Notes de cours - Stat 2

Table des matières

1 Tests et régions de confiance **7**
- 1.1 Modèle statistique . . . 7
- 1.2 Tests statistiques . . . 7
- 1.3 Dissymétrie des hypothèses . . . 8
- 1.4 p-valeur . . . 10
- 1.5 Régions de confiance . . . 11
- 1.6 Dualité entre régions de confiance et tests d'hypothèse nulle simple 12
- 1.7 Construction de tests : méthode du pivot . . . 13

2 Tests asymptotiques **15**
- 2.1 Quelques résultats de probabilités . . . 16
  - 2.1.1 Consistance . . . 16
  - 2.1.2 Normalité asymptotique . . . 17
  - 2.1.3 Estimateurs du maximum de vraisemblance . . . 18
  - 2.1.4 Quantiles empiriques . . . 20
- 2.2 Tests d'adéquation . . . 20
  - 2.2.1 Un exemple dans un modèle de translation . . . 20
  - 2.2.2 Test du χ^2 . . . 21
- 2.3 Test de Kolmogorov Smirnov . . . 23
- 2.4 Tests du rapport de vraisemblance . . . 24

3 Concentration de la moyenne empirique **25**
- 3.1 Méthode de Chernoff . . . 25
- 3.2 Approches génériques . . . 28
  - 3.2.1 Variables aléatoires sous Gaussiennes . . . 29
  - 3.2.2 Variables aléatoires sous Poissonniennes . . . 32
  - 3.2.3 Inégalité de Bernstein . . . 35
- 3.3 Applications à la construction de tests . . . 37
  - 3.3.1 Modèle de translation . . . 37
  - 3.3.2 Modèle de translation et d'échelle . . . 38

4 Tests optimaux **39**
- 4.1 Tests d'hypothèses simples . . . 39
- 4.2 Tests unilatères . . . 41
- 4.3 Tests bilatères . . . 43

5 Théorie de la décision **47**
5.1 Règle de décision, perte, risque 47
5.2 Approche minimax 48
5.3 Approche Bayésienne 50

6 L'algorithme EM **53**
6.1 Cadre général 54
6.2 L'algorithme 54
6.3 Exemple 1 55
6.4 Exemple 2 56

7 Médiane empirique et test du signe **59**
7.1 Le modèle de translation 59
7.2 Test du signe 60
7.3 Efficacité asymptotique relative 62

8 Test du rang de Wilcoxon **67**
8.1 Présentation et premiers résultats 67
8.2 Pente 69

9 Modèles linéaires **71**
9.1 Regression linéaire 71
9.1.1 Estimation par les moindres carrés 71
9.1.2 Régression linéaire Gaussienne 72

10 Tests multiples **75**
10.1 Cadre général 76
10.2 La procédure de Bonferroni 77
10.3 La procédure de Holm 77
10.4 Méthodes récursives "step-down" 78
10.5 FDR 79
10.5.1 Heuristique 79
10.5.2 Résultat général 80

11 Statistiques robustes **83**
11.1 Courbe de sensibilité 83
11.2 Fonction d'influence 84
11.3 Breakdown point d'un estimateur 85
11.4 Estimateurs de Huber 86
11.4.1 Fonctionnelle de Huber 87
11.4.2 Asymptotique de l'estimateur de Huber 87
11.4.3 Concentration de l'estimateur de Huber 88

Notations

Les vecteurs $\mathbf{x} \in \mathbb{R}^d$ sont identifiés aux vecteurs colonnes $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$.

Etant donné $\mathbf{x} \in \mathbb{R}^d$, on note $(x_{i:d})_{i=1,\dots,d}$ les statistiques d'ordre de $\mathbf{x}$, i.e. les valeurs des coordonnées x_i de $\mathbf{x}$ ordonnées par ordre croissant : soit π une permutation de $\{1, \dots, n\}$ (il peut y en avoir plusieurs) telle que

$$x_{\pi(1)} \leqslant \dots \leqslant x_{\pi(n)} \quad .$$

Alors $x_{i:d} = x_{\pi(i)}$. Ainsi, $x_{1:d}$ est le plus petit des x_i, $x_{2:d}$ le second plus petit des x_i, ..., et $x_{d:d}$ le plus grand des x_i.

Si X est une variable aléatoire réelle, F_X désigne sa fonction de répartition et f_X sa densité par rapport à une mesure μ de référence, $\mathbb{E}[X]$ et $\mathrm{Var}(X)$ désignent respectivement l'espérance et la variance de X quand elles sont bien définies.

Si X suit une loi Gaussienne standard $\mathrm{N}(0,1)$ sur $\mathbb{R}$, on note Φ sa fonction de répartition et, pour tout $\alpha \in [0,1]$, $z_\alpha = \Phi^{-1}(1-\alpha)$ le $1-\alpha$ quantile de cette loi.

Chapitre 1

Tests et régions de confiance

1.1 Modèle statistique

On se place dans ce chapitre, et dans la grande majorité de ce cours, dans le cadre de l'inférence statistique. On observe des données $x_1, \dots, x_n$ à valeurs dans un espace mesurable $\mathcal{X}$. On modélise ces observations comme des réalisations de variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribuées (i.i.d.) $X_1, \dots, X_n$ de même loi que X, définie sur un ensemble mesurable $(\Omega, \mathcal{A})$, on dit que $X_1, \dots, X_n$ est un échantillon aléatoire de la variable X. Autrement dit, on considère que $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ pour un certain $\omega \in \Omega$. La loi de X est connue à un paramètre θ près, on la note $\mathbb{P}_\theta$ et on note Θ l'ensemble des valeurs possibles pour ce paramètre θ. Dans l'essentiel du cours, on va en outre supposer qu'on connait une mesure de référence μ par rapport à laquelle toutes les lois $(\mathbb{P}_\theta)_{\theta \in \Theta}$ sont absolument continues. On notera alors la densité de la loi $\mathbb{P}_\theta$ par $f(x, \theta)$. Dans cette écriture, la fonction f est connue mais le paramètre θ est inconnu. L'ensemble des lois $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ est appelé le modèle statistique. Dans la suite, on notera $Z_n = (X_1, \dots, X_n)$ l'échantillon aléatoire observé et $z_n = (x_1, \dots, x_n)$ sa réalisation.

Remarque 1. *Si Θ est un sous-ensemble de $\mathbb{R}^d$ pour un certain d, le modèle est dit paramétrique. La plupart du temps dans ce cours, nous nous placerons dans ce cadre. Si ce n'est pas le cas, par exemple si $\mathcal{P}$ est l'ensemble des lois sur $\mathbb{R}$ ayant une variance finie, le modèle est dit non-paramétrique.*

1.2 Tests statistiques

Dans le problème des tests en statistique, on fait une hypothèse a priori sur le paramètre inconnu θ et il s'agit de décider à l'aide des observations si cette hypothèse est vérifiée ou non.

Exemple 1. *Supposons que le modèle statistique est l'ensemble des lois Gaussiennes $\mathrm{N}(\theta, 1)$, avec $\theta \in \Theta_0 \cup \Theta_1$, où Θ_0 et Θ_1 sont deux sous-ensembles disjoints de $\mathbb{R}$. On veut savoir si $\theta \in \Theta_0$ (c'est notre hypothèse) ou si $\theta \in \Theta_1$.*

Definition 2. *Soit $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ un modèle statistique, une hypothèse H_0 est un sous-ensemble de $\mathcal{P}$. On peut identifier H_0 à un sous-ensemble $\Theta_0 \subset \Theta$ en notant simplement $H_0 : \theta \in \Theta_0$ le sous-ensemble $H_0 = \{\mathbb{P}_\theta, \theta \in \Theta_0\}$.*

Quand on fait un test statistique, on se donne deux hypothèses disjointes H_0 et H_1, H_0 est appelée l'hypothèse nulle, H_1 l'hypothèse alternative.

Definition 3. *Un test est une statistique $\phi = \phi(Z_n) \in \{0,1\}$. ϕ peut être une fonction déterministe de l'observation Z_n, on dit alors qu'on fait un test pur. ϕ peut aussi être une fonction aléatoire, c'est à dire $\phi = \phi(Z_n, U)$ dépendant des observations Z_n et d'une variable aléatoire U indépendante de Z_n, on dit qu'on fait un test randomisé.*

Par exemple, la variable aléatoire $\phi \sim B(\alpha)$ qui vaut 1 avec probabilité α et 0 avec probabilité $1-\alpha$, indépendamment de Z_n, est un test de n'importe quelle hypothèse H_0 contre H_1.

Lorsque $\phi = 1$, on dit qu'on rejette H_0, lorsque $\phi = 0$, on dit qu'on ne rejette pas H_0. On peut commettre deux types d'erreur quand on fait un test :

- rejeter H_0 alors qu'elle est vraie : c'est l'erreur de *première* espèce,
- ne pas rejeter H_0 alors qu'elle est fausse : c'est l'erreur de *seconde* espèce.

Le tableau suivant récapitule l'ensemble des situations possibles.

	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$\phi = 0$	☺	2ème espèce
$\phi = 1$	1ère espèce	☺

Definition 4. *La fonction puissance du test ϕ est définie par*

$$\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(Z_n, U)] = \mathbb{P}_\theta(\phi = 1) \ .$$

La fonction puissance sert à évaluer la qualité d'un test. Rappelons que le statisticien ne connait pas le paramètre θ utilisé pour générer les données. Il recherche donc des garanties valables *pour tous les paramètres* $\theta \in \Theta$. En théorie des tests, Θ peut être remplacé par $\Theta_0 \cup \Theta_1$ si cet ensemble est différent de Θ. D'après ce qu'on vient de dire, un test est bon s'il a une petite erreur de première et de seconde espèce. Ainsi, idéalement, on veut construire des tests pour lesquels la fonction puissance est proche de 0 pour tout $\theta \in \Theta_0$ et proche de 1 pour tout $\theta \in \Theta_1$.

1.3 Dissymétrie des hypothèses

Les objectifs d'avoir simultanément de petites erreurs de première et seconde espèce ne sont pas compatibles en général. Réduire l'erreur de première espèce conduit typiquement à choisir une zone de rejet aussi petite que possible alors que réduire l'erreur de seconde espèce conduit à prendre celle-ci aussi grande que possible. Pour comparer les tests de façon intéressante, il faut donc faire un choix. Le plus classique, que nous utiliserons dans ce cours est d'utiliser le principe de Neyman et de contrôler d'abord l'erreur de première espèce. Pour cela, on introduit la taille du test.

Definition 5. *La taille du test ϕ est*

$$\bar{\alpha}(\phi) = \sup_{\theta \in \Theta_0} \beta_\phi(\theta) \ .$$

Le test ϕ est dit de niveau α si sa taille est inférieure à α, i.e. si $\bar{\alpha}(\phi) \leqslant \alpha$.

En pratique, on va donc s'intéresser d'abord à construire, pour un α donné, un ou plusieurs tests de niveau α. Garantir le niveau d'un test est toujours la première chose à faire, mais ça n'est pas suffisant pour garantir qu'on a construit un bon test. En effet, le test $\phi = 0$ a un niveau nul, mais il n'apporte pas d'information. Pour évaluer un test on va aussi regarder son comportement sous l'hypothèse alternative. La notion suivante donne alors un critère pour préférer un test à un autre.

Definition 6. *Un test ϕ est dit uniformément plus puissant que le test ϕ' si*

$$\forall \theta \in \Theta_1, \qquad \beta_\phi(\theta) \geqslant \beta_{\phi'}(\theta) \ .$$

Clairement, si ϕ et ϕ' sont deux tests de niveau α et si ϕ est uniformément plus puissant que ϕ', alors ϕ est préférable à ϕ'. Toutefois, il est rare de pouvoir comparer directement ainsi deux tests.

Une conséquence pratique importante du principe de Neyman est la dissymétrie qu'elle induit entre l'hypothèse nulle et l'hypothèse alternative. Illustrons cela sur un exemple simple.

Exemple 2. *Supposons qu'on observe une variable aléatoire Gaussienne $X \sim \mathrm{N}(\theta, 1)$ et qu'on souhaite savoir si $\theta = 0$ ou si $\theta = x > 0$.*

Supposons d'abord qu'on fasse pour cela le test $H_0 : \theta = 0$ contre $H_1 : \theta = x$. Une idée naturelle est alors de rejeter si l'observation X dépasse un seuil c. On choisit alors le seuil c de manière à assurer que le test est de niveau α et on prend alors pour cela $c = \Phi^{-1}(1-\alpha)$, Φ étant la fonction de répartition de la loi Gaussienne centrée réduite. On voit donc que la forme *de ce test naturel prend en compte l'hypothèse alternative mais* la calibration *de celui-ci, i.e. le choix de c n'en dépend pas. En particulier, même si $c > x/2$, et même si $c > x$, on choisit quand même $\theta = 0$ si $X < c$, ce qui peut arriver si $X \in [x-\epsilon, x+\epsilon]$ dans ce cas !!*

Supposons maintenant qu'on fasse le test $H_0' : \theta = x$ contre $H_1' : \theta = 0$. Alors, en raisonnant comme précédemment (vérifier le !), on rejette H_0' si $X < x - c$. Supposons x et c soient tels que $x - c < c$. On a alors trois cas

- *Si $X < x - c$, alors les deux tests prennent la décision $\theta = 0$.*
- *Si $X > c$, alors les deux tests prennent la décision $\theta = x$.*
- *Si $X \in [x-c, c]$, alors le premier test prend la décision $\theta = 0$ alors que le second prend la décision $\theta = x$.*

Dans les deux premiers cas, le choix de H_0 n'a donc pas d'importance en revanche, dans le troisième, ce choix guide complètement la décision. La raison est que, si les hypothèses ne sont pas assez séparées (x trop petit) par rapport au niveau de confiance exigé (α trop petit) et à l'information disponible (ici représenté par la variance 1 ou le nombre $n = 1$ d'observations), il existe une "zone grise" ici représentée par le segment $[x-c, c]$. Si les observations tombent dans cette zone grise, le principe de Neyman conduit à prendre la décision H_0.

La dissymétrie entre hypothèses peut avoir des conséquences concrètes importantes, industriels et associations de consommateurs peuvent avoir des intérêts divergents par exemple. L'exemple 2 illustre le fait que changer les hypothèses peut conduire à prendre des décisions opposées. Il est judicieux de bien comprendre cet enjeu pour pouvoir avoir un avis objectif sur la méthodologie employée. Voici quelques heuristiques fréquemment utilisées pour choisir H_0 en pratique.

1. Choisir pour H_0 celle qui est en notre défaveur. Si on veut tester un nouveau médicament, on mettra de préférence en hypothèse nulle qu'il est moins efficace que les précédents. Un rejet de cette hypothèse apportera une preuve plus forte de son efficacité. Ce principe de précaution est celui que nous utilisons en mathématiques et en science plus généralement.
2. Choisir pour hypothèse H_0 celle pour laquelle une erreur présente le plus de risque. Si on souhaite tester la sécurité d'un lieu pour y implanter une centrale nucléaire, on mettra en hypothèse H_0 que le lieu est dangereux. Une erreur alors que H_0 est vraie aurait des conséquences désastreuses alors que déclarer dangereux un lieu sûr conduira simplement à rechercher un autre lieu pour l'implantation.

1.4 p-valeur

Il est intéressant dans certains cas de ne pas fixer le niveau α a priori mais de chercher à quantifier l'incertitude sur l'hypothèse nulle. La p-valeur est un outil classique pour cela.

Definition 7. *Soit $\mathcal{P}$ un modèle statistique et soit $(\phi_\alpha)_{\alpha\in[0,1]}$ une famille de tests vérifiant les hypothèses suivantes.*

1. *Pour tout $\alpha \in [0,1]$, ϕ_α est de niveau α.*
2. *Pour tout α et α' de $[0,1]$ tels que $\alpha < \alpha'$, on a $\phi_\alpha \leqslant \phi_{\alpha'}$.*

La p-valeur de l'observation Z_n pour la famille $(\phi_\alpha)_{\alpha\in[0,1]}$ est alors définie par

$$\hat{\alpha}(Z_n) = \inf\{\alpha \in [0,1] : \phi_\alpha(Z_n) = 1\} \ .$$

Etant donnée l'observation Z_n, la p-valeur $\hat{\alpha}(Z_n)$ est la valeur pour laquelle le test rejette H_0 pour tous $\alpha > \hat{\alpha}(Z_n)$ et ne la rejette pas pour tous les $\alpha < \hat{\alpha}(Z_n)$. Une petite p-valeur est informative, elle indique que l'hypothèse H_0 est peu vraisemblable. Une grande p-valeur en revanche ne renseigne pas.

