_0^{-1} \left(\int_0^{t_1} h_0\, ds \right) \end{aligned} \tag{9.41}$$

Par exemple, si $\dim \Theta = K - 1$, cela prend la forme

$$r_{\widehat{\mathbb{B}}}(t_1,\, t_2) = r_{\mathbb{B}}(t_1,\, t_2) + \frac{g_0(t_1) g_0(t_2)}{I_0} - \frac{g_0(t_1)}{I_0} \int_0^{t_2} h_0(s)\, ds - \frac{g_0(t_2)}{I_0} \int_0^{t_1} h_0(s)\, ds \tag{9.42}$$

Nous allons montrer qu'en fait, on a ici, sous H_0 :

$$\int_0^t h_0(s)\, ds = g_0(t), \quad t \in [0,\, 1] \tag{9.43}$$

Par définition de g_0 et sous les hypothèses de régularité, on a :

$$\begin{aligned} g_0(t) &= \nabla_\theta \left(\int_{-\infty}^{F_0^{-1}(t)} f_\theta\left(y\right)\, dy \right) \Bigg|_{\theta = \theta_0} \\ &= \int_{-\infty}^{F_0^{-1}(t)} \nabla_0 f\left(y\right)\, dy \\ &= \int_{-\infty}^{F_0^{-1}(t)} \frac{\nabla_0 f\left(y\right)}{f(y)}\, f(y)\, dy \\ &= \int_{-\infty}^{F_0^{-1}(t)} s_0(y) f(y)\, dy \\ &= \int_0^t \left(s_0 \circ F_0^{-1}\right)(t)\, dt \\ &= \int_0^t h_0\left(s\right)\, ds \end{aligned} \tag{9.44}$$

En définitive,

$$r_{\widehat{\mathbb{B}}}(t_1,\, t_2) = r_{\mathbb{B}}(t_1,\, t_2) - g_0^T(t_1) I_0^{-1} g_0(t_2) \tag{9.45}$$

Par exemple, si $\dim \Theta = 1$, on a

$$r_{\widehat{\mathbb{B}}}(t_1,\, t_2) = r_{\mathbb{B}}(t_1,\, t_2) - \frac{g_0(t_1) g_0(t_2)}{I_0} \tag{9.46}$$

De plus, on peut démontrer que pour des paramètres de position et d'échelle, $r_{\widehat{\mathbb{B}}}(t_1, t_2)$ donné par (9.45) ne dépend pas de la valeur θ_0 du paramètre inconnu (sous H_0). C'est tout l'intérêt de s'être ramené sur [0, 1] par la transformation $U = F_0(X)$. Voir les exercices ci-dessous.

Ainsi, l'effet de l'introduction de $\widehat{\theta}_n$ à la place de θ_0 se traduit par une modification du processus gaussien limite $\widehat{\mathbb{B}}$, qui n'est plus le pont brownien $\mathbb{B}$. Pour pouvoir encore utiliser les tests de Kolmogorov-Smirnov et de Cramér-von Mises (et d'autres basés sur $\mathbb{F}_n$), il faut donc *tabuler*

$$\mathbb{P}\left(\sup_{t\in[0,1]}\left|\widehat{\mathbb{B}}(t)\right| > a\right), \quad a \geq 0 \tag{9.47}$$

et

$$\mathbb{P}\left(\int_0^1 \widehat{\mathbb{B}}^2(t)\,dt > a\right), \quad a \geq 0 \tag{9.48}$$

pour chaque famille $\mathcal{M}_0$. C'est ce qui a été fait, en particulier par D'Agostino et Stephens (1986). On peut alors calculer, pour chaque α, $0 < \alpha < 1$, courant ($\alpha = 0,05, 0,025$, etc.), les quantiles $\widehat{q}^{\,\mathrm{KS}}_{1-\alpha}$ et $\widehat{q}^{\,\mathrm{CM}}_{1-\alpha}$, d'où les régions de rejet asymptotiques.

Le principe d'utilisation de ces tables, lorsque θ est un paramètre de position et échelle que nous allons noter $\theta = (\mu, \sigma)$, est le suivant : partant de l'échantillon $X_1, \ldots, X_n$, on calcule, pour le modèle $\mathcal{M}_0$ que l'on veut tester, l'estimateur MV $\widehat{\theta}_n = (\widehat{\mu}_n, \widehat{\sigma}_n)$, puis on forme

$$\frac{X_i - \widehat{\mu}_n}{\widehat{\sigma}_n}, \quad i = 1, \ldots, n, \tag{9.49}$$

et l'échantillon ordonné correspondant. On en déduit un échantillon ordonné approximativement uniforme sur [0, 1]. Plus généralement, on forme les

$$\widehat{U}_i = F_{\widehat{\theta}_n}(X_i), \quad i = 1, \ldots, n \tag{9.50}$$

Sous H_0, $\widehat{U}_i = F_{\widehat{\theta}_n}\left[F_0^{-1}(U_i)\right] \in [0,1]$, $i = 1, \ldots, n$, et on peut démontrer que le processus empirique associé $\widehat{\mathbb{B}}_n$ converge vers $\widehat{\mathbb{B}}$ sous H_0. On utilise alors la table correspondant à la famille $\mathcal{M}_0$ pour calculer $\widehat{q}^{\,\mathrm{KS}}_{1-\alpha}$ ou $\widehat{q}^{\,\mathrm{CM}}_{1-\alpha}$, et on accepte ou rejette H_0 selon que la statistique de Kolmogorov-Smirnov KS_n basée sur $\widehat{U}_1, \ldots, \widehat{U}_n$ est inférieure à $\widehat{q}^{\,\mathrm{KS}}_{1-\alpha}$ ou supérieure à $\widehat{q}^{\,\mathrm{KS}}_{1-\alpha}$, ou que la statistique de Cramér-von Mises CM_n basée sur $\widehat{U}_1, \ldots, \widehat{U}_n$ est inférieure à $\widehat{q}^{\,\mathrm{CM}}_{\alpha}$ ou supérieure à $\widehat{q}^{\,\mathrm{CM}}_{\alpha}$, respectivement. (Rappelons que $\mathrm{KS}_n = \sqrt{n}\,D_n$ et que $\mathrm{CM}_n = W_n^2$. Voir (9.17) à (9.24).)

Exemple 13 (Cas où $\mathcal{M}_0 = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R},\ \sigma^2$ connu.$\}$) *Ici, $\theta = \mu \in \mathbb{R}$. Il nous faut calculer (sous H_0) les fonctions F_0^{-1}, $h_0 = s_0 \circ F_0^{-1}$ (s_0 est le score en $\theta_0 = \mu_0$), $g_0 = \frac{\partial F_{\mu,\sigma^2}}{\partial \mu}\Big|_{\mu=\mu_0} \circ F_0^{-1}$, et I_0. On pourra alors en déduire la forme de $\mathbb{B}$.*

Nous vérifierons que $\widehat{\mathbb{B}}$ ne dépend pas de la valeur μ_0.

On a ici, pour $\theta = \mu$:

$$\ln f_\theta(x) = -\ln\left(\sigma\sqrt{2\pi}\right) - \frac{(x-\theta)^2}{2\sigma^2},$$

donc

$$s_\theta(x) = \frac{\partial}{\partial\theta} \ln f_\theta(x) = \frac{x-\theta}{\sigma^2}$$

L'information de Fisher se calcule directement à partir de la dérivée partielle seconde par rapport à θ, et on trouve :

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, dy$$
$$= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt \quad = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

donc

$$F_0^{-1}(t) = \mu_0 + \sigma\Phi^{-1}(t), \quad t \in]0,\, 1[,$$

$$\frac{\partial F_\theta}{\partial\theta}(x) = \frac{\partial}{\partial\theta}\Phi\left(\frac{x-\theta}{\sigma}\right) = -\frac{1}{\sigma}\varphi\left(\frac{x-\theta}{\sigma}\right),$$

avec

$$\varphi(y) = \Phi'(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}},$$

d'où

$$h_0(t) = s_0\left[F_0^{-1}(t)\right] = \frac{1}{\sigma}\Phi^{-1}(t)$$

$$g_0(t) = \left.\frac{\partial F_\theta}{\partial\theta}\right|_{\theta=\theta_0}\left[F_0^{-1}(t)\right] = -\frac{1}{\sigma}\varphi\left(\Phi^{-1}(t)\right)$$

Ainsi,

$$\begin{aligned}\int_0^t h_0(s)\, ds &= \frac{1}{\sigma}\int_0^t \Phi^{-1}(s)\, ds \\ &= \frac{1}{\sigma}\int_{-\infty}^{\Phi^{-1}(t)} y\varphi(y)\, dy \\ &= -\frac{1}{\sigma}\varphi\left(\Phi^{-1}(t)\right) = g_0(t), \quad t \in]0,\, 1[\end{aligned}$$

On trouve donc, d'après (9.42) :

$$r_{\widehat{\mathbb{B}}}(t_1,\, t_2) = r_{\mathbb{B}}(t_1,\, t_2) - \frac{g_0(t_1)\, g_0(t_2)}{I_0}$$
$$= r_{\mathbb{B}}(t_1, t_2) - \varphi\left(\Phi^{-1}(t_1)\right)\varphi\left(\Phi^{-1}(t_2)\right)$$

avec, rappelons-le, $r_{\mathbb{B}}(t_1,\, t_2) = t_1 \wedge t_2 - t_1 t_2$. On observe que le résultat ne dépend ni de $\mu = \mu_0$ ni de σ^2 (ici supposé connue).

Remarquons que, par un simple calcul de covariance, on peut en déduire que, dans ce cas, on a la représentation

$$\mathbb{B} = \widehat{\mathbb{B}} + \frac{g_0}{\sqrt{I_0}}\xi^*, \tag{9.51}$$

où $\xi^* \sim \mathcal{N}(0,\,1)$ est indépendante de $\widehat{\mathbb{B}}$ (mais pas de $\mathbb{B}$).

Exercice 9.8 (Généralisation de l'exemple précédent) *On suppose que*

$$f_\theta(x) = g(x-\theta), \quad g \text{ densité sur } \mathbb{R}$$

et donc

$$F_\theta(x) = G(x-\theta), \quad G \text{ f.r. de densité } g$$

1. *Sous quelles hypothèses peut-on écrire :*

$$s_\theta(x) = -\frac{g'(x-\theta)}{g(x-\theta)} \ ?$$

2. *Montrer que* $F_0^{-1}(t) = \theta_0 + G^{-1}(t)$, $t \in]0,\,1[$.
3. *En déduire* $h_0 = s_0 \circ F_0^{-1}$, *puis calculer* $\int_0^t h_0(s)\,ds$ *sous une forme simple.*
4. *Montrer que*

$$g_0(t) = -g\left[G^{-1}(t)\right]$$

5. *En déduire que :*

$$r_{\widehat{\mathbb{B}}}(t_1,t_2) = r_{\mathbb{B}}(t_1,t_2) - \frac{g\left[G^{-1}(t_1)\right] g\left[G^{-1}(t_2)\right]}{I_0}$$

Exercice 9.9 *Reprendre les calculs dans le cas du modèle Exp*(θ), *où* $f_\theta(x) = (1/\theta)e^{-x/\theta}$, $x \geq 0$, $\theta > 0$.

Exercice 9.10 (Généralisation de l'exemple précédent) *On suppose que*

$$f_\theta(x) = \frac{1}{\theta}\, g\left(\frac{x}{\theta}\right), \quad g \text{ densité sur } \mathbb{R},$$

et donc

$$F_\theta(x) = G\left(\frac{x}{\theta}\right), \quad G \text{ f.r. de densité } g$$

avec $\theta > 0$.

1. *Calculer, sous des conditions à préciser, la fonction de score* $s_\theta(x)$, *à l'aide de* g *et* g'.
2. *Calculer* $F_0^{-1}(t)$ *pour* $t \in]0,\,1[$.
3. *En déduire* $h_0 = s_0 \circ F_0^{-1}$, *puis* $\int_0^t h_0(s)\,ds$.
4. *En déduire* $g_0(t)$, *puis* $r_{\widehat{\mathbb{B}}}(t_1,\,t_2)$ *: cette fonction* $r_{\widehat{\mathbb{B}}}$ *dépend-elle de* θ_0 *?*

9.4 Applications pratiques

Nous allons maintenant présenter les tabulations de Stephens[1] qui donnent, grâce à des corrections obtenues par simulation de Monte-Carlo, le moyen de calculer, pour toute taille d'échantillon n, une *bonne approximation du vrai quantile* d'ordre $1-\alpha$, $q_{n,\,1-\alpha}$, des lois de K_n^+, K_n^-, K_n, W_n^2, ceci pour $\alpha = 25\%, 15\%, 10\%, 5\%, 2,5\%, 1\%$.

T	T^*	niveau de signification α					
		0.25	0.15	0.10	0.05	0.025	0.01
D	$D(\sqrt{n}+0.12+0.11/\sqrt{n})$	1.019	1.138	1.224	1.358	1.480	1.628
W^2	$(W^2-0.4/n+0.6/n^2)(1+1/n)$	0.209	0.284	0.347	0.461	0.581	0.743

TABLEAU 9.1 – KS et CvM. Hypothèse nulle simple.