Exemple 3. *Reprenons le test de l'exemple 2. Pour tout α, le le test $\phi_\alpha = \mathbf{1}_{\{X>\Phi^{-1}(1-\alpha)\}}$ est de niveau α. De plus, si $\alpha < \alpha'$, on a $1-\alpha > 1-\alpha'$ donc, comme Φ^{-1} est strictement croissante, $\Phi^{-1}(1-\alpha) > \Phi^{-1}(1-\alpha')$. Dès lors, si $\phi_\alpha = 1$, $X > \Phi^{-1}(1-\alpha)$ donc $X > \Phi^{-1}(1-\alpha')$ et donc $\phi_{\alpha'} = 1$. Ceci implique que la seconde condition dans la définition de la p-valeur est vérifiée. Pour calculer cette p-valeur en l'observation X, on cherche donc*

$$\begin{aligned}\hat{\alpha}(X) &= \inf\{\alpha \in [0,1] : \phi_\alpha(X) = 1\} \\ &= \inf\{\alpha \in [0,1] : X > \Phi^{-1}(1-\alpha)\} \\ &= \inf\{\alpha \in [0,1] : \alpha > 1 - \Phi(X)\} = 1 - \Phi(X) \ .\end{aligned}$$

Proposition 8. *Soit $\mathcal{P}$ un modèle statistique et soit Θ_0, Θ_1 une partition de Θ. On considère le test de $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ et on suppose qu'on dispose d'une famille $(\phi_\alpha)_{\alpha\in[0,1]}$ de tests satisfaisant les propriétés 1 et 2 de la définition 7.*

Alors, pour tout $\theta \in \Theta_0$, on a

$$\forall u \in (0,1), \qquad \mathbb{P}_\theta(\hat{\alpha}(Z_n) \leqslant u) \leqslant u \ .$$

De plus, pour tout $\theta \in \Theta_0$ tel que $\forall \alpha \in (0,1)$, on a $\mathbb{E}_\theta[\phi_\alpha] = \alpha$, on a

$$\forall u \in (0,1), \qquad \mathbb{P}_\theta(\hat{\alpha}(Z_n) \leqslant u) = u \ .$$

Remarque 9. *La seconde propriété est souvent résumée en "la p-valeur suit une loi uniforme sous H_0". Ce résumé n'est pas inutile pour se faire une intuition dans les exercices, mais il n'est pas inutile de connaitre l'énoncé précis du résultat !*

Preuve. Soit $\theta \in \Theta_0$ et $u \in (0,1)$. On a

$$\hat{\alpha}(Z_n) \leqslant u \quad \Rightarrow \quad \forall v > u, \ \phi_v(Z_n) = 1 \ .$$

On a donc, pour tout $v > u$, $\mathbb{P}_\theta(\hat{\alpha}(Z_n) \leqslant u) \leqslant \mathbb{E}_\theta[\phi_v] \leqslant v$. Comme ceci est vrai pour tout $v > u$, le premier point est démontré.

Soit maintenant $\theta \in \Theta_0$ tel que $\forall \alpha \in (0,1)$, on a $\mathbb{E}_\theta[\phi_\alpha] = \alpha$. Comme $\{\phi_u = 1\} \subset \{\hat{\alpha}(Z_n) \leqslant u\}$, on a donc

$$u = \mathbb{P}_\theta(\phi_u = 1) \leqslant \mathbb{P}_\theta(\hat{\alpha}(Z_n) \leqslant u) \ .$$

Avec le premier point, ceci démontre la seconde partie de la proposition. □

1.5 Régions de confiance

Definition 10. *Une région de confiance pour $g(\theta)$ au niveau $1-\alpha$ est un ensemble aléatoire $\mathcal{C}(Z_n)$ tel que*

1. *$\{z_n \in \mathcal{X}^n : g(\theta) \in \mathcal{C}(z_n)\}$ est mesurable,*
2. *$\inf_{\theta \in \Theta} \mathbb{P}_\theta(g(\theta) \in \mathcal{C}(Z_n)) \geqslant 1-\alpha$.*

Dans le cas où $\Theta \subset \mathbb{R}$, une région de confiance est souvent un intervalle qui prend l'une des formes suivantes :

— $\mathcal{C}(Z_n) = [m(Z_n), +\infty[$, $m(Z_n)$ est appelée borne inférieure de confiance.
— $\mathcal{C}(Z_n) =]-\infty, M(Z_n)]$, $M(Z_n)$ est appelée borne supérieure de confiance.
— $\mathcal{C}(Z_n) = [m(Z_n), M(Z_n)]$, on parle alors d'intervalle de confiance bilatère.

Un outil fondamental pour construire des régions de confiance est celui de fonction pivotale que nous introduisons maintenant.

Definition 11. *Soit $\mathcal{P}$ un modèle statistique et G une fonction mesurable*

$$G : \mathcal{X}^n \times g(\Theta) \to (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p)) \ .$$

La fonction G est dite pivotale pour $g(\theta)$ si, pour tout θ, θ' de Θ,

$$\forall A \in \mathcal{B}(\mathbb{R}^p), \qquad \mathbb{P}_\theta(G(Z_n, g(\theta)) \in A) = \mathbb{P}_{\theta'}(G(Z_n, g(\theta')) \in A) \ .$$

Etant donnée une fonction pivotale pour $g(\theta)$, on construit facilement une région de confiance pour $g(\theta)$ de la façon suivante :

1. On fixe $\theta_0 \in \Theta$ et on détermine A_α tel que $\mathbb{P}_{\theta_0}(G(Z_n, g(\theta_0)) \in A_\alpha) \geqslant 1-\alpha$.
2. On pose $\mathcal{C}(Z_n) = \{g(\theta) \in g(\Theta) : G(Z_n, g(\theta)) \in A_\alpha\}$.

3. On a alors, pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta(g(\theta) \in \mathcal{C}(Z_n)) = \mathbb{P}_\theta(G(Z_n, g(\theta)) \in A_\alpha) = \mathbb{P}_{\theta_0}(G(Z_n, g(\theta_0)) \in A_\alpha) \geqslant 1-\alpha \ .$$

Exemple 4. *Pour tout $\theta > 0$, soit $\mathcal{E}(\theta)$ la loi exponentielle de paramètre θ définie comme la loi sur $\mathbb{R}_+$ de densité par rapport à la mesure de Lebesgue*

$$f(x,\theta) = \frac{1}{\theta} \mathrm{e}^{x/\theta} \ .$$

Considérons le modèle statistique des lois exponentielles $\mathcal{P} = \{\mathcal{E}(\theta), \theta > 0\}$. On veut construire un intervalle de confiance pour θ. Un estimateur naturel de $\theta = \mathbb{E}_\theta[X_1]$ est la moyenne empirique $n^{-1}\sum_{i=1}^n X_i$. On a, sous $\mathbb{P}_\theta$, $\frac{1}{n}\sum_{i=1}^n X_i \sim \Gamma(n, \theta/n)$, donc, sous $\mathbb{P}_\theta$,

$$G(Z_n, \theta) = \frac{1}{\theta}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \sim \Gamma(n, 1/n) \ .$$

Autrement dit, G est une fonction pivotale de θ. On peut en déduire une région de confiance pour θ en appliquant l'algorithme précédent.

1. *Fixons $\theta_0 = 1$ et choisissons a et b égaux respectivement aux $\alpha/2$ et $1-\alpha/2$ quantiles de la loi $\Gamma(n, 1/n)$. De cette façon, on a bien*

$$\mathbb{P}_1(G(Z_n, 1) \in [a, b]) = 1 - \alpha \ .$$

2. *On pose alors comme région de confiance*

$$\mathcal{C}(Z_n) = \{\theta > 0 : G(Z_n, \theta) \in [a, b]\} = \left[\frac{\sum_{i=1}^n X_i}{nb}, \frac{\sum_{i=1}^n X_i}{na}\right] .$$

L'intervalle construit est donc un intervalle bilatère, qui est bien un intervalle de confiance au niveau α d'après le point 3.

Exercice : Construire de la même façon des intervalles de confiance dans le cadre d'un modèle statistique Gaussien :

1. pour la moyenne quand la variance est connue,
2. pour la moyenne quand la variance est inconnue,
3. pour la variance,
4. pour la moyenne et la variance.

1.6 Dualité entre régions de confiance et tests d'hypothèse nulle simple

Dans cette section, on fixe $\alpha \in (0,1)$. Dans un premier temps, on suppose que, pour tout $\kappa \in g(\Theta)$, on dispose d'un test ϕ_κ de niveau α de $H_0 : g(\theta) = \kappa$ contre $H_1 : g(\theta) \neq \kappa$. Autrement, pour tout $\theta \in \Theta$, on a $\mathbb{P}_\theta(\phi_{g(\theta)} = 0) \geqslant 1 - \alpha$. A partir de cet ensemble de tests, on peut définir la région de confiance duale par

$$\mathcal{C}(Z_n) = \{\kappa \in g(\Theta) : \phi_\kappa = 0\} \ .$$

On a alors, pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta(g(\theta) \in \mathcal{C}(Z_n)) = \mathbb{P}_\theta(\phi_{g(\theta)} = 0) \geqslant 1 - \alpha \ .$$

Ainsi, $\mathcal{C}(Z_n)$ est une région de confiance pour $g(\theta)$ de niveau de confiance $1-\alpha$.

Réciproquement, supposons qu'on dispose d'une région de confiance $\mathcal{C}(Z_n)$ pour $g(\theta)$ au niveau $1-\alpha$. Pour tout $\kappa \in g(\Theta)$, on peut alors construire le test suivant des hypothèses $H_0 : g(\theta) = \kappa$ contre $H_1 : g(\theta) \neq \kappa$:

$$\phi_\kappa(Z_n) = \mathbf{1}_{\{\kappa \notin \mathcal{C}(Z_n)\}} \ .$$

On a alors, pour tout $\theta \in \Theta$ tel que $g(\theta) = \kappa$,

$$\mathbb{P}_\theta(\phi_\kappa = 1) = \mathbb{P}_\theta(g(\theta) \notin \mathcal{C}(Z_n)) \leqslant \alpha \ .$$

Autrement dit, le test ϕ_κ est de niveau $1-\alpha$.

Exemple 5. *Reprenons le cadre de l'exemple 2. La dualité entre tests et régions de confiance donne un autre regard sur la dissymétrie des hypothèses H_0 et H_1. En effet, nous disposons dans cet exemple, pour chaque $\kappa \in \{0, x\}$ d'un test ϕ_κ de $H_0 : \theta = \kappa$ contre $H_1 : \theta \neq \kappa$. On peut donc construire à partir de ces tests la région de confiance duale. Sous la condition $x - c < c$ de l'exemple 2, cette région de confiance est égale à*

$$\mathcal{C}(X) = \begin{cases} \{0\} & \text{si } X < x - c \ , \\ \{x\} & \text{si } X > c \ , \\ \{0, x\} & \text{si } X \in [x-c, c] \ . \end{cases}$$

La "zone grise" que nous évoquions est celle dans laquelle les données ne permettent pas de décider entre les hypothèses, ce qui se concrétise par le fait que la région de confiance associée contient les deux hypothèses.

1.7 Construction de tests : méthode du pivot

Supposons qu'on observe un échantillon aléatoire de variables Gaussiennes $\mathbb{P}_\theta = \mathrm{N}(\mu, \sigma^2)$, avec $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$. On souhaite tester $H_0 : \mu = 0$ contre $H_1 : \mu \neq 0$. Comme le test porte sur $\mu = \mathbb{E}_\theta[X]$, il est naturel d'estimer ce paramètre et on pose donc $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$. Cet estimateur a pour loi $\mathrm{N}(\mu, \sigma^2/n)$. La statistique $\hat{\mu} - \mu \sim \mathrm{N}(0, \sigma^2/n)$ a une loi dépendant du paramètre inconnu σ^2, ce n'est donc pas un pivot qu'on peut utiliser pour calculer une région de confiance. Pour construire un pivot, une solution est alors d'estimer σ^2 également. L'estimateur du maximum de vraisemblance est donné par

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu})^2 \ .$$

De plus, le théorème de Cochran (rappelé au théorème 23 du chapitre suivant) permet de montrer que $\hat{\sigma}^2$ est indépendant de $\hat{\mu}$ et que $K = n\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-1)$ (vérifier le en exercice !). Par construction, on a donc $N = \sqrt{n}(\widehat{\mu} - \mu)/\sigma \sim \mathrm{N}(0,1)$ et K sont indépendantes avec $K \sim \chi^2(n-1)$, donc

$$\frac{N}{\sqrt{K/(n-1)}} = \frac{\sqrt{n}(\widehat{\mu} - \mu)/\sigma}{\sqrt{n\hat{\sigma}^2/(n-1)\sigma^2}} = \frac{\sqrt{n-1}(\widehat{\mu} - \mu)}{\hat{\sigma}} \sim \mathcal{T}(n-1) \ .$$

Ainsi, $G(Z_n, \mu) = \sqrt{n-1}(\widehat{\mu} - \mu)/\hat{\sigma}$ est une fonction pivotale pour μ qu'on peut utiliser pour construire un test.

Une idée naturelle est de rejeter H_0 lorsque $|\widehat{\mu}|$ dépasse un seuil c. Ce seuil doit être calibrer de façon à ce que le niveau du test soit inférieur à α. On va utiliser pour cela le pivot G. On a

$$\mathbb{P}_0(|\widehat{\mu}| > c) = \mathbb{P}_0\left(|G(Z_n, 0)| > \sqrt{n-1}\frac{c}{\hat{\sigma}}\right) .$$

Pour que le test soit de taille α, il suffit donc de choisir c de façon à ce que

$$\sqrt{n-1}\frac{c}{\hat{\sigma}} = T_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right) ,$$

où T_k désigne la fonction de répartition de la loi de Student à k degrés de liberté.

Chapitre 2

Tests asymptotiques

Dans tout le chapitre, on note $\{X_n, n \in \mathbb{N}\}$ une suite de variables aléatoires telle que, pour tout $n \geqslant 1$, $Z_n = (X_1, \dots, X_n)$ un échantillon aléatoire de loi $\mathbb{P}_\theta$, où $\theta \in \Theta$ est un paramètre inconnu. La loi $\mathbb{P}_\theta$ est toujours supposée absolument continue par rapport à une mesure μ de référence connue, et sa densité est notée $f(\cdot, \theta)$. On note Θ_0 et Θ_1 deux sous-ensembles disjoints de Θ et on souhaite tester $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$. On note $\phi_n = \phi_n(Z_n)$ une suite de tests de ces hypothèses. Dans tout le chapitre, on relâche les conditions qu'on a définies au chapitre précédent sur les tests pour considérer des versions de ces contrôles adaptés à l'utilisation de la loi asymptotique des estimateurs.

Definition 12. *La suite de tests ϕ_n est dite de niveau asymptotique α si, pour tout $\theta \in \Theta_0$, $\limsup_{n\to\infty} \mathbb{P}_\theta(\phi_n = 1) \leqslant \alpha$. Elle est dite consistante si, pour tout $\theta \in \Theta_1$, $\liminf_{n\to\infty} \mathbb{P}_\theta(\phi_n = 1) = 1$.*

Ainsi, si un test est de niveau asymptotique α, cela signifie que son niveau pour une taille d'échantillon fixé s'écrit $\alpha + \epsilon_n$, avec $\epsilon_n \to 0$. Comme on ne dispose pas de contrôle sur ϵ_n, on suppose que n est assez grand pour que l'approximation soit acceptable en pratique. Le point de vue asymptotique est extrêmement puissant et donc très répandu, car on peut dans de très nombreux cas obtenir la loi limite des estimateurs et utiliser cette loi pour approcher la loi inconnue de l'estimateur. Toutefois, il est important de se souvenir que ce n'est qu'une approximation. Au chapitre suivant, nous verrons une technique qui permet dans certains cas d'éviter de recourir à la loi limite, sans toutefois préciser de modèle paramétrique pour les observations.

Exemple 6 (Modèle de translation et d'échelle)**.** *Supposons que $Z_n = (X_1, \dots, X_n)$ est un échantillon aléatoire tel que*

$$X_i = \mu + \sigma\varepsilon_i \ ,$$

où le paramètre $\theta = (\mu, \sigma)$ avec $\mu \in \mathbb{R}$ et $\sigma > 0$, est inconnu et $\varepsilon_1, \dots, \varepsilon_n$ sont des variables aléatoires de loi $\mathbb{P}$ inconnue mais telle que $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[\varepsilon^2] = 1$. Remarquons que ce modèle statistique est non-paramétrique *puisque l'ensemble des lois possibles pour ε est infini dimensionnel. On souhaite tester $H_0 : \mu = 0$ contre $H_1 : \mu \neq 0$. Comme le paramètre sur lequel porte le test est $\mu = \mathbb{E}_\theta[X]$, on peut l'estimer par $\widehat{\mu}_n = n^{-1} \sum_{i=1}^n X_i$. Clairement, la loi de $\widehat{\mu}_n$ est inconnue, mais*

on peut l'approcher par sa loi asymptotique. Le théorème de la limite centrale (rappelé au théorème 15) assure que

$$\sqrt{n}\frac{\widehat{\mu}_n - \mu}{\sigma} \xRightarrow{\mathbb{P}_\theta} \mathrm{N}(0,1) \ .$$

De plus, le lemme de Slutsky (voir le lemme 21) assure que, puisque $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$ *est un estimateur consistant de* σ^2*,*

$$\sqrt{n}\frac{\widehat{\mu}_n - \mu}{\widehat{\sigma}_n} \xRightarrow{\mathbb{P}_\theta} \mathrm{N}(0,1) \ .$$

Puisque $G(Z_n, \mu) = \sqrt{n}\frac{\widehat{\mu}_n - \mu}{\widehat{\sigma}_n}$ *est un pivot asymptotique pour* μ*, on peut l'utiliser pour construire un test. Une idée naturelle est de rejeter* H_0 *si* $|\widehat{\mu}_n| > c$ *et si* $c_\alpha = \widehat{\sigma}_n z_\alpha/\sqrt{n}$ *avec* $z_\alpha = \Phi^{-1}(1-\alpha)$*, on a*

$$\forall \sigma > 0, \qquad \mathbb{P}_{0,\sigma}(|\widehat{\mu}_n| > c) = \mathbb{P}_{0,\sigma}\left(\sqrt{n}\frac{|\widehat{\mu}_n|}{\widehat{\sigma}_n} > z_\alpha\right) \to \alpha \ .$$

Autrement dit, $\phi_\alpha = \mathbf{1}_{\{|\widehat{\mu}_n| > c_\alpha\}}$ *est de niveau asymptotique* α*. Soit maintenant* $M > 0$*, pour tout* $\theta = (\mu, \sigma)$ *avec* $\mu \neq 0$ *et* $\sigma > 0$*, on a*

$$\begin{aligned}
\mathbb{P}_\theta(|\widehat{\mu}_n| > c_\alpha) &\geqslant \mathbb{P}_\theta\left(\sqrt{n}\frac{|\mu| - |\widehat{\mu}_n - \mu|}{\widehat{\sigma}_n} > c_\alpha\right) \\
&\geqslant \mathbb{P}_\theta\left(\sqrt{n}\frac{|\widehat{\mu}_n - \mu|}{\widehat{\sigma}_n} < \sqrt{n}\frac{|\mu|}{2\sigma} - c_\alpha\right) - \mathbb{P}_\theta(|\widehat{\sigma}_n - \sigma| > \sigma) \\
&\geqslant \mathbb{P}_\theta\left(\sqrt{n}\frac{|\widehat{\mu}_n - \mu|}{\widehat{\sigma}_n} < M\right) - \mathbb{P}_\theta(|\widehat{\sigma}_n - \sigma| > \sigma) \ .
\end{aligned}$$

La dernière inégalité est vraie pour tout $n \geqslant n_0$*. Le premier terme dans cette dernière minoration converge vers* $\Phi(M)$ *et le second vers* 0*, de sorte que*

$$\liminf_n \mathbb{P}_\theta(|\widehat{\mu}_n| > c_\alpha) \geqslant \Phi(M) \ .$$

Ce résultat étant vrai pour tout M*, le test* ϕ_α *est bien consistant.*

2.1 Quelques résultats de probabilités

L'exemple 6 montre qu'on va être amené à utiliser régulièrement dans ce chapitre des résultats de probabilité, notamment des théorèmes limites pour les suites de variables aléatoires indépendantes. Cette section rappelle plusieurs de ces résultats qu'il est bon de connaître. On se référera au cours de probabilités et de stat 1 pour des preuves de ces résultats.