Voici comment s'en servir, sur un exemple simple. Supposons que l'on dispose d'un échantillon de taille $n = 16$, et qu'on ait trouvé

$$D_n = \sup_{\mathbb{R}} |\mathbb{F}_n(x) - F_0(x)| = 0,171.$$

Sans correction, on en déduirait que $K_n = \sqrt{n}\, D_n = \sqrt{16} \times 0,171 = 4 \times 0,171 = 0,684$, que l'on doit comparer à $q_{1-\alpha} = 1,358$ au niveau de signification asymptotique $\alpha = 5\%$. Avec correction, on calcule

$$\begin{aligned} K_n^* &= D_n \left(\sqrt{n} + 0,12 + \frac{0,11}{\sqrt{n}}\right) \\ &= 0,171 \times \left(4 + 0,12 + \frac{0,11}{4}\right) \\ &= 0,171 \times (4,12 + 0,027) \\ &= 0,171 \times 4,147 \\ &\simeq 0,70, \end{aligned}$$

que l'on doit aussi comparer à $q_{1-\alpha} = 1,358$ pour $\alpha = 5\%$. On ne rejette donc pas H_0.

Stephens remarque que pour $n \geq 20$, les corrections sont relativement mineures. Elles ne sont importantes que si la statistique mesurée est (éventuellement après multiplication par $\sqrt{n}$, dans le cas des statistiques de type supremum) très proche de la valeur $q_{1-\alpha}$, pour la statistique et le niveau α considérés.

Exemple 14 *Supposons que $n = 20$, et qu'on ait mesuré $W_n^2 = 0,50$ (en tenant compte des précisions des mesures). Sans correction, on doit rejeter H_0 au niveau 5%, mais ne pas rejeter H_0 au niveau 2,5%. Il y a doute. La correction donne*

$$\begin{aligned} W_n^{*2} &\approxeq \left(0,50 - \frac{0,4}{20} + \frac{0,6}{400}\right)\left(1,0 - \frac{1}{20} + \frac{1}{400}\right) \\ &\approxeq (0,50 - 0,02) \times 0,95 \;=\; 0,456\,, \end{aligned}$$

mais la dernière décimale n'est pas significative ici. On peut dire que W_n^{*2} est situé entre 0,45 et 0,46. Par conséquent, on peut ne pas rejeter H_0 au niveau 5%.

[1] D'Agostino et Stephens (1986).

9.5 Goodness-of-fit tests based on the empirical distribution function. Tests of Uniformity

<< Several[2] goodness-of-fit tests are based on a comparison of the hypothesized cumulative distribution function $F(x)$ with the empirical distribution function $\mathbf{F}_n(x)$ obtained from a random sample of n observations

$$x_{(1)} \leq \ldots \leq x_{(n)}$$

Tests are given below in this section, based on five statistics usually associated with the names of Komolgorov-Smirnov (the statistic D), Cramér-von Mises (W^2), Kuiper (V), Watson (U^2) and Anderson-Darling (A). In many problems it has been found that the tests give very similar answers, but because the procedure involved in their application has been put into the simple form devised by M.A. Stephens, set out in the single page of Table 1, we shall describe and illustrate the use of all five statistics.

The tests will be defined for three cases, depending on what is known of $F(x)$:

Case 1 : $F(x)$ is completely specified and we write $z_{(i)} = F\left(x_{(i)}\right)$.

Case 2 : $F(x)$ is the normal distribution function $N(\mu, \sigma)$, with μ and σ specified. Here we shall write $\Phi(x)$ for the standardized function and set

$$z_{(i)} = \Phi\left(\frac{x_{(i)} - \bar{x}}{s}\right) \tag{9.52}$$

where $\bar{x}$ is the sample mean and $s^2 = \sum_i (x_i - \bar{x})^2 / (n-1)$.

Case 3 : $F(x)$ is the negative exponential distribution function which writes as

$$F(x) = 1 - e^{-ax},$$

$x > 0$ and a is unknown. We know set

$$z_{(i)} = 1 - e^{-x_{(i)}/\bar{x}} \tag{9.53}$$

The test statistics defined below are all expressed in terms of the $z_{(i)}$, which will be in ascending order of magnitude. The asymptotic distribution of the statistics are known, but Stephen's appropriate modifications of the finite sample statistics, defined in the second column of Table Exponential case (case 2), may be referred without serious error to their asymptotic distributions, five of the upper percentage points of which are contained in the last column of the table. The basis of this ingenious empirical procedure, based on part theoretical and part Monte Carlo investigation, has been described in the references given below.

[2]Extrait de D'Agostino et Stephens (1996).

9.5.1 The Komolgorov-Smirnov statistics

Define

$$D^+ = \max_{1\leq i\leq n}\left(\frac{i}{n} - z_{(i)}\right),\ \ D^- = \max_{1\leq i\leq n}\left(z_{(i)} - \frac{i-1}{n}\right),\ \ D = \max\left(D^-,\, D^+\right) \tag{9.54}$$

Proceed by calculating the modified statistic, $T(D)$, shown in the second column of the table, using the formula appropriate to case 1, 2 or 3. Finally, compare $T(D)$ with the significance points shown in the last column of the table. For example, if in the test for normality (case 2) we had $n = 25$ observations and found $D = 0.186$, then

$$T(D) = 0.186\left(5 - 0.01 + \frac{0.85}{5}\right) = 0.960$$

The result is seen to be just significant at the 2.5 % level.

For case 1, a modification is given also for the one-sided statistics D^-, D^+ ; this is the same in both cases as also are the percentage points.

9.5.2 The Cramér-von Mises statistic

Calculate

$$W^2 = \sum_{i=1}^{n}\left(z_{(i)} - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n} \tag{9.55}$$

Modify to $T\left(W^2\right)$ and refer to the appropriate significance level. >>

Exemple 15 *Ici, on teste H_0 [F_0 loi uniforme sur $[0,\, \tau]$], avec $\tau = 1\,520$: le paramètre θ est donc connu. On a enregistré $n = 20$ instants de rupture sur une période de $\tau = 1\,520$ heures, au cours d'un essai. On veut tester l'hypothèse nulle que ces données sont uniformément distribuées sur $[0,\, \tau]$.*

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	30	36	104	286	291	658	893	955	1149	1195
i	11	12	13	14	15	16	17	18	19	20
$x_{(i)}$	1208	1240	1277	1282	1363	1384	1421	1477	1504	1510

TABLEAU 9.2 – Données de rupture ; $n = 20$.

Exercice 9.11 *Effectuer les calculs. Quelles conclusions peut-on en déduire ?*

9.6 Tester l'adéquation d'une loi normale $\mathcal{N}(\mu, \sigma^2)$

Cas 1 : μ inconnu, σ^2 connu : $\theta = \mu$.

Cas 2 : μ connu, σ^2 inconnu : $\theta = \sigma^2$.

Cas 3 : $\theta = (\mu, \sigma^2)$.

La procédure est toujours de même nature : il faut former $Z_{(i)} = F(X_{(i)} \,;\, \widehat{\theta}_n)$ $(1 \leq i \leq n)$, $(X_{(i)} : 1 \leq i \leq n)$ *étant l'échantillon ordonné*, puis tester

$$H_0 \; [Z_1, \ldots, Z_n \sim \mathcal{U}\,[0,\, 1]]$$

Pour cela, on forme d'abord les statistiques centrées réduites :

Cas 1 :

$$W_i = \frac{X_{(i)} - \widehat{\mu}_n}{\sigma} \;\; \left(\sigma \text{ connu}; \widehat{\mu}_n = \overline{X}\right)$$

Cas 2 :

$$W_i = \frac{X_{(i)} - \mu}{\widehat{\sigma}_n} \;\; \left(\mu \text{ connu}; \widehat{\sigma}_n^2 = \frac{1}{n}\sum_1^n (X_i - \mu)^2\right)$$

Cas 3 :

$$W_i = \frac{X_{(i)} - \widehat{\mu}_n}{\widehat{\sigma}_n} \;\; \left(\widehat{\mu}_n = \overline{X};\; \widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_1^n (X_i - \widehat{\mu}_n)^2\right)$$

Puis on forme $Z_{(i)} = \Phi(W_i)$, Φ désignant la fonction de répartition de $\mathcal{N}(0,\, 1)$. On forme ensuite D_n^+, D_n^-, puis D_n, ou encore W_n^2, en utilisant les formules (9.19) ou (9.20), avec $U_{i,\,n} \leftarrow Z_{(i)}$, puis on utilise la Table 2 ou la Table 3, selon les cas.

Stat	Cas	niveau de signification α					
		0.25	0.15	0.10	0.05	0.025	0.01
W^2	σ^2 connue	0.094	0.117	0.134	0.165	0.197	0.238
W^2	μ connue	0.190	0.263	0.327	0.442	0.562	0.725

TABLEAU 9.3 – CvM. Normalité, un des paramètres connu.

T	T^*	niveau de signification α					
		0.25	0.15	0.10	0.05	0.025	0.01
D	$D(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$	-	0.775	0.819	0.895	0.995	1.035
W^2	$W^2(1.0 + 0.5/n)$	0.074	0.091	0.104	0.126	0.148	0.179

TABLEAU 9.4 – KS et CvM. Normalité, les deux paramètres inconnus.

Exemple 16 *Tester la normalité, $\theta = (\mu,\, \sigma^2)$ et $n = 15$. Ici, on cherche à tester la normalité, les deux paramètres étant inconnus.*

Exercice 9.12 *Effectuer les calculs. Quelles conclusions peut-on en déduire ?*

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	0.301	0.519	0.653	0.690	0.892	0.964	0.978	0.987	1.017	1.233
i	11	12	13	14	15					
$x_{(i)}$	1.346	1.357	1.562	1.845	1.944					

TABLEAU 9.5 – Logarithme d'une mesure d'endurance mécanique ; $n = 15$.

9.7 Tester l'adéquation d'une loi exponentielle

Ici,

$$F(x\ ;\ \alpha, \beta) = 1 - e^{-(x-\alpha)/\beta},$$

avec : $\beta > 0,\ x \geq \alpha$.

Cas 1 : α inconnu, β connu : $\theta = \alpha$.

Cas 2 : α connu, β inconnu : $\theta = \beta$.

Cas 3 : $\theta = (\alpha,\ \beta)$.

En particulier, le cas 1 peut se ramener, par la transformation $X'_{(i)} = X_{(i)} - X_{(1)}\ (i \geq 2)$, au cas des tests à paramètres connus, avec un $(n-1)$-échantillon : on teste

$$H_0 \left[F_0(x') = 1 - e^{-x'/\beta} \right]\ (x' > 0),$$

avec β connu.

n	niveau de signification α					
	0.25	0.15	0.10	0.05	0.025	0.01
10	0.753	0.833	0.889	0.977	1.048	1.119
20	0.786	0.872	0.927	1.021	1.099	1.195
50	0.813	0.879	0.960	1.061	1.149	1.257
100	0.824	0.911	0.972	1.072	1.171	1.278
∞	0.840	0.927	0.995	1.094	1.184	1.299

TABLEAU 9.6 – KS, Stat $\sqrt{n}D$: Exp(α, β), les deux paramètres inconnus, en fonction de n.

T	T^*	niveau de signification α					
		0.25	0.15	0.10	0.05	0.025	0.01
W^2	$W^2(1 + 2.8/n - 3/n^2)$	0.116	0.148	0.175	0.222	0.271	0.338

TABLEAU 9.7 – CvM. Exponentialité, paramètres inconnus.

Exemple 17 *Tester l'exponentialité, $\alpha = 0$, $\theta = \beta$ inconnu ; $n = 27$. Les données suivantes représentent les intervalles, rangés par ordre croissant et comptés en nombre de jours, entre les pannes consécutives du système d'air conditionné sur un nouveau modèle d'avion de ligne récemment mis en service et en cours de mise au point. (Donc, les durées des intervalles ne*

sont presque certainement pas apparues dans cet ordre : il s'agit de statistiques ordonnées !) On veut tester l'exponentialité de la loi de ces intervalles, seul le paramètre de moyenne étant inconnu, car l'origine est fixée à la valeur 0.

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	1	4	11	16	18	18	18	24	31	39
i	11	12	13	14	15	16	17	18	19	20
$x_{(i)}$	46	51	54	63	68	77	80	82	97	106
i	21	22	23	24	25	26	27			
$x_{(i)}$	111	141	142	163	191	206	216			

TABLEAU 9.8 – Intervalles de temps ordonnés entre pannes, $n = 27$.

Exercice 9.13 *Effectuer les calculs et en tirer des conclusions.*

Exercice 9.14 *Proposer et essayer un test du chi-deux sur ces exemples.*

Exercice 9.15

1. *Soit X une v.a. à valeurs dans $\mathbb{N}$ suivant une loi de Poisson de paramètre $\lambda > 0$. Déterminer sa fonction de répartition F_λ et l'inverse généralisée $F_\lambda^{\leftarrow}$ de F_λ.*
2. *Soit U une v.a. $\mathcal{U}[0, 1]$. Expliquer par quel algorithme on construit la v.a.*

$$Y = F_\lambda^{\leftarrow}(U)$$

qui a même loi que X.