2.1.1 Consistance

Commençons par la loi faible des grands nombres.

Proposition 13. *Si* $\{X_n,\ n \in \mathbb{N}\}$ *est une suite de variables aléatoires indépendantes et de même loi (i.i.d.)* $\mathbb{P}$ *telle que* $\mathbb{E}[|X|] < \infty$*, alors la moyenne empirique* $n^{-1}\sum_{i=1}^n X_i \overset{\mathbb{P}\text{-prob}}{\longrightarrow} \mathbb{E}[X]$*, c'est à dire que, pour tout* $\epsilon > 0$*,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]\right| > \epsilon\right) \to 0 \ .$$

Le théorème suivant est appelé théorème de l'application continue pour la convergence en probabilité.

Proposition 14. *Soit* $\{X_n,\ n \in \mathbb{N}\}$ *une suite de variables aléatoires et soit* X *une variable aléatoire telles que* $X_n \overset{\mathbb{P}-\text{prob}}{\longrightarrow} X$. *Soit* f *une fonction continue, alors* $f(X_n) \overset{\mathbb{P}-\text{prob}}{\longrightarrow} f(X)$.

2.1.2 Normalité asymptotique

Commençons par le théorème de la limite centrale.

Théorème 15. *Soit* $\{X_n,\ n \in \mathbb{N}\}$ *une suite de variables aléatoire indépendantes et de même loi telles que* $\mathbb{E}[X^2] < \infty$. *Soit* $\sigma^2 = \mathrm{Var}(X)$, *alors*

$$\sqrt{n}\frac{n^{-1}\sum_{i=1}^n X_i - \mathbb{E}[X]}{\sigma} \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0,1) \ ,$$

c'est à dire que, pour tout $t \in \mathbb{R}$,

$$\mathbb{P}\left(\sqrt{n}\frac{n^{-1}\sum_{i=1}^n X_i - \mathbb{E}[X]}{\sigma} \leqslant t\right) \to \Phi(t) \ ,$$

où Φ *est la fonction de répartition de la loi Gaussienne standard* $\mathrm{N}(0,1)$.

La vitesse de convergence dans le théorème 15 peut être précisée sous des hypothèses plus fortes sur les moments de X.

Théorème 16 (Berry-Essen)**.** *Soient* $\{X_n,\ n \in \mathbb{N}\}$ *une suite de variables aléatoires i.i.d. telles que* $\mathbb{E}[|X|^3] < \infty$. *Soient* $\sigma^2 = \mathrm{Var}(X)$ *et* $\rho = \mathbb{E}[|X - \mathbb{E}[X]|^3]$. *Alors*

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sqrt{n}\frac{n^{-1}\sum_{i=1}^n X_i - \mathbb{E}[X]}{\sigma} \leqslant t\right) - \Phi(t)\right| \leqslant \frac{\rho}{2\sigma^3\sqrt{n}} \ .$$

Le théorème 15 admet également une version multivariée qui sera utile.

Théorème 17. *Soit* $\{X_n,\ n \in \mathbb{N}\}$ *une suite de vecteurs aléatoires de* $\mathbb{R}^d$ *i.i.d. tels que* $\mathbb{E}[\|X\|^2] < \infty$. *Soit* $\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$ *la matrice de covariance de la loi de* X. *Alors on a*

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^n X_i - \mathbb{E}[X]\right) \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0,\Sigma) \ .$$

Comme pour la convergence en probabilité, il existe un théorème de continuité pour la convergence en loi.

Proposition 18. *Soit* $\{X_n,\ n \in \mathbb{N}\}$ *une suite de variables aléatoires et soit* X *une variable aléatoire telles que* $X_n \overset{\mathbb{P}}{\Longrightarrow} X$. *Soit* f *une fonction continue, alors* $f(X_n) \overset{\mathbb{P}}{\Longrightarrow} f(X)$.

Le résultat suivant est connu sous le nom de "méthode Delta" en statistique.

Théorème 19 (Méthode Delta)**.** *Soit* $\{X_n,\ n \in \mathbb{N}\}$ *une suite de variables aléatoires i.i.d. de même loi que* X, *où* $\mathbb{E}[X^2] < \infty$. *Soit* $\sigma^2 = \mathrm{Var}(X)$ *et soit* f *une fonction dérivable en* $\mathbb{E}[X]$, *telle que* $f'(\mathbb{E}[X]) > 0$. *Alors,*

$$\sqrt{n}\left(f\left(n^{-1}\sum_{i=1}^n X_i\right) - f(\mathbb{E}[X])\right) \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0,\sigma^2(f'(\mathbb{E}[X]))^2) \ .$$

Ce résultat admet également une version multivariée.

Théorème 20 (Méthode Delta multivariée)**.** *Soit* $\{X_n,\ n \in \mathbb{N}\}$ *une suite de vecteurs aléatoires i.i.d. de même loi que* X*, où* $\mathbb{E}[\|X\|^2] < \infty$*. Soit* $\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$ *et soit* $f : \mathbb{R}^d \to \mathbb{R}$ *une fonction différentiable en* $\mathbb{E}[X]$*, telle que* $\nabla f(\mathbb{E}[X]) \neq 0$*. Alors,*

$$\sqrt{n}\left(f\left(n^{-1}\sum_{i=1}^{n} X_i\right) - f(\mathbb{E}[X])\right) \xRightarrow{\mathbb{P}} \mathrm{N}(0, \nabla f(\mathbb{E}[X])^T \Sigma \nabla f(\mathbb{E}[X]))\ .$$

Le résultat suivant est connu sous le nom de lemme de Slutsky.

Lemme 21 (Slutsky)**.** *Soit* $\{X_n,\ n \in \mathbb{N}\}$ *et* $\{Y_n,\ n \in \mathbb{N}\}$ *deux suites de variables aléatoires telles que* $X_n \xRightarrow{\mathbb{P}} X$*,* $Y_n \xrightarrow{\mathbb{P}\text{-prob}} c$*, où* X *est une variable aléatoire et* c *une constante. Alors* $(X_n, Y_n) \xRightarrow{\mathbb{P}} (X, c)$*. En particulier* $X_n + Y_n \xRightarrow{\mathbb{P}} X + c$ *et* $X_n Y_n \xRightarrow{\mathbb{P}} cX$*.*

Le théorème de Lindeberg-Feller est un théorème de normalité asymptotique pour les tableaux de variables aléatoires indépendantes mais pas nécessairement de même loi. On en donne directement la version multivariée.

Théorème 22 (Lindeberg-Feller)**.** *Soit* k_n *une suite croissante d'entiers. Soit* $X_{n,1}, \ldots, X_{n,k_n}$ *une suite de vecteurs aléatoires indépendants de* $\mathbb{R}^d$*, centrés et vérifiant les conditions suivantes :*

1. $\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}[X_{n,i} X_{n,i}^T] = \Sigma$*,*
2. *pour tout* $\epsilon > 0$*,* $\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}[\|X_{n,i}\|^2 \mathbf{1}_{\{\|X_{n,i}\|^2 > \epsilon\}}] = 0$*.*

Alors, on $\sum_{i=1}^{k_n} X_{n,i} \xRightarrow{\mathbb{P}} \mathrm{N}(0, \Sigma)$*.*

Enfin, on mentionne le théorème de Cochran qu'on utilise pour manipuler les limites.

Théorème 23 (Cochran)**.** *Soit* $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ *un vecteur Gaussien standard de* $\mathbb{R}^d$ *et* Π *un projecteur orthogonal sur un espace de dimension* k*. Alors,*

1. $\Pi\boldsymbol{X}$ *et* $(\boldsymbol{I}_d - \Pi)\boldsymbol{X}$ *sont indépendants.*
2. $\|\Pi\boldsymbol{X}\|^2 \sim \chi^2(k)$*.*

Exercice : Montrer que, si $\mathbf{Y} \sim \mathrm{N}(0, \Pi)$, alors $\|\mathbf{Y}\|^2 \sim \chi^2(k)$.

Exercice : Montrer que, si $\mathbf{X} = (X_1, \ldots, X_n)^T \sim \mathrm{N}(\mu, \sigma^2 \mathbf{I}_d)$, alors $\widehat{\mu} = n^{-1} \sum_{i=1}^n X_i$ et $\sum_{i=1}^n (X_i - \widehat{\mu})^2$ sont indépendants et $\sigma^{-2} \sum_{i=1}^n (X_i - \widehat{\mu})^2 \sim \chi^2(n-1)$.

2.1.3 Estimateurs du maximum de vraisemblance

Les résultats de cette section sont issus du cours de stat 1.

Definition 24. *Soit* $(\mathbb{P}_\theta, \theta \in \Theta)$ *un ensemble de lois indexé par un sous-ensemble* $\Theta \subset \mathbb{R}^d$*. Le modèle est dit régulier s'il satisfait les propriétés suivantes :*

1. *Il existe une mesure* μ *σ-finie telle que toutes les lois* $\{\mathbb{P}_\theta, \theta \in \Theta\}$ *soient dominées par* μ*. On note* $f(\cdot, \theta)$ *la densité de* $\mathbb{P}_\theta$ *par rapport à* μ *et* $\ell(x, \theta) = \log f(x, \theta)$*.*

2. *Pour μ-presque tout $x \in \mathcal{X}$, la fonction $\theta \mapsto f(x,\theta)$ est de classe $\mathcal{C}^2$ sur Θ.*
3. *Pour tout $\theta_0 \in \Theta$, il existe un voisinage V de θ_0 et deux fonctions g et h positives telles que, pour tout $\theta \in V$, et pour μ presque tout $x \in \mathcal{X}$,*

$$\|\nabla \ell(x,\theta)\|^2 \leqslant h(x), \qquad \|H_\ell(x,\theta)\| \leqslant h(x), \qquad f(x,\theta) \leqslant g(x)$$

et $\int_{\mathcal{X}}(1+h)g\mathrm{d}\mu < \infty$, où H_ℓ est la matrice Hessienne de la log-vraisemblance $H_\ell(x,\theta) = \nabla^2_\theta \ell(x,\theta)$.
4. *Pour tout $\theta \in \Theta$, la matrice $\mathbb{E}_\theta[H_\ell(X,\theta)]$ est inversible.*

Dans les modèles réguliers, on peut montrer la normalité asymptotique de l'estimateur du maximum de vraisemblance.

Théorème 25. *Si le modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$ est régulier, alors toute suite consistante d'estimateurs du maximum de vraisemblance $\widehat{\theta}_n$ vérifie*

$$\forall \theta \in \Theta, \qquad \sqrt{n}(\widehat{\theta}_n - \theta) \xRightarrow{\mathbb{P}_\theta} \mathrm{N}(0, \mathbb{I}^{-1}(\theta)) \ ,$$

où la matrice $\mathbb{I}(\theta)$ est la matrice d'information de Fisher définie par

$$\mathbb{I}(\theta) = \mathbb{E}_\theta[\nabla \ell(X,\theta) \nabla \ell(X,\theta)^T] = -\mathbb{E}_\theta[H_\ell(X,\theta)] \ .$$

On peut déduire de ce résultat un théorème important connu sous le nom de théorème de Wilks.

Théorème 26. *Soit $\{\mathbb{P}_\theta, \theta \in \Theta\}$ un modèle statistique régulier et $\widehat{\theta}_n$ une suite consistante d'estimateurs du maximum de vraisemblance. Alors, pour tout $\theta \in \Theta$,*

$$2n(\ell(Z_n, \widehat{\theta}_n) - \ell(Z_n,\theta)) \xRightarrow{\mathbb{P}_\theta} \chi^2(d) \ .$$

Démonstration. Dans un modèle régulier, on a $\nabla \ell(Z_n, \widehat{\theta}_n) = 0$ par définition de l'estimateur du maximum de vraisemblance. Par le théorème de Taylor-Lagrange, il existe donc un point θ_n^* du segment $[\theta, \widehat{\theta}_n]$ tel que

$$\ell(Z_n, \widehat{\theta}_n) - \ell(Z_n,\theta) = \frac{1}{2}(\widehat{\theta}_n - \theta)^T \mathbb{I}_n(\theta_n^*)(\widehat{\theta}_n - \theta) \ .$$

Par le théorème de normalité asymptotique des estimateurs du maximum de vraisemblance $\sqrt{n}\mathbb{I}(\theta)^{1/2}(\widehat{\theta}_n - \theta) \xRightarrow{\mathbb{P}_\theta} \mathrm{N}(0, \mathbf{I}_d)$ donc d'après le théorème de l'application continue

$$n(\widehat{\theta}_n - \theta)^T \mathbb{I}(\theta)(\widehat{\theta}_n - \theta) \xRightarrow{\mathbb{P}_\theta} \|\mathrm{N}(0, \mathbf{I}_d)\|^2 \ .$$

Donc, d'après le théorème de Cochran,

$$n(\widehat{\theta}_n - \theta)^T \mathbb{I}(\theta)(\widehat{\theta}_n - \theta) \xRightarrow{\mathbb{P}_\theta} \chi^2(d) \ .$$

De plus, comme $\widehat{\theta}_n$ est consistant, par le théorème de l'application continue, $\mathbb{I}_n(\theta_n^*) \xrightarrow{\mathbb{P}_\theta-\mathrm{prob}} \mathbb{I}(\theta)$. Donc, par le lemme de Slutsky,

$$2n(\ell(Z_n, \widehat{\theta}_n) - \ell(Z_n,\theta)) = n(\widehat{\theta}_n - \theta)^T \mathbb{I}(\theta_n^*)(\widehat{\theta}_n - \theta) \xRightarrow{\mathbb{P}_\theta} \chi^2(d) \ .$$

□

2.1.4 Quantiles empiriques

Le théorème suivant donne la normalité asymptotique des quantiles empiriques. Soit $\{x_n, n \geqslant 1\}$ une suite de réels. Pour tout $n \geqslant 1$, on note π une permutation de $\{1, \dots, n\}$ telle que

$$x_{\pi(1)} \leqslant \dots \leqslant x_{\pi(n)} \ .$$

Pour tout $k \in \{1, \dots, n\}$, on note alors $x_{k:n} = x_{\pi(k)}$. En d'autres termes, $x_{k:n}$ est la k-ième plus petite valeur de l'ensemble $\{x_1, \dots, x_n\}$.

Théorème 27. *Soit $\{X_n, n \geqslant 1\}$ une suite de variables aléatoires réelles, indépendantes et de même loi de fonction de répartition F et de fonction quantile F^{-1}. Soit $p \in (0,1)$ et k_n une suite d'entiers tels que*

$$\sqrt{n}(k_n/n - p) \to 0 \ .$$

Si F est dérivable au point $F^{-1}(p)$ avec $F'(F^{-1}(p)) > 0$ alors

$$\sqrt{n}(X_{k_n:n} - F^{-1}(p)) \xRightarrow{\mathbb{P}} \mathrm{N}\left(0, \frac{p(1-p)}{F'(F^{-1}(p))^2}\right) \ .$$

Le théorème 27 établit la normalité asymptotique des quantiles empiriques sous des hypothèses faibles au sens ou elles suffisent à donner un sens au résultat.

2.2 Tests d'adéquation

Soit $\theta_0 \in \Theta$. Un test d'adéquation ("goddness-of-fit" test) est un test de $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Autrement dit, on suppose que les données sont issues de la loi $\mathbb{P}_{\theta_0}$ et on veut savoir si l'échantillon est en adéquation avec cette hypothèse.

2.2.1 Un exemple dans un modèle de translation

Supposons qu'on observe un échantillon $X_1, \dots, X_n$ de variables aléatoires réelles i.i.d. dont la densité par rapport à la mesure de Lebesgue est donnée par

$$f(x, \theta) = \frac{1}{\pi(1 + (x-\theta)^2)} \ .$$

On veut tester $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$. On dispose de deux estimateurs naturels de θ. Comme la fonction $f(x, \theta) = g(x - \theta)$ avec $g(x) = [\pi(1 + x^2)]^{-1}$ paire, on a $\theta = \text{mediane}(X)$, donc on peut l'estimer par

$$\hat{\theta}_n = \text{mediane}(X_1, \dots, X_n) \ .$$

D'autre part, on peut aussi estimer θ par maximum de vraisemblance en posant

$$\hat{\theta}_{\text{mv}} \in \text{argmax}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \log f(X_i, \theta) \ .$$

$\hat{\theta}_n$ a une forme explicite mais sa loi n'est pas connue. $\hat{\theta}_{\text{mv}}$ n'est quant à lui qu'implicite, en pratique, on l'approche numériquement. Sa loi exacte est également inconnue.

En revanche, pour ces deux estimateurs, on peut établir la distribution asymptotique. On va appliquer le théorème de normalité asymptotique des quantiles, théorème 27 avec $p = 1/2$. Comme la fonction de répartition de $\mathbb{P}_\theta$ est partout dérivable, de dérivée partout strictement positive, le théorème s'applique. On a $F'(x) = f(x,\theta)$ et $F^{-1}(1/2) = \theta$, donc le théorème assure que

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{\mathbb{P}_\theta}{\Longrightarrow} \mathrm{N}\left(0, \frac{1}{4f(\theta,\theta)^2}\right) = \mathrm{N}\left(0, \frac{\pi^2}{4}\right) \ .$$

La loi asymptotique de $G(Z_n,\theta) = \sqrt{n}(\hat{\theta}_n - \theta)$ est indépendante de θ, on dit que c'est un pivot asymptotique pour θ. On utilise alors ce résultat comme si G était un pivot. On veut rejeter H_0 si $|\hat{\theta}_n| > c$. On a

$$\mathbb{P}_0(|\hat{\theta}_n| > c) = \mathbb{P}_0(2\pi|G(Z_n,0)| > 2\sqrt{n}c/\pi) \ .$$

Donc, $\mathbb{P}_0(|\hat{\theta}_n| > c) \to \alpha$ si $2\sqrt{n}c = \pi\Phi^{-1}(1-\alpha/2)$.

Pour étudier l'estimateur du maximum de vraisemblance, on peut utiliser la normalité asymptotique de cet estimateur dans les modèles réguliers, voir le théorème 25. On a

$$\sqrt{n}(\hat{\theta}_{\mathrm{mv}} - \theta) \overset{\mathbb{P}_\theta}{\Longrightarrow} \mathrm{N}(0, \mathbb{I}(\theta)^{-1}) \ ,$$

où $\mathbb{I}(\theta)$ est l'information de Fisher du modèle, définie par

$$\mathbb{I}(\theta) = \mathrm{Var}_\theta(\partial_\theta \log f(X,\theta)) = \mathbb{E}_\theta[(\partial_\theta \log f(X,\theta))^2] = -\mathbb{E}_\theta[\partial_\theta^2 \log f(X,\theta)] \ .$$

On va déterminer cette information de Fisher ici. On a

$$\partial_\theta \log f(x,\theta) = \frac{2(\theta - x)}{1 + (\theta - x)^2} \ .$$

Donc

$$\mathbb{I}(\theta) = \frac{4}{\pi}\int_{-\infty}^{\infty} \frac{(x-\theta)^2}{[1+(x-\theta)^2]^3}\mathrm{d}x = \frac{4}{\pi}\int_{-\infty}^{\infty} \frac{x^2}{[1+x^2]^3}\mathrm{d}x = \frac{1}{2} \ .$$

On en déduit $G_{\mathrm{mv}}(Z_n,\theta) = \sqrt{n}(\hat{\theta}_{\mathrm{mv}} - \theta)$ est un autre pivot asymptotique pour θ. On peut terminer comme dans l'exemple précédent en rejetant H_0 si

$$|\hat{\theta}_{\mathrm{mv}}| > \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{2n}} \ .$$

La loi asymptotique peut aussi être utilisée pour comparer les estimateurs, la variance asymptotique de $\hat{\theta}_{\mathrm{mv}}$ vaut 2, elle est inférieure à $\pi^2/4$ qui est la variance asymptotique de $\hat{\theta}_n$, donc $\hat{\theta}_{\mathrm{mv}}$ est préférable.

2.2.2 Test du χ^2

Supposons d'abord que $\mathcal{X} = \{0,\dots,M\}$. Introduisons

$$\Theta = \{(\theta_1,\dots,\theta_M) \in (\mathbb{R}_+^*)^M : \sum_{i=1}^M \theta_i < 1\} \ ,$$

et, pour tout $\theta \in \Theta$, soit $\mathbb{P}_\theta$ la loi sur $\mathcal{X}$ telle que

$$\forall i \in \{1,\dots,M\}, \qquad \mathbb{P}_\theta(X = i) = \theta_i \ .$$

On introduit, pour tout $i \in \{1, \dots, M\}$,

$$\widehat{\theta_i} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{X_j = i\}} \ .$$

Soit enfin

$$G(Z_n, \theta) = \sqrt{n} \begin{bmatrix} \frac{\widehat{\theta}_1 - \theta_1}{\sqrt{\theta_1}} \\ \vdots \\ \frac{\widehat{\theta}_M - \theta_M}{\sqrt{\theta_M}} \end{bmatrix} \ .$$

Proposition 28. *On a, pour tout $\theta \in \Theta$,*

$$\|G(Z_n, \theta)\|^2 \xRightarrow{\mathbb{P}_\theta} \chi^2(M-1) \ .$$

Démonstration. Le théorème de la limite centrale multivarié, théorème 17, assure que

$$G(Z_n, \theta) \xRightarrow{\mathbb{P}_\theta} \mathrm{N}(0, \Sigma(\theta)) \ .$$

La matrice $\Sigma(\theta)$ est la matrice de coefficient générique

$$\Sigma(\theta)_{i,j} = \frac{\mathbb{E}_\theta[(\mathbf{1}_{\{X=i\}} - \theta_i)(\mathbf{1}_{\{X=j\}} - \theta_j)]}{\sqrt{\theta_i \theta_j}} = \begin{cases} 1 - \theta_i & \text{si } i = j \\ -\sqrt{\theta_i \theta_j} & \text{si } i \neq j \end{cases} \ .$$

Par le théorème de l'application continue, on a donc

$$\|G(Z_n, \theta)\|^2 \xRightarrow{\mathbb{P}_\theta} \|\mathrm{N}(0, \Sigma(\theta))\|^2 \ .$$

Soit $\sqrt{\theta} = (\sqrt{\theta_1}, \dots, \sqrt{\theta_M})^T$, on a donc $\Sigma(\theta) = \mathbf{I}_M - \sqrt{\theta}\sqrt{\theta}^T$. Comme $\sqrt{\theta}$ est un vecteur de norme Euclidienne égale à 1, on a $\sqrt{\theta}\sqrt{\theta}^T$ est la matrice de la projection orthogonale sur $D = \mathrm{Vect}(\sqrt{\theta})$, donc $\Sigma(\theta)$ est la matrice de la projection orthogonale sur $D^\perp$. Donc, par le théorème de Cochran, théorème 23 (et l'exercice qui le suit directement),

$$\|\mathrm{N}(0, \Sigma(\theta))\|^2 = \chi^2(M-1) \ .$$

□

Le test d'adéquation du χ^2 consiste à rejeter $H_0 : \theta = \theta_0$ si $\|G(Z_n, \theta_0)\|^2 > \chi^2_{1-\alpha}(M-1)$, où $\chi^2_{1-\alpha}(M-1)$ est le $(1-\alpha)$-quantile de la loi $\chi^2(M-1)$.

Proposition 29. *Le test d'adéquation du χ^2 est consistant et de niveau asymptotique α.*

La preuve est à faire à titre d'exercice.

Si maintenant $\mathcal{X}$ est un espace mesurable quelconque et $\{\mathbb{P}_\theta, \theta \in \Theta\}$ un modèle statistique. On peut se donner une partition $A_0, \dots, A_M$ de $\mathcal{X}$ en sous-ensembles tels que, pour tout $\theta \in \Theta$, $\theta_i = \mathbb{P}_\theta(A_i) \neq 0$. On peut alors estimer chaque θ_i avec $i \in \{1, \dots, M\}$ par $\widehat{\theta_i} = n^{-1} \sum_{j=1}^{n} \mathbf{1}_{\{X_j \in A_i\}}$ et construire la statistique $G(Z_n, \theta)$ comme dans l'exemple précédent. Le test du χ^2 associé est toujours de niveau asymptotique α. En revanche, il n'est consistant que si la partition A_i vérifie que, pour tout $\theta \neq \theta_0$, il existe i_0 tel que

$$\mathbb{P}_\theta(A_{i_0}) \neq \mathbb{P}_{\theta_0}(A_{i_0}) \ .$$

Là encore, la preuve de ce résultat est laissée en exercice.

2.3 Test de Kolmogorov Smirnov

Supposons que $\mathcal{X} \subset \mathbb{R}$. Pour tout $\theta \in \Theta$ et tout $t \in \mathbb{R}$, soit

$$F_\theta(t) = \mathbb{P}_\theta(X \leqslant t), \qquad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leqslant t\}} \ .$$

On a d'après le théorème de la limite centrale

$$\forall t \in \mathbb{R}, \qquad \sqrt{n}(F_n(t) - F_\theta(t)) \overset{\mathbb{P}_\theta}{\Longrightarrow} \mathrm{N}(0, F_\theta(t)(1 - F_\theta(t))) \ .$$

Ce résultat peut être amélioré en utilisant des techniques de processus empirique qui dépassent le cadre de ce cours (on pourra par exemple se référer au livre de van der Vaart "asymptotic statistics" pour une preuve).

Théorème 30. *Pour tout $\theta \in \Theta$ et tout $u > 0$,*

$$\mathbb{P}_\theta(\sup_{t\in\mathbb{R}} \sqrt{n}|F_n(t) - F_\theta(t)| \leqslant u) \to \mathsf{K}(u) \ ,$$

où K est la fonction de répartition de loi de Kolmogorov définie par

$$\mathsf{K}(u) = 1 + 2\sum_{k=1}^{+\infty} (-1)^k \mathrm{e}^{-k^2t^2} \ .$$

Remarque 31. *Une version non-asymptotique de ce résultat a été obtenue par Kieffer, Dvoretski et Wolfowitz, puis raffinée de manière à obtenir des constantes précises par Massart.*

Théorème 32 (Inégalité DKW-M)**.** *Pour tout $\theta \in \Theta$ et tout $u > 0$, on a*

$$\mathbb{P}_\theta\left(\sup_{t\in\mathbb{R}} \sqrt{n}|F_n(t) - F_\theta(t)| > u\right) \leqslant 2\mathrm{e}^{-2u^2} \ .$$

Ce résultat est remarquable car les constantes sont exactes au premier ordre.

Le test d'adéquation de Kolmogorov-Smirnov consiste à rejeter $H_0 : \theta = \theta_0$ si $G(Z_n, \theta_0) > \mathsf{K}^{-1}(1-\alpha)$, où

$$G(Z_n, \theta) = \sup_{t\in\mathbb{R}} \sqrt{n}|F_n(t) - F_\theta(t)| \ .$$

Proposition 33. *Le test d'adéquation de Kolmogorov-Smirnov est consistant et de niveau asymptotique α.*

La preuve est laissée là encore à titre d'exercice.

L'avantage du test de Kolmogorov-Smirnov sur le test du χ^2 est qu'il est toujours consistant. Il ne demande pas de choisir a priori une partition. C'est encore un test non-paramétrique car il est valable pour l'ensemble des lois sur $\mathbb{R}$. Pour pouvoir mettre en place en pratique le test, il est nécessaire de pouvoir évaluer $\sup_{t\in\mathbb{R}} |F_n(t) - F_{\theta_0}(t)|$. Or, comme F_{θ_0} est croissante et F_n est constante sur les intervalles $[X_{i:n}, X_{i+1:n}[$, égale à i/n, on a

$$\begin{aligned}\sup_{t\in\mathbb{R}} |F_n(t) - F_{\theta_0}(t)| &= \max_{i=1,\dots,n} \max(|F_n(X_{i:n}) - F_{\theta_0}(X_{i:n})|, |F_n(X_{i:n}^-) - F_{\theta_0}(X_{i:n}^-)|) \\ &= \max_{i=1,\dots,n} \max\left(\left|\frac{i}{n} - F_{\theta_0}(X_{i:n})\right|, \left|\frac{i-1}{n} - F_{\theta_0}(X_{i:n}^-)\right|\right) \ .\end{aligned}$$

2.4 Tests du rapport de vraisemblance

Le premier résultat est une conséquence immédiate du théorème de Wilks.

Théorème 34. *Soit $\theta_0 \in \Theta$. Le test d'adéquation de Wilks est défini par*

$$\phi_n(Z_n) = \mathbf{1}_{\{2n(\ell(Z_n,\widehat{\theta}_n)-\ell(Z_n,\theta_0))>\chi^2_{1-\alpha}(d)\}} \ .$$

Si le modèle $(\mathbb{P}_\theta, \theta \in \Theta)$ est régulier, alors le test est consistant et de niveau asymptotique α.

La preuve du résultat est laissée en exercice. On peut généraliser le théorème de Wilks à certains cas d'hypothèses H_0 composites.

Definition 35 (Contraintes régulières)**.** *Soit $\boldsymbol{h} = (h_1, \dots, h_r)^T$ un vecteur de r fonctions $h_i : \mathbb{R}^d \to \mathbb{R}$. La fonction $\boldsymbol{h}$ définit une contrainte régulière d'ordre r si*

1. *$\boldsymbol{h}$ est de classe $\mathcal{C}^1$ sur Θ.*
2. *Pour tout $\theta \in \Theta$, la Jacobienne $J_{\boldsymbol{h}}(\theta)$ est de rang r.*

Le théorème suivant étend le théorème de Wilks.

Théorème 36. *Soit $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^r$ une fonction définissant une contrainte régulière. On veut tester $H_0 : \boldsymbol{h}(\theta) = 0$ contre $H_1 : \boldsymbol{h}(\theta) \neq 0$. Le test du rapport de vraisemblance de H_0 contre H_1 est défini par*

$$\phi_{rv,n}(Z_n) = \mathbf{1}_{\{2n(\ell(Z_n,\widehat{\theta}_n)-\ell(Z_n,\widehat{\theta}_{0,n}))>\chi^2_{1-\alpha}(r)\}} \ ,$$

où $\widehat{\theta}_n$ est une suite consistante d'estimateurs du maximum de vraisemblance, $\widehat{\theta}_n \in argmax_{\theta\in\Theta} \ell_n(Z_n, \theta)$ et $\widehat{\theta}_{0,n}$ est une suite d'estimateurs du maximum de vraisemblance sur le sous modèle induit par Θ_0 : $\widehat{\theta}_{0,n} \in argmax_{\theta\in\Theta_0} \ell_n(Z_n, \theta)$.

Si le modèle $(\mathbb{P}_\theta, \theta \in \Theta)$ est régulier, alors le test est consistant et de niveau asymptotique α.

Remarque 37. *Le test est appelé test du rapport de vraisemblance car il rejette H_0 si $\Lambda_n < c_\alpha$, avec*

$$\Lambda_n = \frac{\sup_{\theta\in\Theta_0} L(Z_n, \theta)}{\sup_{\theta\in\Theta} L(Z_n, \theta)}, \qquad c_\alpha = \mathrm{e}^{-\chi^2_{1-\alpha}(r)/(2n)} \ .$$

La statistique Λ_n est appelée rapport de vraisemblance généralisé.

Chapitre 3

Concentration de la moyenne empirique

Une inégalité de concentration pour une variable aléatoire X est une borne sur les probabilités $\mathbb{P}(X - \mathbb{E}[X] > t)$, pour tout $t > 0$. C'est donc naturellement un outil privilégié pour donner des intervalles de confiance et construire des tests sur des paramètres de la forme $\theta = \mathbb{E}[T(X)]$. Dans ce chapitre, nous expliquerons d'abord comment démontrer de telles inégalités, notamment lorsque X est la moyenne empirique de variables i.i.d. avant de donner quelques exemples d'application à la construction de tests.

3.1 Méthode de Chernoff

Un outil de base pour obtenir des inégalités de concentration est la méthode de Chernoff qu'on développe dans cette section.

Théorème 38. *Soit X une variable aléatoire telle que $\mathbb{E}[e^{sX}] < \infty$, pour tout $s \in (0, b)$. Alors*

$$\forall t > 0, \qquad \mathbb{P}(X - \mathbb{E}[X] > t) \leqslant e^{-\psi^*(t)} \ ,$$

où $\psi^*(t) = \sup_{s\in(0,b)}\{st - \log \mathbb{E}[e^{s(X-\mathbb{E}[X])}]\}$.

Remarque 39. *Le résultat implique directement le corollaire suivant que nous utiliserons abondamment. Si $f(s) \geqslant \log \mathbb{E}[e^{s(X-\mathbb{E}[X])}]$, et $f^*(t) = \sup_s\{st - f(s)\}$, on a*

$$\forall t > 0, \qquad \mathbb{P}(X - \mathbb{E}[X] > t) \leqslant e^{-f^*(t)} \ .$$

Démonstration. Fixons $t > 0$ et $s \in (0, b)$. Comme la fonction $x \mapsto e^{sx}$ est croissante, on a

$$\mathbb{P}(X - \mathbb{E}[X] > t) = \mathbb{P}(e^{s(X-\mathbb{E}[X])} > e^{st}) \ .$$

D'après l'inégalité de Markov, on a donc

$$\mathbb{P}(X - \mathbb{E}[X] > t) \leqslant \frac{\mathbb{E}[e^{s(X-\mathbb{E}[X])}]}{e^{st}} = e^{-\{st - \log \mathbb{E}[e^{s(X-\mathbb{E}[X])}]\}} \ .$$

Comme le résultat est vrai pour tout $s \in (0, b)$, on peut optimiser en s, ce qui donne le résultat. □

Calculons maintenant la (log)-transformée de Laplace $\psi(s) = \log \mathbb{E}[e^{sX}]$ et la transformée de Fenchel Legendre $\psi^*(t)$ pour quelques lois bien connues.