3. *Déterminer la f.r. G_λ de la v.a. $Z = F_\lambda(Y)$. Cette v.a. est-elle uniforme sur $[0, 1]$? La f.r. $G_\lambda$dépend-elle de λ ?*
4. *Peut-on utiliser les tests de Kolmogorov-Smirnov ou de Cramér-von Mises pour tester l'adéquation du modèle de Poisson ?*
5. *Comment faire pour utiliser commodément le test du chi-deux pour tester l'adéquation du modèle de Poisson ?*

Conclusion

– Fonctions de répartition continues, ici.
– Fonction de répartition empirique et processus empirique.
– Caractère simplement asymptotique ($n \to +\infty$).
– Aucune garantie d'optimalité.
– Tests d'adéquation : Kolmogorov-Smirnov et Cramér-von Mises, entre autres.
– Pont brownien sur [0, 1].
– Importance de la prise en compte des estimateurs utilisés pour le modèle sous H_0.

Eléments bibliographiques pour ce chapitre

Association pour la Statistique et ses Utilisations (1996) *Inférence non paramétrique. Les statistiques de rangs*, Jean-Jacques Droesbeke et Jeanne Fine éditeurs, Collection "Ellipses", Editions de l'Université Libre de Bruxelles : Bruxelles.

Capéraà, P. & Van Cutsem, B. (1988) *Méthodes et modèles en Statistique non paramétrique*. Dunod : Paris.

Conover, W. J. (1971) *Practical Nonparametric Statistics*. Wiley : New York.

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 2. Masson : Paris, 2e édition.

D'Agostino, R. B. & Stephens, M. A. (1986) *Goodness-of-Fit Techniques*, Statistics, textbooks and monographs **68**. Marcel Dekker : New York and Basel.

Dudley, R. M. (1989) *Real Analysis and Probability*, Mathematics Series. Chapman and Hall : New York and London.

Eubank, R. L., Hart, J. D. & LaRiccia, V. N. (1993) Testing goodness-of-fit via nonparametric function estimation techniques. *Commun. Statist.- Theory Meth.* **22**, 3327–3354.

Gouriéroux, C. & Monfort, A. (1989) *Statistique et modèles économétriques*. Collection "Economie et statistiques avancées", Economica : Paris.

Lecoutre, J.-P. & Tassi, P. (1987) *Statistique non paramétrique et robustesse*. Collection "Economie et statistiques avancées", Economica : Paris.

Rayner, J. C. W. & Best, D. J. (1989) *Smooth Tests of Goodness-of-Fit*. Oxford University Press : Oxford.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley : New York.

10

Tests bayésiens

Les tests bayésiens sont les tests mis en œuvre dans le cadre de la Statistique bayésienne. La Statistique bayésienne est une théorie "concurrente" de la Statistique dite classique[1], en ce sens que chacune d'elles propose vis-à-vis d'un même problème une approche et une résolution complètement différentes. Il faut bien comprendre que ces deux théories ne peuvent être discriminées sur un strict plan mathématique : leur cohérence interne n'est pas à remettre en cause. Ce sont leurs *axiomatiques* qui diffèrent, car fondées sur deux démarches scientifiques distinctes, comme on le verra plus loin.

Nous avons supposé que la Statistique bayésienne n'était pas ou peu connue des étudiants auxquels s'adresse ce polycopié. Pour clarifier le propos de ce chapitre, nous présentons donc très succinctement les fondements de cette théorie dans une première partie (§10.1). La nouveauté des concepts introduits (loi sur les paramètres, optique décisionnelle, etc.) risque de dérouter le lecteur au premier abord. Cependant, ces fondements bien compris, on verra que le problème des tests d'hypothèses admet une résolution simple et élégante dans un tel cadre (§10.2). Résolution qui s'avère en revanche incompatible avec l'approche de Neyman-Pearson, comme on pourra le constater en §10.3.

10.1 Fondements de la statistique bayésienne

10.1.1 Le paradigme bayésien

On considère à nouveau, par souci de cohérence avec les chapitres précédents, que les variables aléatoires $X_1, \ldots, X_n$ sont i.i.d., de loi μ_{θ_0} appartenant à une famille paramétrique $\{\mu_\theta : \theta \in \Theta\}$, à laquelle correspond une famille de densités $\{f_\theta : \theta \in \Theta\}$. La vraisemblance des observations $x_1, \ldots, x_n$ sera désormais notée $f(\underline{x}|\theta)$, où $\underline{x}$ désigne le n-uple $(x_1, \ldots, x_n)$.

[1] dite aussi Statistique "fréquentiste", c'est notamment la théorie statistique mise en œuvre dans les autres chapitres.

L'originalité de l'approche bayésienne est d'exprimer l'incertitude sur la "vraie" valeur du paramètre θ_0 en conférant un caractère *aléatoire* à ce paramètre[2] (au contraire de la Statistique fréquentiste, qui considère que θ_0 est inconnu mais fixe). Ainsi, θ fera désormais référence à une variable aléatoire à valeurs dans Θ, l'espace des paramètres.

Définition 10.1 *Un modèle statistique bayésien est la double donnée d'un modèle paramétrique* $\{f(x|\theta) : \theta \in \Theta\}$ *et d'une loi de probabilité, de densité* π *par rapport à une mesure de référence, sur l'espace des paramètres* Θ*, dite loi a priori, qui est la loi marginale de la variable aléatoire* θ*. Par souci de commodité, on notera simplement* $d\theta$ *cette mesure de référence, qui coïncide d'ailleurs le plus souvent avec la mesure de Lebesgue.*

Théorème 10.1 (Bayes) *Pour un modèle bayésien* (f, π)*, la loi conditionnelle du paramètre* θ *sachant les observations* $\underline{x}$ *(dite loi a posteriori) admet pour densité :*

$$\pi\left(\theta|\,\underline{x}\right) = \frac{\pi(\theta) f\left(\underline{x}|\,\theta\right)}{\int \pi(\theta) f\left(\underline{x}|\,\theta\right)\, d\theta}$$

La démonstration de ce théorème est trés aisée : on la retrouve en appliquant la formule de Bayes, qui permet l'inversion de probabilités conditionnelles entre deux événements A et B (ou de densités conditionnelles, cette formule étant aussi valable dans le cas continu) :

$$\mathbb{P}(A|\,B) \;=\; \frac{\mathbb{P}(A)\mathbb{P}(B|\,A)}{\mathbb{P}(B)}$$

On remplace A par θ et B par $\underline{x}$. Le terme au dénominateur $\mathbb{P}(\underline{x})$ ne dépend pas de θ : il s'agit d'une constante multiplicative, dont la valeur peut se retrouver par "normalisation" (en écrivant que $\int \pi(\theta|\,\underline{x})\, d\theta = 1$). On écrira d'ailleurs de manière équivalente :

$$\pi\left(\theta|\,\underline{x}\right) \propto \pi(\theta) f\left(\underline{x}|\,\theta\right)$$

($\propto$ signifiant "proportionnel à"). Cette écriture conditionnelle, propre à la Statistique baysienne, permet d'inverser les "causes" (θ) et les "effets" ($\underline{x}$), et d'obtenir à partir de la relation "θ donne $\underline{x}$" (exprimée par le modèle paramétrique $f(\underline{x}|\,\theta)$) la relation réciproque "$\underline{x}$ donne θ" (exprimée à l'aide de la loi a posteriori $\pi(\theta|\,\underline{x})$).

Revenons sur la nature de chacune des quantités en présence dans l'équation $\pi(\theta|\,\underline{x}) \propto \pi(\theta) f(\underline{x}|\,\theta)$:

1. la loi a priori $\pi(\theta)$ est la loi marginale du paramètre θ, ou, en d'autres termes, la loi du paramètre, *avant que* $\underline{x}$ *ne soit observé*. Elle résume donc à ce titre l'information sur θ disponible *a priori* (avant observation). Si aucune information n'est disponible a priori, il faut spécifier une loi *non informative* (voir exemple plus loin).

[2] Ceci dit, le modèle sous-jacent suppose en général que, du point de vue théorique au moins, les observations sont issues d'un mécanisme probabiliste correspondant à une valeur θ_0 "vraie", mais inconnue, du paramètre, et que c'est notre connaissance des valeurs possibles de θ qui est représentée par des lois de probabilité. Bien entendu, il ne s'agit là que d'idéalisations et de modèles idéaux/idéels. Dans la réalité, il n'existe jamais de paramètre ni de vraie valeur du paramètre, et notre connaissance des valeurs possibles du paramètre n'est pas une loi de probabilité.

2. La vraisemblance du modèle $f(\underline{x}|\,\theta)$ s'interprète, dans un cadre bayésien, comme une vraisemblance *conditionnelle* des observations, sachant la valeur du paramètre aléatoire θ.
3. La loi a posteriori $\pi(\theta|\,\underline{x})$ donne alors l'information dont on dispose sur θ, *après observation*. Elle représente un compromis entre l'information a priori (donnée par π), et l'information tirée de l'observation de $\underline{x}$ (donnée par $f(\underline{x}|\,\theta)$).

Exemple 18 *Simon de Laplace (1749-1827) fournit l'un des premiers exemples (1786) d'un raisonement bayésien. Son but était de déterminer, à partir de la connaissance du nombre x de naissances masculines parmi n naissances à Paris, si la probabilité p de naître de sexe masculin est supérieure à $1/2$. Il assigna à p une loi a priori uniforme sur l'intervalle $[0,\,1]$, puis, partant de $f(x|\,\theta) = C_n^x\, p^x(1-p)^{n-x}$ il calcula la probabilité suivante :*

$$\begin{aligned}\mathbb{P}(p \leq 1/2|\,x) &= \int_0^{1/2} \pi(p|\,x)\,dp \\ &= \frac{\int_0^{1/2} p^x(1-p)^{n-x}\,dp}{\int_0^1 p^x(1-p)^{n-x}\,dp}\end{aligned}$$

Pour $x = 251\,527$ et $n - x = 24\,194$, Laplace obtint $\mathbb{P}(p \leq 1/2|\,x) = 1,15 \times 10^{-42}$ et en conclut que p était très vraisemblablement supérieure à $1/2$. Il vérifia de la même manière que la probabilité de naître de sexe masculin était plus élevée à Londres qu'à Paris.

Le choix par Laplace d'une loi a priori uniforme constitue en fait l'une des premières tentatives d'introduction d'une loi non informative, c'est-à-dire une loi modélisant l'absence d'information a priori sur p. Dans cette optique, une loi uniforme revient à attribuer "une chance égale" à toutes les valeurs de p, puisque rien ne permet de les départager a priori. Cette justification n'est qu'intuitive, et peut être contestée, notamment parce qu'elle pose des problèmes d'invariance par reparamétrage. Nous reviendrons plus amplement sur les difficults de la modélisation a priori en 10.1.3.

Nous verrons dans la partie suivante que les estimateurs bayésiens sont obtenus à partir de la loi a posteriori. Mais cette loi de densité $\pi(\theta|\,\underline{x})$ ne doit pas être considérée comme un simple résultat intermédiaire. Elle constitue en fait la première et la plus générale des réponses bayésiennes au problème de l'inférence statistique. En tant que loi de probabilité sur Θ, l'espace des paramètres, elle représente une information plus riche qu'un simple point de Θ (estimateur ponctuel), et permet notamment de quantifier l'incertitude restante sur le paramètre θ, après prise en compte des observations.

Notons enfin qu'au fur et à mesure que le nombre n d'observations augmente, l'information apportée par les observations devient prépondérante sur l'information a priori. Cependant, l'utilisation de la loi a posteriori ne repose pas sur des justifications d'ordre asymptotique, contrairement à beaucoup de méthodes classiques : le conditionnement par rapport aux observations $\underline{x}$ permet de raisonner "à distance finie", pour le nombre n de variables X_i réellement observées, et non pour n tendant vers l'infini.

Exercice 10.1 *Voici un des nombreux Pillow problems inventés par Lewis Caroll (1832-1898), l'auteur de Alice au pays des merveilles :*

"Un sac contient une boule qui est soit blanche, soit noire. Une boule blanche est ensuite ajoutée dans le sac. On tire alors (au hasard et sans remise) une boule dans ce sac, qui se trouve être noire. Quelle est la probabilité d'obtenir au deuxième tirage une boule blanche ?"

Montrer qu'une résolution rigoureuse de ce problème passe par un raisonnement bayésien (à la fois en termes d'a priori, en attribuant une probabilité $1/2$ *à chacun des événements* $\{$*la boule initiale est blanche/noire*$\}$*, et en termes de probabilités conditionnelles, conditionnellement à l'événement* $\{$*la première boule tirée est noire*$\}$*).*

NB : la réponse correcte est $2/3$*.*

10.1.2 L'approche décisionnelle : estimateurs de Bayes

La théorie de la décision reformule le problème de l'inférence statistique en ces termes : l'observateur (ou *décideur*) doit prendre une décision (choix d'un élément δ dans un ensemble de décisions $\mathcal{D}$) en contexte incertain (caractère aléatoire de θ). Ce décideur se caractérise de plus par une fonction de perte $L(\theta,\, \delta)$, qui permet d'évaluer la *perte* qu'il encourt à prendre la décision δ, lorsque le paramètre correspondant au modèle observé est θ. Plus $L(\theta,\, \delta)$ est élevé, plus le choix de δ est dommageable pour le décideur.