Exemple 7 (Lois Gaussiennes)**.** *Soit $X \sim N(0, \sigma^2)$. Si $\sigma^2 = 1$, on a*

$$\mathbb{E}[e^{sX}] = \int_{\mathbb{R}} e^{sx - \frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} = e^{\frac{s^2}{2}} \int_{\mathbb{R}} e^{-\frac{(x-s)^2}{2}} \frac{dx}{\sqrt{2\pi}} = e^{\frac{s^2}{2}} \ .$$

Si $\sigma^2 > 0$ est quelconque, on a que $X/\sigma \sim N(0,1)$, donc

$$\forall s > 0, \qquad \mathbb{E}[e^{sX}] = \mathbb{E}[e^{(s\sigma)(X/\sigma)}] = e^{\frac{(s\sigma)^2}{2}} \ .$$

On en déduit que

$$\psi^*(t) = \sup_{s \in \mathbb{R}_+^*} \left\{ st - \frac{s^2\sigma^2}{2} \right\} = \sup_{s \in \mathbb{R}_+^*} \left\{ -\frac{t^2}{2\sigma^2} - \frac{(s\sigma - t/\sigma)^2}{2} \right\} = -\frac{t^2}{2\sigma^2} \ .$$

En particulier, si $X_1, \dots, X_n$ sont i.i.d. $N(\mu, \sigma^2)$, $n^{-1} \sum_{i=1}^n X_i - \mu \sim N(0, \sigma^2/n)$, donc

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu > t \right) \leqslant e^{-\frac{nt^2}{2\sigma^2}} \ .$$

Ceci implique en particulier que

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X] > \sqrt{\frac{2\,\mathrm{Var}(X)t}{n}} \right) \leqslant e^{-t} \ .$$

Exemple 8 (Lois de Poisson)**.** *Soit $X \sim \mathcal{P}(\theta)$. On a, pour tout $s > 0$.*

$$\mathbb{E}[e^{sX}] = e^{-\theta} \sum_{k=0}^{\infty} \frac{e^{sk}\theta^k}{k!} = e^{\theta(e^s - 1)} \ .$$

Rappelons que $\mathbb{E}[X] = \theta = \mathrm{Var}(X)$. On a

$$\mathbb{E}[e^{s(X - \mathbb{E}[X])}] = e^{\theta(e^s - 1 - s)} \ .$$

Soit $f(s) = st - \theta(e^s - 1 - s)$. On a $f'(s) = t - \theta(e^s - 1)$, donc f atteint son maximum en $s = \log(1 + t/\theta)$ et ce maximum vaut

$$(\theta + t)\log(1 + t/\theta) - t = \theta h(t/\theta) \ ,$$

avec $h(u) = (1+u)\log(1+u) - u$. Si $X_1, \dots, X_n$ sont i.i.d. $\mathcal{P}(\theta)$, on a $\sum_{i=1}^n X_i \sim \mathcal{P}(n\theta)$, donc

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \theta > t \right) \leqslant e^{-n\theta h(t/\theta)} \ .$$

On pourra vérifier que

$$\forall t > 0, \qquad h(t) \geqslant \frac{t^2}{2(1 + t/3)} \ .$$

On en déduit que

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \theta > t \right) \leqslant e^{-\frac{nt^2}{2(\theta + t/3)}} \ .$$

Ceci implique aussi

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X] > \sqrt{\frac{2\,\mathrm{Var}(X)t}{n}} + \frac{2t}{3n} \right) \leqslant e^{-t} \ .$$

Exemple 9 (Lois Gamma)**.** *Soit $X \sim \Gamma(a,b)$, la loi de densité par rapport à la mesure de Lebesgue*

$$f(x,(a,b)) = \frac{x^{a-1}e^{-x/b}}{b^a\Gamma(a)} \ .$$

Rappelons que $\mathbb{E}[X] = ab$, $\mathrm{Var}(X) = ab^2$. On a

$$\forall s \in (0,1/b), \qquad \mathbb{E}[\mathrm{e}^{sX}] = \int \frac{x^{a-1}e^{-x(1/b-s)}}{b^a\Gamma(a)}\mathrm{d}x = \frac{1}{(1-bs)^a}$$

Ainsi,

$$\forall s \in (0,1/b), \qquad \mathbb{E}[\mathrm{e}^{s(X-\mathbb{E}[X])}] = \frac{\mathrm{e}^{-sab}}{(1-bs)^a} \ .$$

Soit $f(s) = st + sab + a\log(1-bs)$, on a $f'(s) = t + ab - \frac{ab}{1-bs}$, donc f atteint son maximum en $s = t/(b(ab+t))$ et ce maximum vaut $t/b - a\log(1+t/(ab))$.

Si $X_1, \dots, X_n$ sont i.i.d. de lois $\Gamma(a,b)$, on a $\sum_{i=1}^n X_i \sim \Gamma(na,b)$, donc

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - ab > abt\right) \leqslant \mathrm{e}^{-na(t-\log(1+t))} \ .$$

En développant en série entière, on obtient, pour tout $t \in (0,1)$,

$$t - \log(1+t) = \sum_{t=2}^{+\infty} \frac{(-1)^k t^k}{k} \leqslant \frac{t^2}{2(1-t)} \ .$$

Ainsi,

$$\forall t \in (0,1), \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - ab > abt\right) \leqslant \exp\left(-\frac{nat^2}{2(1-t)}\right) \ .$$

Ceci implique en particulier que

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - ab > \sqrt{\frac{2ab^2t}{n}} + \frac{2bt}{n}\right) \leqslant \mathrm{e}^{-t} \ .$$

Ceci peut être reformulé

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X] > \sqrt{\frac{2\,\mathrm{Var}(X)t}{n}} + \frac{2bt}{n}\right) \leqslant \mathrm{e}^{-t} \ .$$

Exemple 10 (Lois binomiales)**.** *Soit $X \sim B(r,\theta)$, on a*

$$\forall s > 0, \qquad \mathbb{E}[\mathrm{e}^{sX}] = \sum_{k=0}^r \binom{r}{k}(\mathrm{e}^s\theta)^k(1-\theta)^{r-k} = (1-\theta+\theta\mathrm{e}^s)^r \ .$$

Donc

$$\forall s > 0, \qquad \mathbb{E}[\mathrm{e}^{s(X-\mathbb{E}[X])}] = (\mathrm{e}^{-s\theta}(1-\theta+\theta\mathrm{e}^s))^r \ .$$

Soit alors $f(s) = s(t+r\theta) - r\log(1-\theta+\theta\mathrm{e}^s)$. On a

$$f'(s) = t + r\theta - \frac{\theta r\mathrm{e}^s}{1-\theta+\theta\mathrm{e}^s} \ .$$

Donc f atteint son maximum en $s = \log[(r\theta+t)(1-\theta)/\theta(r-r\theta-t)]$ et ce maximum vaut

$$(r\theta+t)\log\left[\frac{r\theta+t}{r\theta}\right]+(r(1-\theta)-t)\log\left[\frac{r(1-\theta)-t}{r(1-\theta)}\right] .$$

En introduisant, pour tout p,q dans $(0,1)$, $\mathrm{kl}(p,q) = p\log(p/q)+(1-p)\log[(1-p)/(1-q)]$, on en déduit que, pour $t \in (0, r(1-\theta))$ $\psi^*(t) = r\,\mathrm{kl}(\theta+t/r,\theta)$.

Si $X_1,\dots,X_n$ sont i.i.d. de loi $B(r,\theta)$, on a $\sum_{i=1}^n X_i \sim B(nr,\theta)$, donc

$$\forall t>0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - r\theta > t\right) \leqslant \mathrm{e}^{-nr\,\mathrm{kl}(\theta+t/r,\theta)} .$$

On peut vérifier que

$$\mathrm{kl}(\theta+t/r,\theta) \geqslant \frac{(t/r)^2}{2}, \qquad \mathrm{kl}(\theta+t/r,\theta) \geqslant \frac{(t/r)^2}{2\theta(1-\theta)+t/r} ,$$

on en déduit

$$\forall t>0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - r\theta > rt\right) \leqslant \mathrm{e}^{-\frac{nrt^2}{2}}, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - r\theta > rt\right) \leqslant \mathrm{e}^{-\frac{nrt^2}{2\theta(1-\theta)+t}} .$$

Ceci implique que, pour tout $t>0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X] > \sqrt{\frac{2rt}{n}}\right) \leqslant \mathrm{e}^{-t}, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - r\theta > \sqrt{\frac{2\,\mathrm{Var}(X)t}{n}}+\frac{t}{n}\right) \leqslant \mathrm{e}^{-t} .$$

Dans tous les exemples, on voit qu'on arrive à obtenir une déviation de la forme

$$\forall t>0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X] > \sqrt{\frac{2\,\mathrm{Var}(X)t}{n}} + C\frac{t}{n}\right) \leqslant \mathrm{e}^{-t} .$$

Ce résultat étant vrai avec une constante $C = 0$ pour la Gaussienne. Ce résultat se réécrit, avec $\sigma^2 = \mathrm{Var}(X)$,

$$\forall t>0, \qquad \mathbb{P}\left(\sqrt{n}\frac{n^{-1}\sum_{i=1}^n X_i - \mathbb{E}[X]}{\sigma} > \sqrt{2t} + \frac{Ct}{\sigma\sqrt{n}}\right) \leqslant \mathrm{e}^{-t} .$$

Ce type de résultat précise le théorème de la limite centrale, en montrant pour ces lois que la statistique de ce théorème dévie de 0 comme une Gaussienne standard, à un terme correctif $Ct/\sigma\sqrt{n}$ près.

3.2 Approches génériques

Les exemples précédents montrent qu'on pourrait utiliser la méthode de Chernoff pour majorer les quantiles de lois usuelles. Cela dit, les quantiles de ces lois usuelles étant facilement disponibles sur R ou python, ce nest pas la principale utilisation des inégalités de concentration en statistiques (il est toutefois intéressant de comparer dans ces modèles les valeurs des quantiles, de leur

majoration obtenue par inégalité de concentration et de leur approximation obtenue par théorème limite du chapitre précédent). La puissance de la méthode de Chernoff est qu'elle permet d'obtenir des résultats non-asymptotiques, c'est à dire valables pour toute valeur de n, sous des conditions sur la loi des variables aléatoires qui sont moins restrictives que de préciser un modèle paramétrique pour la loi de la variable aléatoire. L'utilité des résultats non-asymptotique sera illustrée l'année prochaine dans les cours d'apprentissage et de statistiques en grande dimension.

Dans cette section, on va donc donner des conditions "génériques" sur la loi d'une variable aléatoire permettant de montrer des inégalités de concentration. On illustrera systématiquement les résultats la moyenne empirique de variables indépendantes.

3.2.1 Variables aléatoires sous Gaussiennes

Definition 40. *Une variable aléatoire X est dite sous Gaussienne si elle vérifie, pour une certaine constante v^2 :*

$$\forall s > 0, \qquad \mathbb{E}[e^{s(X-\mathbb{E}[X])}] \leqslant e^{\frac{s^2 v^2}{2}} \quad .$$

On écrit alors $X \in \mathrm{sGau}(v^2)$ et on appelle v^2 le proxy pour la variance de X.

Nous avons vu dans l'exemple 7 que les variables Gaussiennes étaient sous-Gaussiennes avec leur vraie variance comme proxy.

Exercice : Vérifier que, si $X \sim \mathrm{sGau}(v^2)$, on a $\sigma^2 \leqslant v^2$.

En appliquant la méthode de Chernoff, on obtient facilement le résultat suivant.

Proposition 41. *Si $X \in \mathrm{sGau}(v^2)$, alors*

$$\forall t > 0, \qquad \mathbb{P}(X - \mathbb{E}[X] > t) \leqslant e^{-t^2/(2v^2)}, \qquad \mathbb{P}(X - \mathbb{E}[X] > \sqrt{2v^2 t}) \leqslant e^{-t} \quad .$$

De plus, si $X_1, \ldots, X_n$ sont indépendantes et sous Gaussiennes, il en va de même pour la moyenne empirique comme le montre le résultat suivant.

Proposition 42 (Tensorisation pour les variables sous Gaussiennes)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et sous Gaussiennes telles que $X_i \in \mathrm{sGau}(v_i^2)$ pour tout $i \in \{1, \ldots, n\}$. Alors $\sum_{i=1}^n X_i \in \mathrm{sGau}(v^2)$, avec $v^2 = \sum_{i=1}^n v_i^2$. En particulier, on a donc,*

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > \sqrt{\frac{2v^2 t}{n}}\right) \leqslant e^{-t} \quad .$$

La démonstration est élémentaire et laissée en exercice.

Exemples de variables aléatoires sous Gaussiennes

En plus des variables aléatoires Gaussiennes, nous avons vu dans l'exemple 10 que les variables binomiales étaient sous-Gaussiennes (de proxy pour la variance $v^2 = r$). Donnons un autre exemple.

Proposition 43. *Soit ε une variable aléatoire de Rademacher, c'est à dire de loi uniforme sur $\{-1, 1\}$. Alors, $\varepsilon \in \mathrm{sGau}(1)$.*

Démonstration. On a clairement, pour tout $s \in \mathbb{R}$, $\mathbb{E}[e^{s\varepsilon}] = (e^s + e^{-s})/2$. On développe en série entière cette expression pour obtenir

$$\mathbb{E}[e^{s\varepsilon}] = \sum_{k=0}^{+\infty} \frac{s^{2k}}{(2k)!} \ .$$

On vérifie alors que, pour tout k, $(2k)! \geqslant 2^k k!$, et on en déduit

$$\mathbb{E}[e^{s\varepsilon}] \leqslant \sum_{k=0}^{+\infty} \frac{(s^2/2)^k}{k!} = e^{s^2/2} \ .$$

□

On peut utiliser la proposition 43 pour démontrer que toutes les variables aléatoires presque sûrement bornées sont sous Gaussiennes. Ce résultat est connu sous le nom de lemme d'Hoeffding. Pour procéder, on va utiliser le résultat suivant qui utilise le principe de symétrisation.

Lemme 44 (symétrisation)**.** *Soit X une variable aléatoire et $s > 0$ tel que $\mathbb{E}[e^{sX}] < \infty$. Alors, si X' est une variable aléatoire indépendante de X et de même loi que X et ε est une variable aléatoire de Rademacher, indépendante de X et X', on a*

$$\mathbb{E}[e^{s(X-\mathbb{E}[X])}] \leqslant \mathbb{E}[e^{s\varepsilon(X-X')}] \ .$$

Le nom du lemme vient du fait qu'on majore la transformée de Laplace de $(X - \mathbb{E}[X])$ par cette de la variable aléatoire symétrique $\varepsilon(X - X')$.

Démonstration. On utilise les faits suivants :

1. $\mathbb{E}[X] = \mathbb{E}[X'|X]$ qui est vrai par indépendance de X et X'.
2. $X - X'$ a même loi que $\varepsilon(X - X')$, ce qui se voit par exemple en calculant la transformée de Laplace de ces deux variables.

On a alors, en utilisant la convexité de $x \mapsto e^{sx}$ et l'inégalité de Jensen,

$$\mathbb{E}[e^{s(X-\mathbb{E}[X])}] = \mathbb{E}[e^{s\mathbb{E}[X-X'|X]}] \leqslant \mathbb{E}[e^{s(X-X')}] = \mathbb{E}[e^{s\varepsilon(X-X')}] \ .$$

□

Lemme 45 (Hoeffding)**.** *Soit X une variable aléatoire à valeur dans $[a, b]$. Alors $X \in \mathrm{sGau}((b-a)^2)$.*

Remarque 46. *Cette version du résultat d'Hoeffding est sous-optimale, on peut montrer avec une preuve plus subtile que $X \in \mathrm{sGau}((b-a)^2/4)$, voir le lemme 49 à la fin de cette section. Le résultat optimal est intéressant car il redonne une version précisée du théorème limite dans le cas où les variables sont de loi de Bernoulli de paramètre $1/2$. Cette version est beaucoup plus élémentaire à obtenir et largement suffisante dans beaucoup d'applications théoriques.*

Démonstration. D'après la proposition 43, on a

$$\mathbb{E}[e^{s\varepsilon(X-X')}|X, X'] \leqslant e^{s^2(X-X')^2/2} \ .$$

Or comme X et X' appartiennent à $[a, b]$, on a $(X - X')^2 \leqslant (b - a)^2$, donc

$$\mathbb{E}[e^{s\varepsilon(X-X')}|X, X'] \leqslant e^{s^2(b-a)^2/2} \ .$$

On déduit le résultat du lemme de symétrisation. □

En combinant les résultats précédents, on peut montrer le résultat dont la preuve est laissée en exercice (on pourra utiliser la version du lemme d'Hoeffding donnée au lemme 49).

Théorème 47 (Inégalité d'Hoeffding)**.** *Si $X_1, \dots, X_n$ sont des variables aléatoires indépendantes, X_i étant à valeurs dans $[a_i, b_i]$. Alors on a*

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > \sqrt{\frac{v^2 t}{2n}}\right) \leqslant \mathrm{e}^{-t} \ ,$$

avec $v^2 = n^{-1}\sum_{i=1}^{n}(b_i - a_i)^2$.

Les variables bornées ne sont pas les seules sous Gaussiennes. Le résultat suivant donne une condition simples sur les moments d'une variables aléatoire permettant de montrer qu'elle est sous Gaussienne.