On définit le *coût a posteriori,* associé à une décision δ, comme l'espérance de la perte encourue pour cette décision δ, conditionellement aux observations :

$$\begin{aligned}\rho(\pi,\, \delta|\, \underline{x}) &= \mathbb{E}^{\pi}\left[L(\theta,\, \delta)|\, \underline{x}\right] \\ &= \int_{\Theta} L(\theta,\, \delta)\pi(\theta|\, \underline{x})\, d\theta\end{aligned}$$

On appellera alors *estimateur de Bayes* associé à la loi a priori π et à la fonction de perte L, la fonction qui associe à un $\underline{x}$ donné la décision $\delta^{\pi}(\underline{x})$ qui *minimise* le coût a posteriori :

$$\delta^{\pi} : \underline{x} \mapsto \delta^{\pi}(\underline{x}) = \arg\min_{\delta\in\mathcal{D}} \rho(\pi,\, \delta|\, \underline{x})$$

L'estimateur de Bayes donne donc la décision *optimale*, au vu de l'observation de $\underline{x}$, et pour la fonction de perte L. Notons qu'il est aussi possible de caractériser le décideur par une fonction d'utilité, plutôt que par une fonction de perte. Dans ce cas, en posant $L(\theta,\, \delta) = -\,U(\theta,\, \delta)$, on voit facilement que l'estimateur de Bayes correspondant maximise l'espérance de l'utilité du décideur.

Le problème de l'estimation d'un paramètre se présente comme un cas particulier de cette approche décisionnelle : l'ensemble de décision devient alors l'espace des paramètres ($D = \Theta$). Pour certaines fonctions de coût usuelles, les estimateurs bayésiens s'expriment simplement en fonction de la loi a posteriori. Citons deux cas importants :

– Pour le coût quadratique $L(\theta,\, \delta) = \|\, \theta - \delta\|^{\,2}$, l'estimateur de Bayes associé est l'espérance de la loi a posteriori
$$\delta^{\pi}(\underline{x}) = \mathbb{E}^{\pi}\left[\theta|\, \underline{x}\right]$$
– Pour le coût absolu $L(\theta,\, \delta) = |\, \theta - \delta|\,$, (en supposant $\Theta = \mathbb{R}$) l'estimateur de Bayes associé est la médiane de la loi a posteriori.

Les tests admettent aussi une formulation décisionnelle, comme nous allons le voir.

10.1.3 Modélisation a priori

La modélisation a priori est sans doute le point le plus délicat de l'analyse bayésienne. Deux types d'approches sont généralement considérés :
- une approche dite *subjective*, ou *informative*, qui revient à tenir compte (lorsqu'elles existent) d'informations a priori sur le paramètre (expériences précédentes, avis d'experts, connaissances extérieures au processus d'observation, etc.),
- une approche dite *objective*, ou *non informative*, qui revient à modéliser l'absence d'information a priori.

Dans le cas subjectif, traduire une information a priori (parfois formulée de façon assez vague) en une loi de probabilité proprement définie n'est pas chose aisée. Dans ce cas, la solution la plus souvent retenue est de réduire l'ensemble des lois a priori considérées à une famille paramétrique donnée (par exemple la famille des lois normales), et de choisir dans cette famille une loi qui semble compatible avec l'information a priori. Ainsi, l'avis d'un expert du type "Pour cette expérience, je doute que le paramètre θ puisse prendre une valeur absolue plus grande que 10" justifie l'utilisation de la loi a priori $\mathcal{N}(0,\, 5^2)$ (qui donne pour θ une probabilité a priori d'appartenir à l'intervalle $[-10,\, 10]$ d'environ 95%).

Dans un tel cadre, les lois dites *conjuguées* jouent un rôle important. Leur définition est la suivante :

Définition 10.2 *Une famille de lois $\mathcal{F}$ est dite conjuguée pour le modèle paramétrique $f(x|\theta)$ si et seulement si pour toute loi a priori π de $\mathcal{F}$, la loi a posteriori $\pi(\theta|x)$ correspondante appartient à $\mathcal{F}$.*

Par abus de langage, on qualifie aussi de conjuguée toute loi appartenant à une famille conjuguée. Les seules familles conjuguées réellement intéressantes sont les familles paramétriques. Dans un tel cas, la prise en compte de l'information apportée par les observations (correspondant au passage de l'a priori $\pi(\theta)$ à l'a posteriori $\pi(\theta|x)$) se traduit aisément par une mise à jour des paramètres de la loi suivie par θ.

Exemple 19 *Pour un modèle normal $\mathcal{N}(\theta, \sigma^2)$ (variance σ^2 connue), la famille des lois normales est conjuguée. Ainsi, pour n observations $x_1, \ldots, x_n$ supposées issues de ce modèle, et une loi a priori $\theta \sim \mathcal{N}(\mu,\, \tau^2)$, la loi a posteriori est de la forme*

$$\pi\left(\theta \,|x_1, \ldots, x_n\right) \sim \mathcal{N}\left(\frac{\overline{x}\tau^2 \,+\, \mu\sigma^2/n}{\tau^2 \,+\, \sigma^2/n},\, \frac{\tau^2\sigma^2/n}{\tau^2 \,+\, \sigma^2/n}\right),$$

où $\overline{x}$ est la moyenne des observations.

L'utilisation de lois a priori conjuguées présente plusieurs avantages. Tout d'abord, une loi a posteriori correspondant à une loi a priori conjuguée suit une loi de probabilité connue, et est donc simple à manier[3]. De plus, elle permet une évaluation simple des poids relatifs de

[3]Ceci est loin d'être un avantage mineur ! Tout autre type de loi a priori mène généralement à une loi a posteriori dont on ne connaît que la densit é à une constante multiplicative près. Le calcul des estimateurs de Bayes correspondants est alors en général beaucoup plus difficile à effectuer, et nécessite le plus souvent des méthodes numériques complexes, de type Monte-Carlo.

l'information a priori et de l'information apportée par les observations. Ainsi, dans l'exemple précédent, comparer l'influence respective de ces deux informations revient à comparer τ^2 (la variance de la loi a priori) à σ^2/n : plus τ^2 est faible devant σ^2/n, plus le poids de l'a priori est faible (et pour τ^2 suffisamment petit, $\pi(\theta|x)$ est approximativement la loi $\mathcal{N}(\overline{x},\, \sigma^2/n)$, qui ne dépend plus de l'a priori). Plus précisément, si τ^2 est de l'ordre de σ^2/k (k entier), la loi a priori a en quelque sorte le poids de k observations supplémentaires. Toutes ces considérations sont de précieux guides pour la détermination de la loi a priori.

Il faut noter que seuls les modèles à structure exponentielle admettent une famille conjuguée. De plus, le recours aux lois conjuguée n'est pas toujours satisfaisant, car sa justification repose plus sur des aspects pratiques (détermination aisée de la loi a posteriori) que véritablement informationnels. Néanmoins, l'approche conjuguée reste la solution la plus standard dans un cadre informatif.

Exercice 10.2 *Montrer que l'ensemble des loi Gamma forme une famille conjuguée pour le modèle $\mathcal{N}(\mu,\, 1/\theta)$ (moyenne μ connue). Retrouver à partir de ce résultat une famille conjuguée pour le modèle normal "complet" $\mathcal{N}(\mu,\, \sigma^2)$ (paramètre $\theta = (\mu,\, 1/\sigma^2)$) pour un échantillon d'au moins deux observations.*

Exercice 10.3 *Montrer que l'ensemble des lois Beta forme une famille conjuguée pour un modèle binomial Bin$(n,\, \theta)$.*

Lorsque aucune information a priori n'est disponible, ou même lorsque l'information est trop diffuse ou trop difficilement traduisible en termes de loi a priori, on est amené à construire une loi a priori dite non informative (approche dite objective). Ainsi, la loi loi uniforme sur l'intervalle $[0,\, 1]$ proposée par Laplace dans l'exemple 18 répond à cette volonté de ne pas discriminer a priori les valeurs possibles de p, et donc de leur attribuer une "chance égale".

Cependant, bien qu'intuitif, le recours a une loi uniforme dans le cas non informatif n'est pas complètement justifié : il pose notamment le problème essentiel de l'invariance par reparamétrage. Ainsi, dans l'exemple de Laplace, si l'on reparamètre le modèle en prenant $q = p^2$ comme nouveau paramètre, on trouve qu'une loi a priori uniforme pour p ne devient pas (par changement de variable) une loi uniforme sur q, et vice-versa. Un tel choix d'a priori fait donc dépendre, en général, l'inférence du paramétrage choisi, et n'est donc pas parfaitement "objectif". Une analyse objective plus convaincante revient alors à construire des lois a priori qui respectent une invariance complète (lois de Jeffreys), ou au moins partielle (pour un sous-groupe donné de transformations du paramètre, par exemple les translations), mais nous n'aborderons pas ici ces constructions en détail (voir cependant les exercices à la fin de cette partie pour une introduction à ces questions d'invariance).

Il faut noter aussi qu'une approche non informative est plus délicate quand l'espace des paramètres n'est pas compact. Dans ce cas, une loi de probabilité donnée sur l'espace des paramètres est toujours un peu informative, car elle revient à "localiser" le paramètre dans une certaine partie de cet espace, et ce aussi grande que soit sa variance (ou son étendue). Une analyse vraiment objective requiert alors l'utilisation de lois dites *impropres*.

Une loi a priori *impropre* est une loi qui définit sur l'espace Θ une mesure non plus finie, mais simplement σ-finie. La densité correspondante n'est alors pas intégrable :

$$\int_\Theta \pi(\theta)\, d\theta = +\infty.$$

Cette approche reste cependant valide tant que la loi a posteriori correspondante définit bien une loi de probabilité sur Θ, ce qui impose que

$$\int_\Theta \pi(\theta) f(x|\theta)\, d\theta < +\infty.$$

Exemple 20 *Soit $x \sim \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$, et soit la loi a priori impropre de densité constante sur $\mathbb{R}$, $\forall \theta$, $\pi(\theta) = 1$ (c'est-à-dire la mesure de Lebesgue sur $\mathbb{R}$) ; la loi a posteriori correspondante est alors bien définie :*

$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta) \propto (2\pi)^{-\frac{1}{2}} \exp\left\{-(x-\theta)^2/2\right\},$$

d'où $\theta|x \sim \mathcal{N}(x, 1)$.

La loi a priori impropre $\pi(\theta) = 1$ de l'exemple précédent est en quelque sorte une généralisation de la loi a priori uniforme au cas d'un espace des paramètres non compact. Par contre, si Θ était le groupe multiplicatif des réels strictement positifs, la loi a priori impropre non informative invariante par changement d'échelle serait définie par $\pi(\theta) = 1/\theta$. On retiendra qu'une approche non informative impose le plus souvent l'utilisation de lois impropres.

Exercice 10.4 *Soit un modèle paramétrique de la forme $x \sim f(x-\theta)$, $\theta \in \mathbb{R}$. Le paramètre θ est dit alors paramètre de position. Pour un tel modèle, on impose généralement aux lois a priori d'être invariantes par translations (c'est-à-dire par tout reparamétrage de la forme $\theta' = \theta + c$). Expliquer pourquoi cette restriction est naturelle. Montrer alors que la seule loi vérifiant cette invariance est la loi impropre $\pi(\theta) = 1$ pout tout θ.*

Exercice 10.5 *Soit un modèle paramétrique avec paramètre d'échelle : $x \sim f(x/\theta)/\theta$, $\theta > 0$. En utilisant un raisonnement analogue à celui de l'exercice précédent, justifier pour un tel modèle l'utilisation de la loi a priori impropre $\pi(\theta) = 1/\theta$ pout tout $\theta > 0$ (définie sur $\mathbb{R}^{+*}$).*

Exercice 10.6 *La loi de Jeffreys associée à un modèle paramétrique donné est définie par*

$$\pi_J(\theta) \propto \det\left(I(\theta)\right)^{1/2},$$

où $I(\theta)$ est l'information de Fisher du modèle en θ. Pour simplifier, on se restreint ici au cas où $\Theta = \mathbb{R}$. Montrer qu'alors $\pi_J(\theta)$ est une loi a priori invariante par tout difféomorphisme de classe C^1. Vérifier que $\pi_J(\theta)$ est le plus souvent une loi impropre. Calculer la loi de Jeffreys dans l'exemple de Laplace, et montrer qu'elle diffère de la loi uniforme.

10.2 Tests bayésiens : facteurs de Bayes

Tester une hypothèse nulle $H_0 : \theta \in \Theta_0$ contre l'hypothèse alternative $H_1 : \theta \in \Theta_1 = \Theta \backslash \Theta_0$ (le complémentaire de Θ_0 dans Θ) revient à se donner une fonction de perte L définie sur un espace des décisions $\mathcal{D} = \{0, 1\}$, où 1 vaut pour l'acceptation de H_0, et 0 pour son rejet. Un premier exemple d'une telle fonction L est le coût $0-1$:

$$L(\theta, \delta) = \begin{cases} 1 & \text{si } \delta = 1_{\Theta_0}(\theta) \\ 0 & \text{sinon} \end{cases}$$

qui donne une perte nulle pour une décision correcte, et égale à 1 pour une décision erronée. On vérifie aisément que l'estimateur de Bayes associé à cette fonction de perte est :

$$\delta^\pi(\underline{x}) = \begin{cases} 1 & \text{si } \mathbb{P}(\theta \in \Theta_0 | \, \underline{x}) \, > \, \mathbb{P}(\theta \in \Theta_1 | \, \underline{x}) \\ 0 & \text{sinon} \end{cases}$$

ce qui revient (en toute logique !) à choisir l'hypothèse la plus probable a posteriori. On peut généraliser le coût $0-1$ au coût dissymétrique suivant :

$$L(\theta, \delta) = \begin{cases} a_0 & \text{si } \delta = 0, \quad \theta \in H_0 \\ a_1 & \text{si } \delta = 1, \quad \theta \in H_1 \\ 0 & \text{sinon} \end{cases}$$

qui donne cette fois un estimateur de Bayes de la forme :

$$\delta^\pi(\underline{x}) = \begin{cases} 1 & \text{si } \frac{\mathbb{P}(\theta \in \Theta_0 | \, \underline{x})}{\mathbb{P}(\theta \in \Theta_1 | \, \underline{x})} \, > \, \frac{a_0}{a_1} \\ 0 & \text{sinon} \end{cases}$$

Les réels a_0 et a_1 permettent de différencier les coûts des deux types de décision erronée (erreurs de première et de seconde espèce).

Le rapport des probabilités $\mathbb{P}(\theta \in \Theta_0 | \, \underline{x}) / \mathbb{P}(\theta \in \Theta_1 | \, \underline{x})$ intervient donc naturellement comme un outil de décision pour les tests. Il est possible de réduire la dépendance de ce rapport à la loi a priori π, en définissant le *facteur de Bayes* :

Définition 10.3 *Le facteur de Bayes $B^\pi(\underline{x})$ associé au test des hypothèses Θ_0 contre Θ_1 pour le modèle (π, f), est défini comme :*

$$B^\pi(\underline{x}) = \frac{\mathbb{P}(\theta \in \Theta_0 | \, \underline{x})}{\mathbb{P}(\theta \in \Theta_1 | \, \underline{x})} \frac{\pi(\theta \in \Theta_1)}{\pi(\theta \in \Theta_0)}.$$

Pour mieux comprendre l'intérêt de cette nouvelle quantité, considérons le cas simple où Θ_0 et Θ_1 se réduisent à deux singletons $\{\theta_0\}$ et $\{\theta_1\}$ ($\Theta = \{\theta_0, \theta_1\}$). Le facteur de Bayes $B^\pi(\underline{x})$ devient alors

$$B^\pi(\underline{x}) = \frac{\mathbb{P}(\theta = \theta_0 | \, \underline{x})}{P(\theta = \theta_1 | \, \underline{x})} \frac{\pi(\theta_1)}{\pi(\theta_0)} = \frac{f(\underline{x} | \, \theta_0)}{f(\underline{x} | \, \theta_1)},$$

et ce rapport des vraisemblances, qui intervient aussi dans les tests classiques, ne dépend pas de π. Plus généralement, on peut toujours écrire une loi a priori π sous la forme d'un mélange de lois du type : $\pi(\theta) = \pi_0 g_0(\theta) + \pi_1 g_1(\theta)$, avec $\pi_0 = \pi(\theta \in \Theta_0)$, $\pi_1 = \pi(\theta \in \Theta_1)$,

et g_0 et g_1 deux densités de lois de supports Θ_0 et Θ_1, respectivement. Le facteur de Bayes correspondant s'écrit alors

$$B^\pi(\underline{x}) = \frac{\int_{\Theta_0} \pi(\theta) f(\underline{x}|\,\theta)\, d\theta}{\int_{\Theta_1} \pi(\theta) f(\underline{x}|\,\theta)\, d\theta} \frac{\pi_1}{\pi_0} = \frac{\int_{\Theta_0} f(\underline{x}|\,\theta) g_0(\theta)\, d\theta}{\int_{\Theta_1} f(\underline{x}|\,\theta) g_1(\theta)\, d\theta}\,,$$

et on voit que $B^\pi(\underline{x})$ ne dépend pas de π_0 ni de π_1, l'importance relative accordée par la loi a priori aux régions Θ_0 et Θ_1 (en revanche il dépend de π, à travers les densités g_0 et g_1). On note de plus que la règle de décision associée au coût dissymétrique introduit précédemment peut s'écrire, en fonction de $B^\pi(\underline{x})$:

$$\delta^\pi(\underline{x}) = \begin{cases} 1 & \text{si } B^\pi(\underline{x}) > \frac{a_1}{a_0}\frac{\pi_0}{\pi_1} \\ 0 & \text{sinon} \end{cases}$$

Le facteur de Bayes $B^\pi(\underline{x})$ doit donc être conçu comme un indicateur "universel", pour le test de H_0 contre H_1. Il peut à ce titre servir à plusieurs décideurs, chacun déterminant sa propre règle de décision, en fonction de ses a priori (π_0 et π_1) et de sa fonction de perte (a_0 et a_1). Notons à ce propos que le seuil $(a_1\pi_0)/(a_0\pi_1)$ fait apparaître une dualité entre coûts et poids a priori. Il revient en effet au même de donner un poids a priori identique aux hypothèses H_0 et H_1 ($\pi_0 = pi_1 = 1/2$) et de différencier les coûts a_0 et a_1, ou de prendre les mêmes coûts et d'affecter des poids a priori différents à H_0 et H_1. Cette dualité introduit une souplesse supplémentaire dans la procédure de test. Au-delà de l'optique strictement décisionnelle, le facteur de Bayes est en soi un indicateur graduel de la plausibilité de l'hypothèse nulle, à l'instar de la p-valeur pour les tests fréquentistes.

Exemple 21 *(suite de l'exemple 18) Le problème de Laplace peut s'interpréter comme un test de l'hypothèse $H_0 : p > 1/2$ contre $H_1 : p \leq 1/2$. Le facteur de Bayes associé est alors :*

$$\begin{aligned} B^\pi(x) &= \frac{\mathbb{P}(p > 1/2|\,x)}{\mathbb{P}(p \leq 1/2|\,x)} \frac{\pi(]1/2,\,1])}{\pi([0,\,1/2])} \\ &= \frac{\int_{1/2}^1 p^x(1-p)^{n-x}\, dp}{\int_0^{1/2} p^x(1-p)^{n-x}\, dp} \\ &= 8,7 \times 10^{41} \text{ pour } x = 25\,152\,, \end{aligned}$$

valeur qui privilégie (très nettement !) l'hypothèse H_0 contre H_1.

10.2.1 Test d'une hypothèse ponctuelle

On a vu dans les chapitres précédents que le principe d'un test fréquentiste varie selon la nature de l'hypothèse nulle. Le cas de référence est celui d'une hypothèse nulle simple $H_0 : \theta = \theta_0$ (dite aussi hypothèse *ponctuelle*), pour lequel des résultats d'optimalité sont donnés par le lemme de Neyman-Pearson. Cette optimalité peut être généralisée à d'autres types d'hypothèse, au prix de restrictions et de considérations techniques (parfois complexes !) qui diffèrent au cas par cas. Ainsi, les tests unilatéraux ($H_0 : \theta \leq \theta_0$) requièrent l'hypothèse de monotonie du rapport des vraisemblances. Les tests bilatéraux ($H_0 : \theta \in [\theta_1,\, \theta_2]$) imposent de se restreindre à la classe des tests sans biais. Les tests sur une composante réduisent à nouveau l'ensemble des tests considérés à celui des tests α-semblables, etc.

La Statistique bayésienne en revanche propose une procédure de test universelle, qui convient pour toute hypothèse $H_0 : \theta \in \Theta_0$, où Θ_0 est une partie mesurable de l'espace des paramètres. Les distinctions selon la nature de Θ_0 ne sont alors plus nécessaires. Ainsi, dans l'exemple de Laplace, mener un test de l'hypothèse $H_0 : p \in [1/4, 3/4]$ contre $H_1 : p \in [0, 1/4[\bigcup]3/4, 1]$ (test bilatéral) relèverait de la même démarche que celle employée pour le test unilatéral $H_0 : p > 1/2$ contre $H_1 : p \leq 1/2$, le facteur de Bayes correspondant prenant cette fois pour valeur :

$$\frac{\int_{1/4}^{3/4} p^x(1-p)^{n-x}\,dp}{\int_0^{1/4} p^x(1-p)^{n-x}\,dp \;+\; \int_{3/4}^1 p^x(1-p)^{n-x}\,dp}$$

Les tests d'hypothèse ponctuelle méritent cependant une attention particulière. Si la loi a priori est continue, l'événement $\{\theta = \theta_0\}$ est de mesure nulle, et donc de probabilité (a priori et posteriori) nulle : $\pi(\theta = \theta_0) = \pi(\theta = \theta_0 |\, \underline{x}) = 0$. Une hypothèse ponctuelle devrait donc être systématiquement rejetée dans le cas continu ! [4]

Ce résultat ne doit pas surprendre ; il n'y a effectivement rien de logique à tester la validité d'une hypothèse, à l'aide d'une procédure construite sur l'a priori que cette hypothèse ne sera (presque sûrement) pas vérifiée.

Cette difficulté peut être contournée en modifiant la loi a priori π, afin qu'elle affecte une probabilité π_0 non nulle à l'événement $\{\theta = \theta_0\}$. La densité de la nouvelle loi a priori $\widetilde{\pi}$ s'écrit alors

$$\widetilde{\pi}(\theta) = \pi_0 1_{\{\theta_0\}}(\theta) \;+\; (1-\pi_0)\pi(\theta)$$

La probabilité a posteriori de l'hypothèse nulle prend bien cette fois une valeur non nulle :

$$\pi(\theta = \theta_0 |\, \underline{x}) = \frac{\pi_0 f(\underline{x} |\, \theta_0)}{\int \widetilde{\pi}(\theta) f(\underline{x} |\, \theta)\,d\theta} = \frac{\pi_0 f(\underline{x} |\, \theta_0)}{\pi_0 f(\underline{x} |\, \theta_0) \;+\; (1-\pi_0)\int_\Theta \pi(\theta) f(\underline{x} |\, \theta)\,d\theta}$$

Cette approche a le défaut de particulariser fortement la valeur θ_0 par rapport aux autres valeurs possibles du paramètre. Mais cette remarque s'applique en fait plus généralement à la problématique du test d'hypothèse ponctuelle elle-même. Rappelons la fameuse remarque de I.J. Good (1980) : qui croit raisonnable de tester si la probabilité de pluie pour demain vaut $0,7163891256$? Un nombre fini d'observations n'apporte qu'une information imprécise sur la valeur du paramètre, ce qui est incompatible avec une hypothèse aussi restrictive qu'une hypothèse ponctuelle. Il serait plus judicieux de tester une hypothèse de la forme $H_0 : \theta \in [\theta_0 - \varepsilon,\, \theta_0 + \varepsilon]$, intervalle qui contiendrait tous les points que l'on peut se permettre de confondre avec θ_0, pour le problème considéré.

Cependant, il existe bien des cas où une certaine valeur du paramètre se distingue naturellement des autres. Ainsi, dans un modèle de régression $Y = a + bX + u$ (u résidu centré, de loi normale $\mathcal{N}(0,\, \sigma^2)$ par exemple), l'hypothèse $b = 0$ revient à donner un pouvoir explicatif nul à la variable X. Cette valeur du paramètre joue donc bien un rôle distinct des autres valeurs de b, puisqu'elle modifie la nature du modèle considéré. En d'autres termes, l'hypothèse nulle concerne alors un aspect qualitatif du problème considéré, auquel il est admissible d'affecter une probabilité a priori non nulle. Dans un tel cadre, un test d'hypothèse ponctuelle devient raisonnable.

[4] Notons par ailleurs que le facteur de Bayes n'est plus défini dans un tel cas.

De plus, il est possible de montrer que le test de l'hypothèse ponctuelle $\theta = \theta_0$ mène à des conclusions quasi identiques à ceux relatifs à une hypothèse composée $\theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]$, pour des valeurs faibles de ε.

Enfin, le choix d'une valeur pour π_0, la probabilité a priori de l'événement $\{\theta = \theta_0\}$, peut toujours sembler arbitraire. Or la probabilité à posteriori de l'hypothèse nulle en dépend fortement. On voit à nouveau ici l'intérêt du facteur de Bayes, qui permet d'éliminer cette dépendance :

$$B^\pi(\underline{x}) = \frac{f(\underline{x}|\,\theta_0)\,\pi_0}{\left(\int_\Theta f(\underline{x}|\,\theta)\pi(\theta)\,d\theta\right)(1-\pi_0)}\,\frac{1-\pi_0}{\pi_0} = \frac{f(\underline{x}|\,\theta_0)}{\int_\Theta f(\underline{x}|\,\theta)\pi(\theta)\,d\theta}$$

Comme nous l'avons déjà indiqué, ce facteur de Bayes donne alors un indicateur "objectif" de la plausibilité de l'hypothèse nulle $H_0 : \theta = \theta_0$. Une démarche courante consiste alors à prendre $\pi_0 = 1/2$, et à déterminer la règle de décision en fonction des coûts respectifs des hypothèses nulle et alternative.