Proposition 48. *Supposons que X est centrée et que ses moments sont "sous géométriques" au sens où il existe $b > 0$ tel que*

$$\forall k \geqslant 2, \qquad |\mathbb{E}[X^k]| \leqslant b^k \ ,$$

alors X est sous Gaussienne, de proxy pour la variance $2b^2$, c'est à dire que

$$\mathbb{E}[\mathrm{e}^{sX}] \leqslant \mathrm{e}^{s^2 b^2} \ .$$

En particulier, si $X_1, \dots, X_n$ sont i.i.d. de même loi que X, on a

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > 2b\sqrt{\frac{t}{n}}\right) \leqslant \mathrm{e}^{-t} \ .$$

Démonstration. L'existence de $\mathbb{E}[\mathrm{e}^{sX}]$ pour tout $s > 0$ vient de la condition sur les moments et du théorème de Fubbini-Tonelli. Soit $s > 0$, on a par le lemme de symétrisation

$$\mathbb{E}[\mathrm{e}^{sX}] \leqslant \mathbb{E}[\mathrm{e}^{s\varepsilon(X - X')}] \ .$$

La variable $\varepsilon(X - X')$ étant symétrique, on a, pour tout k entier, $\mathbb{E}[(\varepsilon(X - X'))^{2k+1}] = 0$. On en déduit, en développant en série entière

$$\mathbb{E}[\mathrm{e}^{sX}] \leqslant \sum_{k=0}^{+\infty} \frac{s^{2k}\mathbb{E}[(X - X')^{2k}]}{(2k)!} \ .$$

Or

$$\mathbb{E}[(X - X')^{2k}] = \binom{2k}{k}\mathbb{E}[X^k]^2 + 2\sum_{\ell=0}^{k-1}\binom{2k}{\ell}\mathbb{E}[X^{2\ell}]\mathbb{E}[X^{2(k-\ell)}] \leqslant b^{2k}2^k \ .$$

Donc, en utilisant l'inégalité $(2k)! \geqslant 2^k k!$,

$$\mathbb{E}[\mathrm{e}^{sX}] \leqslant \sum_{k=0}^{+\infty} \frac{s^{2k}b^{2k}}{k!} = \mathrm{e}^{s^2 b^2} \ .$$

□

Terminons cette section par la preuve du lemme d'Hoeffding avec les constantes optimales.

Lemme 49 (Hoeffding)**.** *Soit X une variable aléatoire à valeur dans $[a,b]$. Alors $X \in \mathrm{sGau}((b-a)^2/4)$.*

Démonstration. Soit $\psi(s) = \log \mathbb{E}[\mathrm{e}^{sX}]$, on a

$$\psi'(s) = \frac{\mathbb{E}[X\mathrm{e}^{sX}]}{\mathbb{E}[\mathrm{e}^{sX}]}, \qquad \psi''(s) = \mathbb{E}\left[X^2 \frac{\mathrm{e}^{sX}}{\mathbb{E}[\mathrm{e}^{sX}]}\right] - \left(\mathbb{E}\left[X \frac{\mathrm{e}^{sX}}{\mathbb{E}[\mathrm{e}^{sX}]}\right]\right)^2 .$$

On en déduit que $\psi(0) = \psi'(0) = 0$ et que ψ'' est la variance de la loi $\mathbb{P}$ de densité $\mathrm{d}\mathbb{P}(x) = \frac{\mathrm{e}^{sx}}{\mathbb{E}[\mathrm{e}^{sX}]} \cdot \mathrm{d}\mathbb{P}_X(x)$. La loi $\mathbb{P}$ étant absolument continue par rapport à la loi de X, elle est à support dans $[a,b]$, donc

$$\psi''(s) = \mathrm{Var}_{\mathbb{P}}(X) \leqslant \mathbb{E}_{\mathbb{P}}\left[\left(X - \frac{b-a}{2}\right)^2\right] \leqslant \frac{(b-a)^2}{4} .$$

Ainsi,

$$\begin{aligned}\psi(s) = \psi(s) - \psi(0) &= \int_0^s \psi'(t)\mathrm{d}t = \int_0^s (\psi'(t) - \psi'(0))\mathrm{d}t = \int_0^s \int_0^t \psi''(u)\mathrm{d}u\mathrm{d}t \\ &\leqslant \int_0^s \int_0^t \frac{(b-a)^2}{4}\mathrm{d}u\mathrm{d}t = \frac{(b-a)^2}{4}\int_0^s t\mathrm{d}t = \frac{(b-a)^2}{8} .\end{aligned}$$

□

3.2.2 Variables aléatoires sous Poissonniennes

Le résultat sur les variables sous Gaussiennes permet de préciser le théorème de la limite centrale quand le proxy pour la variance et la variance sont comparables. Or, l'exemple des lois Bernoulli (binomiales avec $r = 1$ nous montre que la variance $\theta(1-\theta)$ peut être beaucoup plus petite que le proxy pour la variance 1 si θ est proche de 0 ou 1. Le but de cette section est de donner des conditions sous lesquelles retrouver un résultat plus précis.

Definition 50. *Une variable aléatoire X est dite sous Poissonnienne s'il existe $v^2 > 0$ et $b \geqslant 0$ tels que*

$$\forall s > 0, \qquad \log \mathbb{E}[\mathrm{e}^{s(X-\mathbb{E}[X])}] \leqslant \frac{v^2}{b^2}(\mathrm{e}^{bs} - 1 - bs) .$$

On notera alors $X \in \mathrm{sPoi}(v^2, b)$. Si $b = 0$, la condition précédente doit être comprise comme

$$\forall s > 0, \qquad \log \mathbb{E}[\mathrm{e}^{s(X-\mathbb{E}[X])}] \leqslant \frac{v^2 s^2}{2} .$$

Autrement dit, elle signifie que la variable $X \in \mathrm{sGau}(v^2)$.

On a vu que les variables aléatoires de Poisson étaient sous Poissonniennes. Comme pour les variables sous Gaussiennes, on peut donner une condition sur les moments d'une variable aléatoire permettant de montrer qu'elle est sous Poissonnienne.

Proposition 51. *Soit X une variable aléatoire telle que*

$$\forall k \geqslant 2, \qquad \mathbb{E}[(X - \mathbb{E}[X])^2 (X - \mathbb{E}[X])_+^{k-2}] \leqslant v^2 b^{k-2} \ ,$$

alors $X \in \mathrm{sPoi}(v^2, b)$.

Démonstration. Comme $(\mathrm{e}^x - 1 - x)/x^2$ est croissante, on a, pour tout $x \in \mathbb{R}$,

$$\mathrm{e}^x \leqslant 1 + x + \sum_{k=2}^{+\infty} \frac{x^2 x_+^{k-2}}{k!} \ .$$

Soit $s > 0$, on a donc

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{s(X-\mathbb{E}[X])}] &= 1 + \sum_{k=2}^{+\infty} \frac{s^k \mathbb{E}[(X - \mathbb{E}[X])^2 (X - \mathbb{E}[X])_+^{k-2}]}{k!} \\ &\leqslant 1 + \frac{v^2}{b^2} \sum_{k=2}^{+\infty} \frac{s^k b^k}{k!} = 1 + \frac{v^2}{b^2}(\mathrm{e}^{sb} - 1 - sb) \ .\end{aligned}$$

On conclut la preuve avec l'inégalité de convexité $\log(1 + x) \leqslant x$. □

La proposition 51 assure en particulier que toute variable aléatoire telle que $X - \mathbb{E}[X] \leqslant b$ presque sûrement est sous Poissonnienne. Elle implique en particulier le résultat suivant.

Proposition 52. *Soit X une variable aléatoire positive alors*

$$-X \in \mathrm{sPoi}(\mathrm{Var}(X), \mathbb{E}[X]) \ .$$

Démonstration. Comme $X \geqslant 0$ presque sûrement, on a $0 \leqslant (\mathbb{E}[X] - X)_+ \leqslant \mathbb{E}[X]$ presque sûrement, donc, pour tout $k \geqslant 2$, $\mathbb{E}[(X - \mathbb{E}[X])^2 (X - \mathbb{E}[X])_+^{k-2}] \leqslant \mathrm{Var}(X)\mathbb{E}[X]^{k-2}$. □

Les variables sous Poissonniennes vérifient le résultat suivant.

Proposition 53 (Concentration des variables sous Poissonniennes)**.** *Si $X \in \mathrm{sPoi}(v^2, b)$, alors on a*

$$\forall t > 0, \qquad \log \mathbb{P}(X - \mathbb{E}[X] > t) \leqslant -\frac{v^2}{b^2} h\left(\frac{bt}{v^2}\right) \leqslant -\frac{t^2}{2(v^2 + bt/3)} \ . \tag{3.1}$$

Ce résultat implique que

$$\forall t > 0, \qquad \mathbb{P}\left(X - \mathbb{E}[X] > \sqrt{2v^2 t} + \frac{2bt}{3}\right) \leqslant \mathrm{e}^{-t} \ .$$

Démonstration. Supposons d'abord que $b = 1$. D'après l'exemple 8, la transformée de Laplace de X est majorée par celle d'une variable aléatoire de Poisson de paramètre v^2. Les calculs menés dans cet exemple assurent alors que $\psi^*(t) \geqslant v^2 h(t/v^2)$. Ainsi, on a

$$\forall t > 0, \qquad \mathbb{P}(X - \mathbb{E}[X] > t) \leqslant \mathrm{e}^{-v^2 h\left(\frac{t}{v^2}\right)} \ .$$

Supposons maintenant $b > 0$ quelconque. Alors $X/b \in \mathrm{sPoi}(v^2/b^2, 1)$, donc, pour tout $t > 0$,

$$\mathbb{P}(X - \mathbb{E}[X] > t) = \mathbb{P}\left(\frac{X}{b} - \mathbb{E}\left[\frac{X}{b}\right] > \frac{t}{b}\right) \leqslant \mathrm{e}^{-\frac{v^2}{b^2} h\left(\frac{t}{b(v^2/b^2)}\right)} = \mathrm{e}^{-\frac{\sigma^2}{b^2} h\left(\frac{bt}{\sigma^2}\right)} \quad .$$

Comme de plus,

$$\forall t > 0, \qquad h(t) \geqslant \frac{t^2}{2(1 + t/3)} \quad .$$

On en déduit

$$-\frac{\sigma^2}{b^2} h\left(\frac{bt}{\sigma^2}\right) \leqslant -\frac{\sigma^2}{b^2} \frac{b^2 t^2}{2\sigma^4(1 + bt/3\sigma^2)} = -\frac{t^2}{2(\sigma^2 + bt/3)} \quad .$$

□

Proposition 54 (Tensorisation pour les variables sous Poissonniennes)**.** *Soient $X_1, \ldots, X_n$ des variables indépendantes telles que $X_i \in \mathrm{sPoi}(v_i^2, b_i)$. Alors $\sum_{i=1}^n X_i \in \mathrm{sPoi}(\sum_{i=1}^n v_i^2, b_{\max})$, où $b_{\max} = \max\{b_i, i = 1, \ldots, n\}$. En particulier,*

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_i] > \sqrt{\frac{2v^2 t}{n}} + \frac{2b_{\max} t}{3n}\right) \leqslant \mathrm{e}^{-t} \quad ,$$

où $v^2 = n^{-1}\sum_{i=1}^n v_i^2$.

Démonstration. La clé de la preuve est de remarquer que la fonction

$$x \mapsto \frac{\mathrm{e}^x - 1 - x}{x^2}$$

est croissante sur $\mathbb{R}_+$ (comme le montre facilement un développement en série entière. On en déduit que, pour tout i et tout $s > 0$,

$$\frac{\mathrm{e}^{sb_i} - 1 - sb_i}{b_i^2} \leqslant \frac{\mathrm{e}^{sb_{\max}} - 1 - sb_{\max}}{b_{\max}^2} \quad .$$

On a alors, en notant $nv^2 = \sum_{i=1}^n v_i^2$,

$$\begin{aligned}
\log \mathbb{E}[\mathrm{e}^{s\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}] &= \sum_{i=1}^n \log \mathbb{E}[\mathrm{e}^{s(X_i - \mathbb{E}[X_i])}] \\
&\leqslant \sum_{i=1}^n v_i^2 \frac{\mathrm{e}^{sb_i} - 1 - sb_i}{b_i^2} \\
&\leqslant nv^2 \frac{\mathrm{e}^{sb_{\max}} - 1 - sb_{\max}}{b_{\max}^2} \quad .
\end{aligned}$$

Ceci démontre le premier résultat. Le second est alors une conséquence du résultat de concentration des variables sous Poissonniennes. □

3.2.3 Inégalité de Bernstein

Definition 55. *Une variable aléatoire X vérifie la condition de Bernstein s'il existe $v^2 > 0$, $b \geqslant 0$ tels que*

$$\forall s \in (0, 1/b), \qquad \log \mathbb{E}[e^{s(X-\mathbb{E}[X])}] \leqslant \frac{v^2 s^2}{2(1-bs)} \ .$$

On écrit alors $X \in \mathrm{Bern}(v^2, b)$.

Exemple 11. *Supposons que $X \in \mathrm{sPoi}(v^2, b)$ alors, comme, pour tout $k \geqslant 2$, $k! \geqslant 2 3^{k-2}$, on a*

$$\log \mathbb{E}[e^{s(X-\mathbb{E}[X])}] \leqslant \frac{v^2}{b^2} \sum_{k=2}^{+\infty} \frac{b^k s^k}{k!} = \frac{v^2 s^2}{2(1-bs/3)} \ .$$

Autrement dit, $X \in \mathrm{Bern}(v^2, b/3)$.

Proposition 56. *Posons $h_1(x) = 1 + x - \sqrt{1+2x}$. Si $X \in \mathrm{Bern}(v^2, b)$, on a*

$$\forall t > 0, \qquad \mathbb{P}(X - \mathbb{E}[X] > t) \leqslant e^{-\frac{v^2}{b^2} h_1\left(\frac{bt}{v^2}\right)}, \qquad \mathbb{P}(X - \mathbb{E}[X] > \sqrt{2v^2 t} + bt) \leqslant e^{-t} \ .$$

Exercice 1. *Montrer que la proposition 56 implique la dernière inégalité de l'équation 3.1.*

Démonstration. Fixons $t > 0$. On va appliquer la méthode de Chernoff et on définit pour cela la fonction

$$f(s) = st - \frac{v^2 s^2}{2(1-bs)} = s\left(t + \frac{v^2}{2b}\right) + \frac{v^2}{2b^2} - \frac{v^2}{2b^2(1-bs)} \ .$$

On a

$$f'(s) = t + \frac{v^2}{2b} - \frac{v^2}{2b(1-bs)^2} \ .$$

Donc f atteint son maximum en

$$s = \frac{1}{b}\left(1 - \frac{1}{\sqrt{1 + 2bt/v^2}}\right) \ .$$

Ce maximum vaut

$$f^*(t) = \frac{v^2}{b^2}\left(1 + \frac{bt}{v^2} - \sqrt{1 + \frac{2bt}{v^2}}\right) = \frac{v^2}{b^2} h_1\left(\frac{bt}{v^2}\right) \ .$$

Le premier résultat est donc une conséquence de la méthode de Chernoff.

Pour le second, on écrit d'abord $h_1(x) = \frac{1+2x}{2} - \sqrt{1+2x} + \frac{1}{2} = \frac{(\sqrt{1+2x}-1)^2}{2}$, de sorte que, pour tout $u > 0$, on a $h_1(x) = u$ si $x = [(1+\sqrt{2u})^2 - 1]/2 = \sqrt{2u} + u$. On a donc, en posant $h_1^{-1}(u) = \sqrt{2u} + 2u$ et

$$u = \frac{v^2}{b} h_1^{-1}\left(\frac{b^2 u}{v^2}\right) = \sqrt{2v^2 u} + bu \ ,$$

d'après le premier résultat

$$\forall u > 0, \qquad \mathbb{P}(X - \mathbb{E}[X] > \sqrt{2v^2 u} + bu) \leqslant e^{-u} \ .$$

□

Proposition 57 (Tensorization)**.** *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes telles que, pour tout $i \in \{1, \dots, n\}$, $X_i \in \mathrm{Bern}(v_i^2, b_i)$. Alors en posant $v^2 = n^{-1} \sum_{i=1}^n v_i^2$, $b = \max\{b_i, i = 1, \dots, n\}$, on a $\sum_{i=1}^n X_i \in \mathrm{Bern}(nv^2, b)$. En particulier, on a donc,*

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > \sqrt{\frac{2v^2 t}{n}} + \frac{bt}{n} \right) \leqslant \mathrm{e}^{-t} \ .$$

Démonstration. Remarquons que toutes les variables $X_i \in \mathrm{Bern}(v_i^2, b)$, de sorte que, par indépendance, pour tout $s \in (0, 1/b)$,

$$\log \mathbb{E}\big[\mathrm{e}^{s \sum_{i=1}^n (X_i - \mathbb{E}[X_i])}\big] = \sum_{i=1}^n \log \mathbb{E}\big[\mathrm{e}^{s(X_i - \mathbb{E}[X_i])}\big] \leqslant \sum_{i=1}^n \frac{s^2 v_i^2}{2(1-bs)} = \frac{n s^2 v^2}{2(1-bs)} \ .$$

Ceci montre le premier point et le second vient de l'inégalité de Bernstein donnée à la proposition 56. □

La proposition suivante donne alors une condition suffisante sur les moments de X permettant de vérifier que $X \in \mathrm{Bern}(v^2, b)$.

Proposition 58. *Soit X une variable aléatoire telle que,*

$$\forall k \geqslant 2, \qquad \mathbb{E}[(X - \mathbb{E}[X])^2 (X - \mathbb{E}[X])_+^{k-2}] \leqslant \frac{v^2 b^{k-2} k!}{2} \ .$$

Alors $X \in \mathrm{Bern}(v^2, b)$.

Démonstration. Comme $(\mathrm{e}^x - 1 - x)/x^2$ est croissante, on a

$$\forall x \in \mathbb{R}, \qquad \sum_{k=2}^{+\infty} \frac{x^k}{k!} \leqslant \sum_{k=2}^{+\infty} \frac{x^2 x_+^k}{k!} \ .$$

Soit $s \in (0, 1/b)$. On a donc

$$\mathbb{E}\big[\mathrm{e}^{s(X - \mathbb{E}[X])}\big] = 1 + \frac{s^2 v^2}{2} \sum_{k=2}^{+\infty} (sb)^{k-2} = 1 + \frac{s^2 v^2}{2(1-bs)} \ .$$

On conclut alors avec l'inégalité de convexité $1 + x \leqslant \mathrm{e}^x$. □

Un exemple d'application de cette propriété classique (et très utile pour la suite) est donné par le résultat suivant.