10.2.2 Tests et lois a priori impropres

Une loi a priori *impropre* est une loi qui définit sur l'espace Θ une mesure non plus finie, mais simplement σ-finie. La densité correspondante n'est alors pas intégrable :

$$\int_\Theta \pi(\theta)\,d\theta = +\infty$$

Cette approche reste valide tant que la loi a posteriori correspondante définit bien une loi de probabilité sur Θ, ce qui impose

$$\int_\Theta \pi(\theta) f(\underline{x}|\,\theta)\,d\theta < +\infty$$

Exemple 22 *Soit $x \sim \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$, et soit la loi a priori de densité constante sur $\mathbb{R}$, $\pi(\theta) = 1$. La loi a posteriori correspondante est alors bien définie :*

$$\pi(\theta|\,x) \propto \pi(\theta) f(x|\,\theta) \propto (2\pi)^{-\frac{1}{2}} \exp\left\{-(x-\theta)^2/2\right\},$$

d'où $\theta|\,x \sim \mathcal{N}(x, 1)$.

Les lois a priori impropres forment une extension nécessaire des lois a priori propres. Dans un contexte non informatif, elles permettent en général une meilleure modélisation de l'absence d'information a priori, notamment lorsque l'espace des paramètres n'est pas compact. Ainsi, la loi a priori $\pi(\theta) = 1$ de l'exemple précédent, qui n'est autre que la mesure de Lebesgue sur $\mathbb{R}$, peut être vue comme une extension de la loi a priori uniforme proposée par Laplace.

Malheureusement, l'utilisation d'une loi impropre rend impossible le calcul de facteurs de Bayes. En effet, les probabilités a priori des hypothèses nulle et alternative ne sont pas définies pour une telle loi. Il reste cependant possible d'évaluer le rapport des probabilités a posteriori, et donc de mener un test bayésien dans un tel cadre. Mais de tels tests sont moins satisfaisants car, comme nous l'avons déjà indiqué, il n'est alors plus possible de

réduire (ou d'évaluer aisément) l'influence de l'a priori dans la procédure de décision. Ce problème se pose plus particulièrement dans le cas d'une d'hypothèse ponctuelle $H_0 : \theta = \theta_0$ (voir §10.2.1), où la probabilité a priori π_0 de l'événement $\theta = \theta_0$ peut être considérée comme arbitraire (voir paragraphe précédent).

La résolution générale de cette incompatibilité entre tests bayésiens (correspondant à une fonction de coût de type 0-1) et lois a priori impropres reste un problème ouvert, même si plusieurs solutions partielles ont déjà été proposées, via la définition de "pseudo-facteurs de Bayes" notamment, que nous n'aborderons pas ici.

10.3 Critiques et extensions

La très brève présentation des tests bayésiens que nous venons d'effectuer révèle indirectement certains défauts des tests fréquentistes, que nous évoquons dans les parties suivantes.

10.3.1 Tests fréquentistes et dissymétrie des hypothèses

Les tests fréquentistes font jouer un rôle fortement dissymétrique aux hypothèses nulle et alternative. Notamment, le test d'une hypothèse H_0 contre son alternative H_1 n'est pas équivalent au test de H_1 contre H_0 :

Exemple 23 *Soient $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. L'hypothèse $\theta \leq 0$ est rejetée au seuil α si la moyenne des observations $\overline{x}$ vérifie :*

$$\overline{x} > n^{-1/2} u_\alpha,$$

où u_α est le quantile d'ordre $1-\alpha$ d'une loi normale centrée réduite. De même, l'hypothèse $\theta \geq 0$ est rejetée au seuil α' si :

$$\overline{x} < -n^{-1/2} u_{\alpha'}$$

Si $\overline{x} \in [-n^{-1/2} u_{\alpha'},\ n^{-1/2} u_\alpha]$, les deux procédures de tests que nous venons d'exposer mènent à deux conclusions contradictoires (acceptation simultanée des deux hypothèses).

Ce paradoxe tient à la nature distincte des risques de première et seconde espèce : si le premier peut être librement choisi (seuil α), le second est lui indirectement déterminé par le choix de ce seuil. En fait, ce second risque est en général asymptotiquement nul (du moins pour les tests dits *convergents*). L'hypothèse nulle H_0 est donc fortement défavorisée vis-à-vis de l'hypothèse alternative H_1 : lorsque le nombre d'observations tend vers l'infini, la probabilité d'accepter à tort H_0 tend vers 0, mais la probabilité de rejeter à tort H_0 reste constante et égale à α. D'un point de vue décisionnel, cela revient à supposer qu'une erreur de première espèce est infiniment plus coûteuse qu'une erreur de seconde espèce. Les tests bayésiens permettent une pondération plus réaliste entre ces deux coûts (via le rapport a_0/a_1), et conservent la symétrie entre hypothèse nulle et alternative. Un test bayésien conduira donc généralement à des décisions plus favorables à l'hypothèse nulle, comme le montre l'exemple suivant :

Exemple 24 (paradoxe de Lindley-Jeffreys) *Pour le même modèle que dans l'exemple précédent, Le test fréquentiste uniformément plus puissant de l'hypothèse* $H_0 : \theta = 0$ *admet pour région critique* $\{|\, n^{1/2}\overline{x}| > 1.96\}$*) au seuil* 5%*. L'hypothèse nulle est alors systématiquement rejetée dès que* $z_n = n^{1/2}\overline{x} = 1,97$*, quelle que soit la valeur de* n*. En revanche, pour une loi a priori* $\pi \sim \mathcal{N}(0,\, 1)$*, la probabilité a posteriori de* H_0 *s'écrit :*

$$\pi(\theta = 0|\, z_n) = 1 \,+\, \frac{1-\pi_0}{\pi_0}(n+1)^{-1/2}\exp\left(z_n^{2n/2(n+1)}\right)$$

et, pour $z_n = 1,97$*, cette probabilité tend vers* 1 *lorsque* n *tend vers l'infini, quelle que soit la probabilité a priori* π_0 *de l'hypothèse nulle.*

10.3.2 Justification asymptotique contre raisonnement conditionnel

La validité des tests fréquentistes repose généralement sur des résultats asymptotiques, c'est-à-dire lorsque le nombre d'observations tend vers l'infini. Dans la pratique, il est difficile, voire impossible, de mesurer à quel point l'échantillon considéré, de taille finie, s'écarte du comportement prévu par l'asymptotique. Pour les tests notamment, les valeurs réelles des risques de première et seconde espèce, pour le nombre n d'observations considérées, ne peut en général être évalué autrement que par simulation. On obtient alors des valeurs de risque *moyennes*, pour tous les échantillons possibles de taille n, et non pas une mesure pour l'échantillon étudié seul. Seul un raisonnement conditionnel (en les observations), tel qu'il est mis en œuvre en Statistique bayésienne, permet une évaluation *pour les observations considérées*, et non en moyenne.

10.3.3 Le principe de vraisemblance

Considérons l'exemple suivant.

Exemple 25 *Un physicien, que nous appelerons Guillaume, mesure à l'aide d'un appareil sophistiqué trois impacts aléatoires de particules sur un axe horizontal. Les abscisses correspondantes* $(x_1, x_2, x_3) = (1,7,\, -0,2,\, 1,65)$ *suivent une loi normale* $\mathcal{N}(\theta,\, 1)$*, et Guillaume veut vérifier si l'hypothèse que les impacts soient centrés (*$H_0 : \theta = 0$*) est acceptable. Un test de cette hypothèse au niveau* 5% *admet pour région critique* $W = \{|\,\overline{x}| > 1,96/\sqrt{3}\}$*. L'hypothèse nulle est donc acceptée (*$\overline{x} = 1,05$ *et* $1,96/\sqrt{3} = 1,13$*). Ce résultat obtenu, Guillaume apprend que l'appareil utilisé est défaillant : les abscisses* x *telles que* $|\, x| > 2$ *ne sont pas correctement mesurées, et sont remplacées par la valeur "saturée"* ± 2 *(selon le signe de* x*). Les observations obtenues ne sont donc pas tirées selon une loi normale* $\mathcal{N}(\theta,\, 1)$*, mais selon une loi normale tronquée, de densité :*

$$f(x|\,\theta) = \frac{\exp\{(x-\theta)^2\}}{\int_{-2}^{2}\exp\{(x-\theta)^2\}dx}\,1_{[-2,\,2]}(x)$$

Guillaume recommence ces calculs, et trouve pour nouvelle région critique $W = \{|\,\overline{x}| > 1,68/\sqrt{3}\}$*, toujours à* 5%*. L'hypothèse nulle* $\theta = 0$ *doit cette fois être rejetée (*$1.68/\sqrt{3} = 0,96$*). Cependant, Guillaume se dit que le résultat de son expérience aurait été le même avec un appareil en parfait état, puisqu'aucune des observations n'est sortie de l'intervalle*

$[-2, 2]$. *Dès lors, pourquoi devoir tenir compte d'une panne qui n'a pas eu lieu ? Guillaume ne sait plus alors s'il doit se tenir à sa seconde décision et rejeter l'hypothèse nulle, ou revenir à la première décision, et ne pas la rejeter.*

Ce paradoxe, et le dilemne de Guillaume qui en découle, provient du fait que les tests fréquentistes ne respectent pas le principe de vraisemblance. Ce principe édicte que l'information que l'on tire d'une expérience $x \sim f(x|\,\theta)$ ne doit dépendre que de la fonction $\theta \mapsto f(x|\,\theta)$, c'est-à-dire la vraisemblance au point x correspondant à la valeur effectivement observée. De plus, deux expériences x_1 et x_2 dont les vraisemblances ne différent que d'une constante multiplicative (i.e. $f(x_1|\,\theta) = cf(x_2|\,\theta)$, pour tout θ) doivent mener aux mêmes conclusions.

Dans l'exemple précédent, la loi de l'observée x varie selon que l'on prend ou non en compte la défaillance de l'appareil (ce qui modifie la région critique), mais la vraisemblance au point x reste toujours la même (à une constante près), que x soit tiré d'une loi tronquée ou non. La défaillance de l'appareil ne devrait donc effectivement pas changer la règle de décision.

L'analyse bayésienne quant à elle respecte le principe de vraisemblance, car elle fonde son raisonnement sur la loi a posteriori, qui dépend bien uniquement (en $\underline{x}$) de la vraisemblance $f(\underline{x}|\,\theta)$.

10.3.4 Conclusion

Sur tous les points évoqués dans les parties précédentes, l'approche bayésienne nous semble préférable. Cependant, comme on a pu le voir, les facteurs de Bayes ne représentent pas non plus une solution entièrement satisfaisante, notamment parce qu'ils ne sont pas compatibles avec les lois a priori impropres (voir §10.2.2).

Nous espérons en tout cas que ce chapitre aura rendu le lecteur curieux de Statistique bayésienne, et convaincu de l'intérêt de cette approche "alternative". Pour compléter cette introduction, nous conseillons la lecture de Robert, C.P., L'analyse statistique bayésienne, Economica, 1992, ouvrage de référence (francophone) dont nous nous sommes largement inspirés. On pourra lire aussi Robert, C.P., The Bayesian choice, Springer-Verlag, 2000, la version anglaise, plus récente et plus complète, Berger, J.O., Statistical decision theory and Bayesian analysis, Springer-Verlag, 1980, ou Schervisch, M.J., Theory of Statistics, Springer-Verlag, 1995, pour une approche plus théorique.

Nicolas CHOPIN

Eléments bibliographiques pour ce chapitre

Berger, J. O. (1980) *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag : New York, Berlin, Heidelberg.

Good, I. J. (1980) *Some history of the hierarchical Bayesian methodology.* In : *Bayesian Statistics* ***2***, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A.F.M Smith (eds.). North-Holland : Amsterdam.

Pratt, J. W. Discussion of A. Birnbaum's "On the foundations of statistical inference". *Journal of the American Statistical Association* **57**.

Robert, C. P. (1992) *L'analyse statistique bayésienne*. Economica : Paris.

Robert, C. P. (2000) *The Bayesian Choice*. Springer-Verlag : New York, Berlin, Heidelberg.

Schervisch, M. J. (1995) *Theory of Statistics*. Springer-Verlag : New York, Berlin, Heidelberg.

Ulmo, J. & Bernier, J. (1973) *Eléments de décision statistique*. Presses Universitaires de France : Paris.

Annexes

Annexe A

Caractérisation de certaines lois

A.1 Lois discrètes

Loi	Paramètres	Probabilité	Moyenne	Variance
Uniforme		$p(k) = \frac{1}{n}\mathbb{1}_{[\![0,n]\!]}(k)$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Bernoulli	$0 \leq p \leq 1$	$p(k) = p \quad k=1$ $= q \quad k=0$ $= 0 \quad$ sinon	p	$p.q$
Binomiale	$n = 1, 2, \ldots$ $0 \leq p \leq 1$	$p(k) = C_n^k p^k q^{n-k} \mathbb{1}_{[\![0,n]\!]}(k)$	np	npq
Hypergéométrique	$N = 1, 2, \ldots$ $n = 1, 2, \ldots, N$ $p = 0, \frac{1}{N}, \frac{2}{N}, \ldots, 1$	$p(k) = \frac{C_{Np}^k C_{Nq}^{n-k}}{C_N^n} \mathbb{1}_{[\![0,n]\!]}(x)$	np	$npq\left(\frac{N-n}{N-1}\right)$
Géométrique	$0 \leq p \leq 1$	$p(k) = q^{k-1}p, \ k \geq 1$	$\frac{1}{p}$	$\frac{q}{p^2}$
Poisson	$\lambda > 0$	$p(k) = e^{-\lambda}\frac{\lambda^k}{k!} 1\mathrm{I}_{[\![0,N]\!]}(k)$	λ	λ

A.2 Lois continues

Loi	Paramètres	Densité de probabilité $f(.)$	moyenne	variance
Uniforme	$-\infty < a < b < +\infty$	$f(x) = \frac{1}{b-a}\mathbb{1}_{]a,b[}(x)$	$\frac{a+b}{2}$	$\frac{b^2-a^2}{12}$
Normale	$-\infty < \mu < +\infty$ $\sigma > 0$	$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2
Exponentielle	$\lambda > 0$	$f(x) = \lambda e^{-\lambda x}\mathbb{1}_{\mathbb{R}_+^*}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$r > 0, \lambda > 0$	$f(x) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1} e^{-\lambda x}\mathbb{1}_{\mathbb{R}_+^*}(x)$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$

Annexe B

Tables

B.1 Table de la loi normale $\mathcal{N}(0,\,1)$

Si X suit une loi normale de moyenne μ et de variance σ^2, $X \sim \mathcal{N}(\mu,\,\sigma^2)$, on écrit la densité

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty,$$

et la fonction de répartition est notée $\Phi(x) = \int_{-\infty}^{x} \varphi(t)\ dt$.

B.2 Table de quantiles de χ^2_ν $(1 \leq \nu \leq 10)$

B.3 Table de quantiles des lois de Student $t_{\nu,\,p}$ $(1 \leq \nu \leq 30)$

x	0	1	2	3	4	5	6	7	8	9
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	0,7291	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	0,8413	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	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	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	0,9370	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	0,9772	0,9778	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	0,988l	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	0,9331	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
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

TABLEAU B.1 – Table de la fonction $\Phi(x)$ $(x \geq 0)$.

$\nu \backslash p$	0,995	0,99	0,975	0,95	0,9
1	7,88	6,63	5,02	3,84	2,71
2	10,60	9,21	7,38	5,99	4,61
3	12,60	11,30	9,35	7,81	6,25
4	14,90	13,30	11,10	9,49	7,78
5	16,70	15,10	12,80	11,10	9,24
6	18,50	16,80	14,40	12,60	10,60
7	20,30	18,50	16,00	14,10	12,00
8	22,00	20,10	17,50	15,50	13,40
9	23,60	21,70	19,00	16,90	14,70
10	25,20	23,20	20,50	18,30	16,00

TABLEAU B.2 – Table des valeurs $\chi^2_{\nu,\,p}$ définies par $\mathbb{P}\left(0 \leq X \leq \chi^2_{\nu,\,p}\right) = p$.

$\nu \backslash p$	0,9	0,95	0,975	0,99	0,995
1	3,076	6,314	12,706	31,621	63,657
2	1,686	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,941	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,716	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,896
18	1,330	1,734	2,101	2,552	2,878
19	1,326	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,516	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
inf.	1,282	1,645	1,960	2,326	2,576

TABLEAU B.3 – Table des valeurs t_ν définies par $\mathbb{P}\left(-\infty < X \leq t_{\nu,\,p}\right) = p$.

Annexe C

Décomposition spectrale des vecteurs gaussiens

Soit $\underline{Y}$ un vecteur aléatoire de $\mathbb{R}^d$ ($d \geq 1$) (vecteur colonne). On note

$$\langle \underline{x}, \underline{y} \rangle = \sum_1^d x_i y_i$$

le produit scalaire canonique de $\mathbb{R}^d$, pour lequel la base canonique est orthonormée.

On suppose que $\underline{Y}$ est d'ordre 2, d'espérance 0 et de matrice variance Σ. Si on identifie (*via* la base canonique), matrice $d \times d$ et endomorphisme de $\mathbb{R}^d$, on a :

$$\begin{aligned} \forall \underline{x} \quad &\in \quad \mathbb{R}^d, \quad \mathbb{E}\left(\langle \underline{Y}, \underline{x} \rangle\right) = 0 \\ \forall \underline{x} \text{ et } \underline{y} \quad &\in \quad \mathbb{R}^d, \quad \mathbb{E}\left(\langle \underline{Y}, \underline{x} \rangle \langle \underline{Y}, \underline{y} \rangle\right) = \langle \Sigma \underline{x}, \underline{y} \rangle \end{aligned}$$

Soit P une matrice $d \times d$. Soit $\underline{Z} = P\underline{Y}$. Alors $\underline{Z}$ est un vecteur aléatoire de $\mathbb{R}^d$, centré et de matrice variance

$$\mathbb{E}(\underline{Z}\underline{Z}^T) = P\mathbb{E}(\underline{Y}\underline{Y}^T)P^T = P\Sigma P^T$$

Si l'on choisit pour P une matrice orthogonale telle que $P\Sigma P^T = P\Sigma P^{-1} = \Lambda = \text{diag}\ (\lambda_1, \ldots, \lambda_d)$ (matrice diagonale), c'est-à-dire si les vecteurs colonnes de $P^T = P^{-1}$ forment une base orthonormée de vecteurs propres de Σ (il est possible de trouver une telle matrice P parce que Σ est symétrique réelle), en sélectionnant $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ (les valeurs propres de Σ sont ≥ 0 car Σ est semi-définie positive), alors, si l'on note $\underline{u}_1, \ldots, \underline{u}_d$ ces vecteurs colonnes, on a les résultats suivants :

Proposition 1 *Les v.a.* $Z_j \overset{def}{=} \langle \underline{Y}, \underline{u}_j \rangle$ $(= \underline{u}_j^T \underline{Y} = \underline{Y}^T \underline{u}_j$ *en notation matricielle),* $1 \leq j \leq d$, *sont non corrélées deux-à-deux, et* Z_j *est centrée et de variance* λ_j *:*

$$\begin{aligned} \mathbb{E}(Z_j^2) &= \mathbb{E}\left(\langle \underline{Y}, \underline{u}_j \rangle^2\right) \\ &= \langle \Sigma \underline{u}_j, \underline{u}_j \rangle \\ &= \langle \lambda_j \underline{u}_j, \underline{u}_j \rangle \\ &= \lambda_j, \end{aligned}$$

puisque $\|\underline{u}_j\|^2 = \langle \underline{u}_j, \underline{u}_j \rangle = 1$.

Définition C.1 *Les v.a.* Z_j *sont les composantes principales de* $\underline{Y}$.

1. *Reconstitution de* $\underline{Y}$ *à partir des* Z_j
   - On a :

$$\begin{aligned} \forall \underline{x} \in \mathbb{R}^d, \quad \Sigma \underline{x} &= P^T \Lambda P x \\ &= \sum_1^d \lambda_j \langle \underline{u}_j, \underline{x} \rangle \underline{u}_j \\ &= \sum_1^d \lambda_j \quad \underline{u}_j (\underline{u}_j^T \underline{x}) \\ &= \sum_1^d \lambda_j (\underline{u}_j \underline{u}_j^T) \underline{x}, \end{aligned}$$

     soit, matriciellement :

$$\Sigma = \sum_1^d \lambda_j \quad \underline{u}_j \underline{u}_j^T$$

   - On reconstitue $\underline{Y}$ ainsi :

$$\begin{aligned} \underline{Y} &= \sum_{j=1}^d \langle \underline{Y}, \underline{u}_j \rangle \underline{u}_j \\ &= \sum_{j=1}^d Z_j \quad \underline{u}_j \\ &= \sum_{j=1}^d \sqrt{\lambda_j} Z_j^* \underline{u}_j, \end{aligned}$$

     où $Z_j^* = Z_j / \sqrt{\lambda_j}$ (si $\lambda_j > 0$) et $Z_j^* = 0$ (si $\lambda_j = 0$). Si $\lambda_j > 0$, Z_j^* a pour espérance 0 et pour variance 1 ; d'autre part, les v.a. Z_j^* sont non corrélées 2 à 2 :

$$\forall 1 \leq j \neq j' \leq d, \quad \mathbb{E}(Z_j^* Z_{j'}^*) = 0$$

– Par conséquent :

$$\|\underline{Y}\|^2 = \underline{Y}^T \underline{Y} = \sum_{j=1}^{d} \lambda_j Z_j^{*2}$$

Dans le cas normal où $\underline{Y} \sim \mathcal{N}(0, \Sigma)$, les v.a. Z_j^* sont *indépendantes* et $\mathcal{N}(0, 1)$ *si* $\lambda_j > 0$ (avec $Z_j^* = 0$ sinon). Ainsi, si le rang de Σ est K, alors la v.a. $\sum_{j=1}^{d} Z_j^{*2}$ suit une loi χ^2_K.

2. Enfin, la décomposition $\underline{Y} = \sum_{j=1}^{d} \sqrt{\lambda_j}\ Z_j^*\ \underline{u}_j$ présente des *caractères d'optimalité* en termes de variance. Cela est lié aux extrema (liés) de la norme sur un ellipsoïde. Intuitivement, pour chaque entier $1 \leq J \leq d$ la v.a. $\sum_{j=1}^{J} \sqrt{\lambda_j}\ Z_j^*\ \underline{u}_j$ est la v.a. J-variée qui résume le mieux (en un sens lié au second ordre) la v.a. $\underline{Y}$ et sa loi. On retrouve ici les idées de l'Analyse en composantes principales.

Annexe D

Convergence en distribution et processus

Pour cette partie, on pourra consulter pour commencer les ouvrages suivants : Billingsley (1968) et Dudley (1989). Voir la bibliographie à la fin de cette annexe.

D.1 Convergence en distribution abstraite

Définition D.1 – (M, δ) *espace métrique complet séparable,*
– $\mathcal{M}$ *tribu borélienne* $\Rightarrow$ *séparable,*
– $\mathbb{X}_n$ *suite de v.a. à valeurs* $(M, \mathcal{M})$*,*
– $\mu_{\mathbb{X}_n}$ = *loi de* $\mathbb{X}_n$ = *mesure de probabilité sur* $(M, \mathcal{M})$*.*

1. *On dit que*

$$\begin{cases} \mathbb{X}_n \xrightarrow{d} \mathbb{X} \text{ ou} \\ \mathbb{X}_n \Rightarrow \mathbb{X} \text{ (en distribution)} \end{cases} \quad \text{si } \mu_{\mathbb{X}_n} \overset{\text{étroitement}}{\longrightarrow} \mu_{\mathbb{X}}$$

c'est-à-dire $(\forall f : M \to \mathbb{R}$ *continue bornée),*

$$\mathbb{E}f(\mathbb{X}_n) = \int_M f \, d\mu_{\mathbb{X}_n} \longrightarrow \mathbb{E}f(\mathbb{X}) = \int_M f \, d\mu_{\mathbb{X}}$$

2. *On dit que* $\mathcal{P}$*, ensemble de mesures de probabilité sur* $(M, \mathcal{M})$*, est faiblement (ou étroitement) relativement compacte si de toute suite d'éléments de* $\mathcal{P}$ *on peut extraire une sous-suite qui converge faiblement (ou étroitement).*

Théorème D.1 (Critère de tension de Prohorov) *Le sous-ensemble $\mathcal{P}$ est relativement compact pour la topologie de la convergence étroite si et seulement si il est tendu :*

$$\forall \varepsilon > 0,\ \exists K \text{ compact de } (\mathcal{M},\ \delta),\ \forall \mu \in \mathcal{P},\ \mu(K) \geq 1 - \varepsilon$$

D.2 Représentation de Skorohod

Théorème D.2 *Si $\mathbb{X}_n \xrightarrow{d} \mathbb{X}$ à valeurs dans $(M, \mathcal{M},\ \delta)$ séparable, alors il existe un espace $(\Omega^*, \mathcal{A}^*, P^*)$ et des copies $\mathbb{X}_n^*$ des $\mathbb{X}_n$ et $\mathbb{X}^*$ de $\mathbb{X}$, définies sur cet espace, telles que*

$$\mathbb{X}_n^* \to \mathbb{X}^* \text{ p.s.}$$

Exemple 26 (dans un cas simple) *$M = [0,\ 1]$, $\delta(x, y) = |x - y|$, $\mathcal{M}$. Soit G_n la f.r. de $\mathbb{X}_n \in [0,\ 1]$ et supposons que $\mathbb{X} \sim \mathcal{U}\,[0,\ 1]$. Supposons pour simplifier que G_n est continue strictement croissante sur $[0,\ 1]$. Posons :*

$$\left.\begin{array}{ll} \mathbb{X}_n^* = G_n^{-1}(u) & u \in]0,\ 1[\\ \mathbb{X}^* = u & u \in]0,\ 1[\end{array}\right\} \begin{array}{l} \Omega^* =]0,\ 1[\\ \mathcal{A}^* = \text{boréliens} \\ P^* = \text{Lebesgue} \end{array}$$

Montrons que $G_n^{-1}(u) \longrightarrow u$ pour presque tout $u \in]0,\ 1[$. Soit $u \in]0,\ 1[$ et soit $\varepsilon > 0$ assez petit pour que $0 < u - \varepsilon < u < u + \varepsilon < u$. On sait que $G_n(u \pm \varepsilon) \longrightarrow u \pm \varepsilon$. Donc, pour n assez grand,

$$G_n(u - \varepsilon) \leq u \leq G_n(u + \varepsilon)$$

D'où

$$u - \varepsilon \leq G_n^{-1}(u) \leq u + \varepsilon$$

Ainsi,

$$\left|G_n^{-1}(u) - u\right| \leq \varepsilon$$

pour n assez grand. ■

Corollaire 6 *Si $\psi : (M, \delta) \to \mathbb{R}$ est continue et si $\mathbb{X}_n \xrightarrow{d} \mathbb{X}$, alors $\psi(\mathbb{X}_n) \xrightarrow{d} \psi(\mathbb{X})$.*

Principe : On passe par $\mathbb{X}_n^* \to \mathbb{X}^*$ p.s., d'où $\psi(\mathbb{X}_n^*) \to \psi(\mathbb{X}^*)$ p.s., donc en distribution, or

$$\psi(\mathbb{X}_n) \overset{d}{=} \psi(\mathbb{X}_n^*) \quad \text{et} \quad \psi(\mathbb{X}) \overset{d}{=} \psi(\mathbb{X}^*) \tag{D.1}$$

■

Applications :

Si $\psi : \left.\begin{array}{l} D[0,1] \\ C[0,1] \end{array}\right\} \to \mathbb{R}$ continue $\left\{\begin{array}{l} d \\ \|.\|_\infty \end{array}\right.$.