Proposition 59. *Soit $X \in \mathrm{sGau}(v^2)$, alors $(X - \mathbb{E}[X])^2 \in \mathrm{Bern}(16v^4, 2v^2)$. En particulier, si $X_1, \dots, X_n$ sont des variables indépendantes telles que pour tout $i \in \{1, \dots, n\}$, $\sigma_i^2 = \mathrm{Var}(X_i)$, $X_i \in \mathrm{sGau}(v^2)$, on a*

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \big\{ (X_i - \mathbb{E}[X_i])^2 - \sigma_i^2 \big\} > 2v^2 \left(\sqrt{\frac{8t}{n}} + \frac{t}{n} \right) \right) \leqslant \mathrm{e}^{-t} \ .$$

Démonstration. On va évaluer les moments de $(X - \mathbb{E}[X])^2$. On utilise pour cela le résultat élémentaire suivant dont la démonstration est laisée à titre d'exercice.

Lemme 60. *Soit X une variable aléatoire* positive *d'espérance finie, alors*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t) \mathrm{d}t \ .$$

Soit maintenant $k \geqslant 2$, on a, d'après le lemme,

$$\mathbb{E}[(X - \mathbb{E}[X])^{2k}] = \int_0^{+\infty} \mathbb{P}(|X - \mathbb{E}[X]| > t^{1/(2k)}) \mathrm{d}t \ .$$

Comme $X \in \mathrm{sGau}(v^2)$, pour tout $t > 0$, on a $\mathbb{P}(|X - \mathbb{E}[X]| > t^{1/(2k)}) \leqslant 2\mathrm{e}^{-t^{1/k}/(2v^2)}$, donc

$$\mathbb{E}[(X - \mathbb{E}[X])^{2k}] \leqslant 2 \int_0^{+\infty} \mathrm{e}^{-t^{1/k}/(2v^2)} \mathrm{d}t \ .$$

On fait alors le changement de variable $u = t^{1/k}/(2v^2)$ qui donne

$$\mathbb{E}[(X - \mathbb{E}[X])^{2k}] \leqslant 2(2^k v^{2k})k \int_0^{+\infty} u^{k-1} \mathrm{e}^{-u} \mathrm{d}u = \frac{16v^4(2v^2)^{k-2}k!}{2} \ .$$

Le résultat découle alors de la proposition 58 car on a

$$((X - \mathbb{E}[X])^2 - \sigma^2)_+ \leqslant (X - \mathbb{E}[X])^2 \ .$$

□

3.3 Applications à la construction de tests

3.3.1 Modèle de translation

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de densité $f(\cdot - \theta)$, où θ est un paramètre inconnu et f est une densité inconnue. On a donc $X_i = \theta + \varepsilon_i$, avec $\varepsilon_1, \ldots, \varepsilon_n$, avec ε de densité f. On suppose en outre que

$$\mathbb{E}[\varepsilon] = 0, \qquad \varepsilon \in \mathrm{sGau}(1) \ .$$

On veut tester $H_0 : \theta \leqslant 0$ contre $H_1 : \theta > 0$. On a le résultat suivant.

Proposition 61. *Le test $\phi(Z_n) = \mathbf{1}_{\{\widehat{\theta} > c_\alpha\}}$, où $\widehat{\theta} = n^{-1} \sum_{i=1}^n X_i$ et $c_\alpha = \sqrt{2\log(1/\alpha)/n}$ est de niveau α. De plus, pour tout $\theta > c_\alpha$, on a $\beta_\phi(\theta) \geqslant 1 - \mathrm{e}^{-n(\theta - c_\alpha)^2/2}$.*

Démonstration. D'après la proposition 42,

$$\forall t > 0, \qquad \mathbb{P}_\theta\left(\widehat{\theta} - \theta > \sqrt{\frac{2t}{n}}\right) \leqslant \mathrm{e}^{-t} \ . \tag{3.2}$$

En appliquant ce résultat avec $\theta = 0$ et $t = \log(1/\alpha)$, on démontre le premier résultat. Pour le second, en vérifiant que $-X$ est également dans $\mathrm{sGau}(1)$, on a

$$\forall t > 0, \qquad \mathbb{P}_\theta\left(\widehat{\theta} - \theta < -t\right) \leqslant \mathrm{e}^{-nt^2/2} \ . \tag{3.3}$$

On en déduit que, pour tout $\theta > c_\alpha = \sqrt{2\log(1/\alpha)/n}$

$$\mathbb{P}_\theta\left(\widehat{\theta} > \sqrt{\frac{2\log(1/\alpha)}{n}}\right) \geqslant 1 - \mathrm{e}^{-n(\theta - c_\alpha)^2/2} \ .$$

□

Exercice : Etudier le cas ou $\varepsilon \in \mathrm{sPoi}(1)$.

3.3.2 Modèle de translation et d'échelle

Soit $X_1, \dots, X_n$ un échantillon aléatoire de densité $\sigma^{-1} f((x-\theta)/\sigma)$, où f est une densité inconnue, $\theta \in \mathbb{R}$ et $\sigma^2 > 0$ sont des paramètres inconnus. Autrement dit, on peut écrire pour tout $i \in \{1, \dots, n\}$,

$$X_i = \theta + \sigma \varepsilon_i \ ,$$

où $\varepsilon_1, \dots, \varepsilon_n$ sont i.i.d. de densité f. On souhaite tester $H_0 : \sigma \leqslant 1$ contre $H_1 : \sigma > 1$. On fait tout au long de cette section l'hypothèse que ε est une variable aléatoire centrée, de variance 1 et $\varepsilon \in \mathrm{sGau}(v^2)$ où v^2 est une constante connue. On a le résultat suivant

Proposition 62. *Soit $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ la moyenne empirique et $\widehat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$ la variance empirique. Soit $\phi(Z_n) = \mathbf{1}_{\{\widehat{\sigma}_n^2 > c_\alpha\}}$, où la valeur critique*

$$c_\alpha = 2v^2 \left(\sqrt{\frac{8 \log(\alpha^{-1})}{n}} + \frac{\log(\alpha^{-1})}{n} \right) \ .$$

Alors ϕ est de niveau α. De plus, si $h_1(x) = 1 + x - \sqrt{1+2x}$, pour tout $\sigma^2 > c_\alpha$, on a

$$\beta_\phi(\theta) \geqslant 1 - \mathrm{e}^{-\frac{n}{64} h_1\left(\frac{8(\sigma^2 - c_\alpha)}{v^2}\right)} \ .$$

La preuve est laissée en exercice. On pourra s'appuyer sur le résultat suivant sur la concentration de la variance empirique.

Proposition 63. *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes et de même loi, dans $\mathrm{sGau}(v^2)$. Soit $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ et*

$$\widehat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2 = n^{-1} \sum_{i=1}^n (X_i - \mathbb{E}[X])^2 - (\overline{X}_n - \mathbb{E}[X])^2 \ .$$

Alors, si $\sigma^2 = \mathrm{Var}(X)$, on a, pour tout $t > 0$,

$$\mathbb{P}\left(\widehat{\sigma}_n^2 - \sigma^2 > 2\sigma^2 v^2 \left(\sqrt{\frac{8t}{n}} + \frac{t}{n} \right) \right) \leqslant \mathrm{e}^{-t}, \qquad \mathbb{P}\left(\widehat{\sigma}_n^2 - \sigma^2 < -2\sigma^2 \left(v^2 \sqrt{\frac{2t}{n}} + \frac{4t}{3n} \right) \right) \leqslant \mathrm{e}^{-t} \ .$$

La seconde inégalité se réécrit, en rappelant que $h_1(x) = 1 + x - \sqrt{1+2x}$,

$$\forall t > 0, \qquad \mathbb{P}\left(\widehat{\sigma}_n^2 - \sigma^2 < -t \right) \leqslant \mathrm{e}^{-\frac{n}{64} h_1\left(\frac{8t}{v^2}\right)} \ .$$

Démonstration. Le premier résultat est une conséquence de la proposition 59 et du fait que $(\overline{X}_n - \mathbb{E}[X])^2 \geqslant 0$ presque sûrement.

Pour le second, comme $(X_i - \mathbb{E}[X])^2 \geqslant 0$, on a $-(X_i - \mathbb{E}[X])^2 \in \mathrm{sPoi}(4\sigma^4 v^4, \sigma^2)$ d'après la proposition 52. D'après la proposition 54, on en déduit

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X])^2 - \sigma^2 < -\sigma^2 \left(v^2 \sqrt{\frac{8t}{n}} + \frac{2t}{3n} \right) \right) \leqslant \mathrm{e}^{-t} \ .$$

D'autre part, on a aussi, d'après la proposition 42, on a

$$\forall t > 0, \qquad \mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X]| > \sqrt{\frac{2v^2 t}{n}} \right) \leqslant 2\mathrm{e}^{-t} \ .$$

En combinant ces deux résultats, on obtient la seconde partie de théorème. □

Chapitre 4

Tests optimaux

Dans ce chapitre, on s'intéresse à la possibilité de construire des tests optimaux du point de vue de Neyman-Pearson. On introduit pour cela la définition suivante.

Definition 64. *Soit $\alpha \in (0,1)$, soient $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$ deux hypothèses disjointes et soit $\mathcal{T}(\alpha)$ l'ensemble des tests de niveau α de H_0 contre H_1. Un test ϕ est dit uniformément plus puissant au niveau α,* UPP(α) *si $\phi \in \mathcal{T}(\alpha)$ et si, pour tout $\phi' \in \mathcal{T}(\alpha)$, on a*

$$\forall \theta \in \Theta_1, \qquad \beta_\phi(\theta) \geqslant \beta_{\phi'}(\theta) \ .$$

Il est rare, mais pas impossible de pouvoir construire des tests uniformément plus puissants. Chaque fois que c'est possible, ceux-ci sont basés sur les rapports de vraisemblance. Nous avons déjà présenté ce test au chapitre 2. Il est basé sur la statistique suivante : soit $L(\theta) = \mathcal{L}(Z_n, \theta) = \prod_{i=1}^n f(X_i, \theta)$ la vraisemblance de l'observation, la statistique du rapport de vraisemblance est définie par

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)} \ .$$

L'idée du test est que, si H_0 est vraie, Λ doit être proche de 1 tandis que si H_1 est vraie, $\Lambda < 1$, le test du rapport de vraisemblance consiste donc à rejeter H_0 lorsque $\Lambda < c$, où c est un seuil critique qu'il reste à calibrer. On pourra se référer à la section 2.4 du chapitre 2 pour des résultats généraux sur le comportement limite de cette statistique.

4.1 Tests d'hypothèses simples

Dans cette section, on considère le cadre élémentaire, dans lequel $\Theta_i = \{\theta_i\}$ pour $i = \{0,1\}$. Le test du rapport de vraisemblance revient dans ce cas à rejeter H_0 lorsque le rapport de la vraisemblance en θ_1 sur la vraisemblance en θ_0 dépasse un seuil critique à définir. Formellement, on introduit la densité de l'échantillon sous la loi θ : $\mathcal{L}(z_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$. La vraisemblance de l'échantillon en tout point $\theta \in \Theta$ est alors définie par $L(\theta) = \mathcal{L}(Z_n, \theta)$. Le test du rapport de vraisemblance rejette H_0 si $R = R(Z_n) = L(\theta_1)/L(\theta_0) > c$, où c est un seuil qu'il reste à déterminer. Par convention, on posera $R = +\infty$ si $L(\theta_0) = 0$.

Dans le cas du test d'hypothèses simples, on choisit c de manière à assurer que le test est bien de niveau α, on pose

$$c_\alpha = \inf\{c \in \mathbb{R} : \mathbb{P}_{\theta_0}(R > c) \leqslant \alpha\} \ .$$

Lorsque la loi $\mathbb{P}_\theta$ est discrète, il est possible que $\mathbb{P}_{\theta_0}(R > c_\alpha) < \alpha$, dans ce cas, il est commode d'introduire le test randomisé suivant. Si $R > c_\alpha$, $\phi = 1$ (c'est à dire qu'on rejette H_0), et si $R < c_\alpha$, $\phi = 0$ (c'est à dire qu'on ne rejette pas H_0). Si $R = c_\alpha$, ϕ est une variable aléatoire de Bernoulli de paramètre γ, indépendante de Z_n, γ étant choisi de façon à ce que

$$\mathbb{P}_{\theta_0}(R > c_\alpha) + \gamma \mathbb{P}_{\theta_0}(R = c_\alpha) = \alpha \ . \tag{4.1}$$

Si $R = c_\alpha$, on choisit donc de rejeter H_0 en tirant notre décision au hasard avec une probabilité de succès γ. Dans la pratique, on considère dans l'immense majorité des exemples seulement des tests purs. L'intérêt des tests randomisés est théorique, puisque les tests purs sont des cas particuliers de tests randomisés. En effet, le test randomisé ϕ convient aussi dans le cas continu, mais il se simplifie alors en un test pur $\phi = \mathbf{1}_{\{R<c_\alpha\}}$. Un autre avantage théorique à considérer cette généralisation est que le test randomisé ϕ du rapport de vraisemblance est de taille α, c'est à dire qu'il vérifie, grâce à l'équation (4.1), $\mathbb{E}_{\theta_0}[\phi(Z_n, U)] = \alpha$. Le théorème suivant garantit l'optimalité du test ϕ parmi les tests de niveau α.

Théorème 65 (Neyman-Pearson)**.** *Le test randomisé du rapport de vraisemblance est* $\mathrm{UPP}(\alpha)$*. De plus, tout test* ϕ' $\mathrm{UPP}(\alpha)$ *vérifie, pour* μ*-presque tout* $z_n \in \mathcal{X}^n$*,*

$$\phi'(z_n) = \begin{cases} 1 & \text{si } \mathcal{L}(z_n, \theta_1) > c_\alpha \mathcal{L}(z_n, \theta_0) \\ 0 & \text{si } \mathcal{L}(z_n, \theta_1) < c_\alpha \mathcal{L}(z_n, \theta_0) \end{cases} \ .$$

Remarque 66. *La seconde partie du théorème assure que le test du rapport de vraisemblance est l'unique test* $\mathrm{UPP}(\alpha)$ *(à des modifications sur des ensembles de mesure nulle près) lorsque les variables sont continues. Lorsqu'elles sont discrètes, il peut exister différents tests* $\mathrm{UPP}(\alpha)$ *mais ceux ci ne diffèrent que sur l'événement* $\mathcal{L}(z_n, \theta_1) = c_\alpha \mathcal{L}(z_n, \theta_0)$*.*

Preuve. Soit ϕ' un test de niveau α et soit $\mathcal{X}_0 = \{z_n \in \mathcal{X}^n : \mathcal{L}(z_n, \theta_0) = 0\}$. Par convention, on a $R = +\infty$, donc $R > c_\alpha$ si $z_n \in \mathcal{X}_0$, de sorte que $\phi \geqslant \phi'$ si $z_n \in \mathcal{X}_0$. Supposons donc que $z_n \notin \mathcal{X}_0$. On a alors

$$\begin{aligned} &(\phi(z_n) - \phi'(z_n))\mathcal{L}(z_n, \theta_1) \\ &\quad = (\phi(z_n) - \phi'(z_n))(R(z_n) - c)\mathcal{L}(z_n, \theta_0) + c(\phi(z_n) - \phi'(z_n))\mathcal{L}(z_n, \theta_0) \ . \end{aligned}$$

L'intégrale

$$\int (\phi(z_n) - \phi'(z_n))\mathcal{L}(z_n, \theta_0) \mathrm{d}\mu(z_n) = \alpha - \mathbb{E}_{\theta_0}[\phi'(Z_n, U)] \geqslant 0 \ .$$

D'autre part, pour tout $z_n \in \mathcal{X}^n$, on a

$$\begin{aligned} R(z_n) > c \quad &\Rightarrow \quad \phi(z_n) = 1 \quad \Rightarrow \quad \phi(z_n) - \phi'(z_n) \geqslant 0 \ , \\ R(z_n) < c \quad &\Rightarrow \quad \phi(z_n) = 0 \quad \Rightarrow \quad \phi(z_n) - \phi'(z_n) \leqslant 0 \ . \end{aligned}$$

Ainsi, on a, pour tout $x \in \mathcal{X}$,

$$(\phi(z_n) - \phi'(z_n))(R(z_n) - c) \geqslant 0 \ .$$

Donc

$$\mathbb{E}_{\theta_1}[(\phi - \phi')\mathbf{1}_{Z_n \notin \mathcal{X}_0}] = \mathbb{E}_{\theta_0}[(\phi - \phi')(R - c)] + c(\alpha - \mathbb{E}_{\theta_0}[\phi']) \geqslant 0 \ .$$

Ainsi,

$$\mathbb{E}_{\theta_1}[\phi - \phi'] = \mathbb{E}_{\theta_1}[(\phi - \phi')\mathbf{1}_{Z_n \notin \mathcal{X}_0}] + \mathbb{E}_{\theta_1}[(\phi - \phi')\mathbf{1}_{Z_n \in \mathcal{X}_0}] \geqslant 0 \ .$$

Si ϕ' est UPP(α), alors toutes les inégalités sont des égalités. En particulier,

$$(\phi(z_n) - \phi'(z_n))\mathbf{1}_{z_n \in \mathcal{X}_0} = 0, \qquad \mathbb{P}_{\theta_1} - p.s. \ .$$

Donc, si $z_n \in \mathcal{X}_0$ et $\mathcal{L}(z_n, \theta_1) > 0$, on a $\phi'(z_n) = 1 = \phi(z_n)$. De plus, si $z_n \notin \mathcal{X}_0$, on a, pour μ presque tout $z_n \notin \mathcal{X}_0$,

$$\begin{aligned} 0 &= (\phi(z_n) - \phi'(z_n))(R(z_n) - c_\alpha)\mathcal{L}(z_n, \theta_0) \\ &= (\phi(z_n) - \phi'(z_n))(\mathcal{L}(z_n, \theta_1) - c_\alpha \mathcal{L}(z_n, \theta_0)) \ , \end{aligned}$$

ce qui conclut la preuve du second point. □

4.2 Tests unilatères

Dans cette section, on suppose que $\Theta \subset \mathbb{R}$ et pour $\theta_0 \in \theta$, on s'intéresse à tester les hypothèses

$$H_0 : \theta \leqslant \theta_0, \quad \text{contre} \quad H_1 : \theta > \theta_0 \ .$$

Definition 67. *La famille de lois de densités $\{f(x, \theta), \theta \in \Theta\}$ est à rapports de vraisemblances strictement croissants en la statistique $T = T(Z_n)$ si, pour tout θ et θ' de Θ tels que $\theta > \theta'$, il existe une fonction φ strictement croissante telle que*

$$\frac{\mathcal{L}(z_n, \theta)}{\mathcal{L}(z_n, \theta')} = \varphi(T(z_n)) \ .$$

Remarque 68. *Dans cette définition, la fonction φ dépend typiquement de θ et θ'. Notez que cette définition est indépendante de θ_0 et donc des hypothèses H_0 et H_1 qu'on cherche à tester ici.*

Les familles à rapports de vraisemblances monotones vérifient le lemme suivant qui sera utile.