Si $\mathbb{X}_n \xrightarrow{d} \mathbb{X}$, alors $\psi(\mathbb{X}_n) \xrightarrow{d} \psi(\mathbb{X})$.

Par exemple,

$$\begin{aligned} \psi(\mathbb{B}_n) &= \sup_{[0,1]} |\mathbb{B}_n| \xrightarrow{d} \sup_{[0,1]} |\mathbb{B}| \\ \psi(\mathbb{B}_n) &= \int_0^1 \mathbb{B}_n^2 w \quad dt \xrightarrow{d} \int_0^1 \mathbb{B}^2 w \quad dt \end{aligned}$$

D.3 Convergence en distribution des suites de processus continus[1]

- $C = C[0,1]$: fonctions réelles continues sur $[0,1]$; espace muni de la norme de la convergence uniforme.
- $D = D[0,1]$: fonctions réelles admettant en tout point une limite à gauche et une limite à droite, et bornées ; espace muni d'une topologie (dite de Skorohod), proche de la topologie de la convergence uniforme, mais ayant l'avantage d'en faire un espace métrique *séparable*. Voir, par exemple, l'ouvrage de Billingsley (1968).

Pour des suites $(\mathbb{X}_n)$ de processus (à valeurs dans C ou dans D), les convergences finidimensionnelles de $\mathbb{X}_n$, c'est-à-dire la convergence en distribution de tout vecteur aléatoire de la forme

$$(\mathbb{X}_n(t_1), \ldots, \mathbb{X}_n(t_k))^T,$$

où $0 \leq t_1 < \ldots < t_k \leq 1$, $k \in \mathbb{N}^*$, vers des lois limites, *ne suffisent pas* à garantir que :

$$\exists \mathbb{X} \text{ processus (à valeurs dans } C \text{ ou } D) : \ \mathbb{X}_n \xrightarrow{d} \mathbb{X}$$

En effet, ces lois limites pourraient être les projections (sur chaque $\mathbb{R}^k$ *via* $\underline{x} \mapsto (\underline{x}(t_1), \ldots, \underline{x}(t_k))$) d'une mesure de probabilité dont le support serait strictement plus grand que C ou D – par exemple un espace de Sobolev ou un espace de distributions.

Par contre, cela est suffisant *si, en outre,* la suite des mesures $\mu_{\mathbb{X}_n}$ (donc à support dans C ou dans D), est *tendue* (voir critère de Proborov ci-dessus).

Il existe, pour C ou pour D, des critères utilisables de tension.

C'est ainsi qu'on démontre que $\mathbb{S}_n \xrightarrow{d} \mathbb{W}$ ($\mathbb{S}_n$ processus des sommes partielles de v.a. i.i.d. centrées réduites, $\mathbb{W}$ processus de Wiener), et que $\mathbb{B}_n \xrightarrow{d} \mathbb{B}$ ($\mathbb{B}$ pont brownien).

Eléments bibliographiques

[1] ou continus par morceaux.

Billingsley, P. (1968) *Convergence of Probability Measures*. Wiley : New York.

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 1. Masson : Paris, 2e édition.

Dacunha-Castelle D. & Duflo M. (1994) *Probabilités et Statistiques*, Tome 2. Masson : Paris, 2e édition.

Dudley, R. M. (1989) *Real Analysis and Probability*, Mathematics Series. Chapman and Hall : New York and London.

Annexe E

Inverse généralisée d'une fonction de répartition

E.1 Résultats sur $F^{\leftarrow}$

Soit $F(x) = \mathbb{P}\{X \leq x\}$ une fonction de répartition (il s'agit de la version continue à droite). On définit l'inverse généralisée $F^{\leftarrow}$ de F comme la fonction croissante au sens large et continue à gauche

$$F^{\leftarrow}(t) = \inf\{x : F(x) \geq t\}, \quad 0 < t < 1.$$

Voici les propriétés élémentaires de $F^{\leftarrow}$:

(a)
$$F \circ F^{\leftarrow}(t) \geq t \quad \text{pour tout } \ 0 < t < 1.$$

Démonstration : Par définition, et croissance (sens large) et continuité à droite de F, pour tout $0 < t < 1$, l'ensemble des x tels que $F(x) \geq t$ est l'intervalle $[F^{\leftarrow}(t), +\infty[$. Par suite, $F \circ F^{\leftarrow}(t) \geq t$. ■

(b) $F(x) \geq t$ si et seulement si $F^{\leftarrow}(t) \leq x$.

Démonstration : C'est une conséquence directe de ce que l'ensemble des x tels que $F(x) \geq t$ est l'intervalle $[F^{\leftarrow}(t), +\infty[$. ■

(c) $F(x) < t$ si et seulement si $F^{\leftarrow}(t) > x$.

(d) $F(x_1) < t \leq F(x_2)$ si et seulement si $x_1 < F^{\leftarrow}(t) \leq x_2$.

(e) L'égalité $F \circ F^{\leftarrow}(t) = t$ dans (a) a lieu si et seulement si t appartient à l'image de F.

Démonstration : On a vu que l'ensemble des x tels que $F(x) \geq t$ est l'intervalle $[F^{\leftarrow}(t), +\infty[$. Si $F \circ F^{\leftarrow}(t) = t$, alors t appartient à l'image de F. Si $F \circ F^{\leftarrow}(t) > t$, comme $F(x) < t$ pour tout $x < F^{\leftarrow}(t)$, donc t n'appartient pas à l'image de F. ■

(f)
$$F^{\leftarrow} \circ F(x) \leq x \quad \text{pour tout } x.$$

L'inégalité $F^{\leftarrow} \circ F(x) < x$ a lieu si et seulement si existe $\varepsilon > 0$ tel que $F(x-\varepsilon) = F(x)$.

Démonstration : Comme $F(x) \geq t_0 = F(x)$, on a $F^{\leftarrow}(t_0) \leq x$, d'où $F^{\leftarrow} \circ F(x) \leq x$. S'il existe $\varepsilon > 0$ tel que $F(x-\varepsilon) = F(x)$, alors l'ensemble des y tels que $F(y) \geq F(x)$ contient $x - \varepsilon$, donc $F^{\leftarrow} \circ F(x) \leq x - \varepsilon < x$. Si $F^{\leftarrow} \circ F(x) = x$, alors, si on pose $t_0 = F(x)$, x est le plus petit élément de l'ensemble des y tels que $F(y) \geq t_0$. Il est donc alors impossible qu'existe $\varepsilon > 0$ tel que $F(x-\varepsilon) = t_0$. ■

(g) Si X est une variable aléatoire dont la loi admet F pour fonction de répartition, on a

$$\mathbb{P}\{F(X) \leq t\} \leq t \quad \text{pour tout } 0 < t < 1.$$

Démonstration : D'après (b), si U est une variable aléatoire uniforme sur $[0,1]$, alors la loi de la variable aléatoire $Y = F^{\leftarrow}(U)$ admet F pour fonction de répartition, donc Y a même loi que X. D'après (a), $F(Y) \geq U$ p.s., par suite

$$\mathbb{P}\{F(X) \leq t\} = \mathbb{P}\{F(Y) \leq t\} \leq \mathbb{P}\{U \leq t\} = t.$$

Si F est continue, alors l'image de F est un des intervalles (0, 1), donc d'après (e) $F(X)$ est alors uniforme sur [0, 1]. ■

On en déduit le résultat suivant, qui permet en particulier de simuler des variables aléatoires de fonction de répartition F à partir de variables aléatoires uniformes sur [0, 1].

Proposition E.1 *Soit F une fonction de répartition.*

(i) Si U est une variable aléatoire $\mathcal{U}[0,1]$, alors la loi de la variable aléatoire $F^{\leftarrow}(U)$ admet F pour fonction de répartition.

(ii) Soit X une variable aléatoire dont la loi admet F pour fonction de répartition. Si F est continue, alors la variable aléatoire $F(X) \sim \mathcal{U}[0,1]$.

Remarquons que si $F(X)$ est uniforme sur [0, 1], alors $\mathbb{P}\{F(Y) \leq t\} = t$ pour tout $0 < t < 1$, donc $F \circ F^{\leftarrow}(U) = U$ p.s., donc d'après (e) presque tout $0 < t < 1$ appartient à l'image de F, et par suite l'adhérence de l'image de F est alors [0, 1].

E.2 Représentation de Skorohod, suite

Notons maintenant $H^- = F^{\leftarrow}$ et

$$H^+(t) = \inf\{x : F(x) > t\}, \quad 0 < t < 1.$$

On a donc $H^-(t) \leq H^+(t)$ pour tout $0 < t < 1$. On démontre de manière analogue à la Proposition ci-dessus que si U est une variable aléatoire $\mathcal{U}[0,1]$, alors la loi de la variable aléatoire $H^+(U)$ admet F pour fonction de répartition. On démontre alors que la mesure de

Lebesgue de l'ensemble des $0 < t < 1$ tels que $H^-(t) = H^+(t)$ est égale à 1. On en déduit le résultat remarquable suivant, qui constitue un cas particulier du théorème de représentation de Skorohod.

Proposition E.2 *Soit (X_n) une suite de variables aléatoires réelles qui converge en loi vers une variable aléatoire réelle X lorsque $n \to \infty$. Il existe alors un espace probabilisé $(\Omega^\star, \mathcal{A}^\star, P^\star)$ et, sur cet espace, des copies $X_n^\star$ de X_n et $X^\star$ de X telles que $X_n^\star$ converge presque sûrement vers $X^\star$ lorsque $n \to \infty$.*

Démonstration : Prendre $\Omega^\star = [0, 1]$, $\mathcal{A}^\star =$ tribu borélienne de $[0, 1]$, et $P^\star =$ mesure de Lebesgue sur $[0, 1]$, puis $X_n^+(t) = H_n^+(t)$, $X_n^-(t) = H_n^-(t)$, $X^+(t) = H^+(t)$ et $X^-(t) = H^-(t)$. Soit $0 < t < 1$ fixé, et soit x un point de continuité de la fonction de répartition F de la variable aléatoire X, tel que $x > H^+(t)$. Par définition de H^+, on a alors $F(x) > t$. Comme $F_n(x) \to F(x)$ quand $n \to \infty$ (puisque x est un point de continuité de F, et que X_n, de fonction de répartition F_n, converge en loi vers X), on a aussi $F_n(x) > t$ pour tout n assez grand. Donc $X_n^+(t) \leq x$ pour tout n assez grand, et par suite $\limsup_{n\to\infty} X_n^+(t) \leq x$. Comme l'ensemble des points de continuité de F est dense dans l'ensemble des réels (puisque $F \uparrow$), on peut faire $x \downarrow X^+(t)$, x restant point de continuité de F. Par conséquent,

$$\limsup_{n\to\infty} X_n^+(t) \leq X^+(t).$$

De même,

$$\liminf_{n\to\infty} X_n^-(t) \geq X^-(t).$$

Comme $X_n^- \leq X_n^+$, dès que $X^+(t) = X^-(t)$ il vient donc $\liminf_{n\to\infty} X_n^-(t) = \limsup_{n\to\infty} X_n^-(t)$, et de même pour $X_n^+(t)$, avec pour limite commune $X^+(t) = X^-(t)$. Comme la mesure de Lebesgue de l'ensemble des $0 < t < 1$ tels que $X^+(t) = X^-(t)$ est égale à 1, cette situation a lieu presque sûrement. Finalement, il suffit donc de poser, par exemple, $X_n^\star = X_n^-$ et $X^\star = X^-$ pour conclure. ■