Lemme 69. *Supposons la famille de lois de densités $\{f(x, \theta), \theta \in \Theta\}$ à rapports de vraisemblances strictement croissants en la statistique $T = T(Z_n)$ et soit ψ une fonction croissante. Alors la fonction $h : \theta \mapsto \mathbb{E}_\theta[\psi \circ T(Z_n)]$ est croissante.*

Preuve. Soit $\theta > \theta'$. Soit $A = \{z_n \in \mathcal{X}^n : \mathcal{L}(z_n, \theta') > \mathcal{L}(z_n, \theta)\}$. La famille étant à rapports de vraisemblances strictement croissants en $T(Z_n)$, il existe une fonction φ strictement croissante telle que $\mathcal{L}(z_n, \theta)/\mathcal{L}(z_n, \theta') = \varphi(T(z_n))$. Le seuil $s = \sup\{t \in \mathbb{R} : \varphi(t) < 1\}$ vérifie donc

$$A = \{z_n \in \mathcal{X}^n : \mathcal{L}(z_n, \theta') > \mathcal{L}(z_n, \theta)\} = \{z_n \in \mathcal{X}^n : T(z_n) < s\} \ .$$

La fonction ψ étant croissante, on a donc

$$\inf_{z_n \in A^c} \psi \circ T(z_n) \geqslant \sup_{z_n \in A} \psi \circ T(z_n) \ .$$

Alors,

$$\begin{aligned} h(\theta) - h(\theta') &= \int \psi \circ T(z_n)(\mathcal{L}(z_n,\theta) - \mathcal{L}(z_n,\theta'))\mathrm{d}\mu(z_n) \\ &\geqslant \inf_{z_n \in A^c} \psi \circ T(z_n) \int_{A^c} (\mathcal{L}(z_n,\theta) - \mathcal{L}(z_n,\theta'))\mathrm{d}\mu(z_n) \\ &\quad + \sup_{z_n \in A} \psi \circ T(z_n) \int_{A} (\mathcal{L}(z_n,\theta) - \mathcal{L}(z_n,\theta'))\mathrm{d}\mu(z_n) \\ &= (\inf_{z_n \in A^c} \psi \circ T(z_n) - \sup_{z_n \in A} \psi \circ T(z_n)) \int_{A^c} (\mathcal{L}(z_n,\theta) - \mathcal{L}(z_n,\theta'))\mathrm{d}\mu(z_n) \geqslant 0 \ . \end{aligned}$$

□

Nous sommes désormais en mesure d'établir le résultat suivant pour le test du rapport de vraisemblance dans le cadre des familles à rapports de vraisemblances monotones.

Théorème 70. *Supposons la famille de lois de densités $\{f(x,\theta), \theta \in \Theta\}$ à rapports de vraisemblances strictement croissants en la statistique $T = T(Z_n)$, soit $\alpha \in (0,1)$ et soit $\theta_0 \in \Theta$. Alors, il existe un test ϕ uniformément plus puissant au niveau α des hypothèses*

$$H_0 : \theta \leqslant \theta_0 \quad \textit{contre} \quad H_1 : \theta > \theta_0 \ .$$

Par exemple, le test randomisé ϕ suivant convient :

$$\phi = \begin{cases} 1 & \textit{si } T(z_n) > c \\ 0 & \textit{si } T(z_n) < c \\ U & \textit{si } T(z_n) = c \end{cases} ,$$

où $U \sim B(\gamma)$ est indépendante de Z_n et $c \in \mathbb{R}$ et $\gamma \in [0,1]$ vérifient l'équation

$$\mathbb{P}_{\theta_0}(T(Z_n) > c) + \gamma \mathbb{P}_{\theta_0}(T(Z_n) = c) = \alpha \ . \tag{4.2}$$

De plus, la fonction puissance de ce test est une fonction croissante sur $\{\theta \in \Theta : \beta_\phi(\theta) < 1\}$.

Preuve. Introduisons la fonction

$$\psi(t) = \begin{cases} 0 & \text{si } t < c \\ \gamma & \text{si } t = c \\ 1 & \text{si } t > c \end{cases} .$$

La fonction ψ est croissante, donc, d'après le Lemme 69, la fonction $\beta_\phi(\theta) = \mathbb{E}_\theta[\psi \circ T(Z_n)]$ est croissante, ce qui montre le second point du théorème. On en déduit ensuite que, pour tout $\theta \leqslant \theta_0$, on a

$$\beta_\phi(\theta) \leqslant \beta_\phi(\theta_0) = \alpha \ .$$

La dernière égalité étant due à l'hypothèse 4.2. Le test randomisé ϕ est donc de niveau α.

Soit maintenant $\theta_1 > \theta_0$ et ϕ' un autre test de niveau α. Considérons les hypothèses $H_0' : \theta = \theta_0$ et $H_1' : \theta = \theta_1$. Pour ces hypothèses simples, le théorème 65 assure que le test randomisé du rapport de vraisemblance est UPP(α), en particulier, comme ϕ' est aussi de niveau α pour tester H_0' contre H_1', sa puissance en θ_1 est supérieure à celle de ϕ'. Le test randomisé du rapport de vraisemblance vaut

$$\phi^*(z_n) = \begin{cases} 0 & \text{si } R(z_n) > c^* \\ U & \text{si } R(z_n) = c^* \\ 1 & \text{si } R(z_n) < c^* \end{cases} .$$

Ici, $U \sim B(\gamma^*)$ est indépendante de Z_n et c^* et γ^* sont solutions de

$$\mathbb{P}_{\theta_0}(R(Z_n) > c^*) + \gamma^* \mathbb{P}_{\theta_0}(R(Z_n) = c^*) = \alpha \ .$$

La famille étant à rapports de vraisemblance strictement croissants en $T(Z_n)$, il existe une fonction φ strictement croissante telle que $R(z_n) = \varphi(T(z_n))$, donc on a

$$\{z_n \in \mathcal{X}^n : R(z_n) > c^*\} = \{z_n \in \mathcal{X}^n : T(z_n) > c\} \ ,$$
$$\{z_n \in \mathcal{X}^n : R(z_n) = c^*\} = \{z_n \in \mathcal{X}^n : T(z_n) = c\} \ .$$

Ainsi $\phi^* = \phi$, donc ϕ est plus puissant que ϕ' en θ_1. Comme ceci est vrai pour tout $\theta_1 > \theta_0$ et tout test ϕ' de niveau α, on a bien ϕ est UPP(α). □

4.3 Tests bilatères

Dans cette section, on suppose que $X_1, \dots, X_n$ est un échantillon aléatoire de $\mathrm{N}(\theta, 1)$ avec $\theta \in \mathbb{R}$ et on s'intéresse au test de $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$. On a $\mathcal{L}(z_n, \theta) = (2\pi)^{-n/2} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2}$, de sorte que, si $\theta > \theta'$, on a

$$\frac{\mathcal{L}(z_n, \theta)}{\mathcal{L}(z_n, \theta')} = \mathrm{e}^{\frac{1}{2}\sum_{i=1}^n (X_i - \theta')^2 - (X_i - \theta)^2} = \mathrm{e}^{\frac{(\theta')^2 - \theta^2}{2}} \mathrm{e}^{n(\theta - \theta')\hat{\theta}_n} \ ,$$

où $\hat{\theta}_n = n^{-1} \sum_{i=1}^n X_i$. Ainsi, la famille $\mathrm{N}(\theta, 1)$ est à rapports de vraisemblances strictement croissants en la statistique $\hat{\theta}_n$. Il s'en suit que le test du rapport de vraisemblance de H_0 contre $H_1' : \theta = \theta_1$, avec $\theta_1 > 0$, est de la forme $\phi(Z_n) = \mathbf{1}_{\{\hat{\theta}_n > c\}}$. De plus, s'il existait un test ϕ_h qui soit UPP(α) de H_0 contre H_1, il serait également UPP(α) de H_0 contre H_1'. Par le théorème de Neyman-Pearson, on aurait donc $\phi_h(z_n) = \phi(z_n)$ presque-partout. Or la puissance de ce test en tout $\theta < \theta_0$ est strictement inférieure à α, donc à celle du test trivial $\phi_\alpha = B$, avec $B \sim B(\alpha)$ indépendante de Z_n, donc ce test n'est pas uniformément plus puissant pour H_0 contre H_1. Finalement, la discussion précédente montre *qu'il n'existe pas de tests* UPP(α) de H_0 contre H_1.

La propriété UPP étant trop forte, on la relâche dans cette section pour celle d'uniformément plus puissant parmi les tests sans biais de niveau α.

Definition 71. *Un test ϕ de $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$ est dit* sans biais *de niveau α s'il est de niveau α et vérifie*

$$\forall \theta \in \Theta_1, \qquad \beta_\phi(\theta) \geqslant \alpha \ .$$

Soit $\mathcal{T}_S(\alpha)$ l'ensemble des tests sans biais de niveau α. Un test ϕ est dit uniformément plus puissant parmi les tests sans biais de niveau α, UPPS(α) *si $\phi \in \mathcal{T}_S(\alpha)$ et si, pour tout $\phi' \in \mathcal{T}_S(\alpha)$ et tout $\theta \in \Theta_1$, $\beta_\phi(\theta) \geqslant \beta_{\phi'}(\theta)$.*

Notons que, lorsqu'il en existe, un test UPP(α) est sans biais de niveau α, puisqu'il est plus puissant que le test trivial $\phi \sim B(\alpha)$.

Soit ϕ un test sans biais de niveau α de $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$ dans le modèle Gaussien $\{\mathrm{N}(\theta,1)^{\otimes n}, \theta \in \mathbb{R}\}$. On a donc $\beta_\phi(0) \leqslant \alpha$ et pour tout $\theta \neq 0$, en notant $\mathcal{L}(z_n,\theta)$ la densité du produit de n Gaussiennes $\mathrm{N}(\theta,1)$,

$$\alpha \leqslant \beta_\phi(\theta) = \int_{\mathbb{R}} \phi(z_n)\mathcal{L}(z_n,\theta)\mathrm{d}z_n \ .$$

La fonction β_ϕ étant continue, on en déduit donc que, pour tout test ϕ sans biais,

$$\int_{\mathbb{R}} \phi(z_n)\mathcal{L}(z_n,0)\mathrm{d}z_n = \alpha \ . \tag{4.3}$$

De plus, la fonction β_ϕ étant dérivable et atteignant son minimum en 0, on en déduit que, pour tout test ϕ sans biais,

$$\int_{\mathbb{R}} \phi(z_n)\partial_\theta\mathcal{L}(z_n,0)\mathrm{d}z_n = 0 \ .$$

Un test ϕ UPPS(α), s'il en existe, maximise donc en tout point $\theta_1 \neq 0$ la fonction puissance

$$\beta_\phi(\theta_1) = \int_{\mathbb{R}} \phi(z_n)\mathcal{L}(z_n,\theta_1)\mathrm{d}z_n \ ,$$

sous les contraintes

$$\int_{\mathbb{R}} \phi(z_n)\mathcal{L}(z_n,0)\mathrm{d}z_n = \alpha, \qquad \int_{\mathbb{R}} \phi(z_n)\partial_\theta\mathcal{L}(z_n,0)\mathrm{d}z_n = 0 \ .$$

Soient κ_1 et κ_2 deux paramètres à régler, pouvant dépendre de θ_1, et soit ϕ le test suivant

$$\phi(z_n) = \begin{cases} 1 & \text{si } \mathcal{L}(z_n,\theta_1) > \kappa_1\mathcal{L}(z_n,0) + \kappa_2\partial_\theta\mathcal{L}(z_n,0) \\ 0 & \text{si } \mathcal{L}(z_n,\theta_1) \leqslant \kappa_1\mathcal{L}(z_n,0) + \kappa_2\partial_\theta\mathcal{L}(z_n,0) \end{cases} .$$

Il est facile de voir que ϕ maximise la fonction

$$\beta_\phi(\theta_1) - \kappa_1\beta_\phi(\theta_0) - \kappa_2 \int_{\mathbb{R}} \phi(z_n)\partial_\theta\mathcal{L}(z_n,0)\mathrm{d}z_n \ .$$

Donc, si ϕ vérifie les contraintes, ϕ est UPPS(α). Par construction $\phi = 1$ sur un ensemble R de la forme $\mathrm{e}^{C_1\hat{\theta}_n} > C_2\hat{\theta}_n + C_3$. Ces ensembles, selon la valeur des constantes C_1, C_2 et C_3 peuvent être d'une des formes suivantes :

$$R = \begin{cases} \{z_n : \hat{\theta}(z_n) > c_1\} \ , \\ \{z_n : \hat{\theta}(z_n) < c_2\} \ , \\ \{z_n : \hat{\theta}(z_n) \notin [c_3,c_4]\} \ . \end{cases}$$

On a déjà vu que, dans les deux premiers cas, le test ϕ associé ne peut être UPPS(α), on a donc nécessairement $\phi = \mathbf{1}_{\hat{\theta}_n \notin [C_1,C_2]}$. Pour être de niveau α, C_1 et C_2 doivent vérifier

$$\Phi(\sqrt{n}C_1) + 1 - \Phi(\sqrt{n}C_2) = \alpha \ .$$

Comme de plus $\hat{\theta}_n$ suit, sous H_0, une loi Gaussienne $\mathrm{N}(0, 1/n)$, la seconde contrainte s'écrit ensuite

$$n \int_{-\infty}^{C_1} z \mathrm{e}^{-\frac{nz^2}{2}} \mathrm{d}z + n \int_{C_2}^{+\infty} z \mathrm{e}^{-\frac{nz^2}{2}} \mathrm{d}z = 0$$

Comme la fonction $z\mathrm{e}^{-\frac{nz^2}{2}}$ est impaire, ceci implique $C_1 = -C_2$, donc $\Phi(\sqrt{n}C_1) = 1 - \Phi(\sqrt{n}C_2)$ et finalement, avec la première contrainte

$$C_2 = \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}, \qquad C_1 = -C_2 \ .$$

On en déduit que le test suivant est UPPS(α).

$$\phi = \mathbf{1}_{\{|\hat{\theta}_n| > \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}\}} \ .$$

Chapitre 5

Théorie de la décision

5.1 Règle de décision, perte, risque

Un estimateur est une statistique arbitraire. L'estimateur est pertinent s'il est proche en un certain sens de la fonction à estimer. Lorsqu'on dispose de plusieurs estimateurs, il est également naturel de se demander lequel est le meilleur, ou s'il en existe un optimal par rapport à un critère. Les notions de perte et de risque permettent de définir des mesures quantitatives de la performance d'estimateurs et de tests, qui sont des cas particuliers de règles de décision.

Definition 72. *Soit $\mathcal{P}$ un modèle statistique et A un ensemble mesurable appelé ensemble d'actions. Une règle de décision est une application mesurable $a : \mathcal{X}^n \to A$. Soit $\ell : A \times \Theta \to \mathbb{R}_+$ une fonction telle que, pour tout $\theta \in \Theta$, $\ell(\cdot, \theta)$ est mesurable. ℓ est appelée fonction de perte et le risque de la règle de décision a est défini comme l'espérance*

$$R_\ell(a, \theta) = \mathbb{E}_\theta[\ell(a(Z_n), \theta)] \ .$$

L'estimation ponctuelle peut être formulée comme un problème de décision. Soit $g : \Theta \to g(\Theta)$ une fonction à estimer (on suppose $g(\Theta)$ mesurable). Soit $A = g(\Theta)$. Un estimateur $T(Z_n)$ est alors une règle de décision au sens de la définition 72. Dans ce cadre, la fonction de perte est en général une distance $\ell(a, \theta) = d(a, g(\theta))$ ou une puissance de cette distance comme la perte quadratique.

De même, un test peut être vu comme une règle de décision à valeur dans l'ensemble $A = \{0, 1\}$ muni de la tribu de ses parties. Ici, une fonction de perte va généralement vérifier, pour tout $a \neq a' \in A$,

$$\ell(a, \theta) \begin{cases} = 0 & \forall \theta \in \Theta_a \\ > 0 & \forall \theta \in \Theta_{a'} \end{cases} \ .$$

Ainsi, tout ce que nous allons développer dans cette section peut s'appliquer aussi bien aux problèmes d'estimation que de tests.

Definition 73. *Une règle de décision $a : \mathcal{X}^n \to A$ est dite non-admissible pour la perte ℓ s'il existe une autre règle a' telle que*

$$R_\ell(a', \theta) \begin{cases} \leqslant R_\ell(a, \theta) & \forall \theta \in \Theta \ , \\ < R_\ell(a, \theta) & \textit{pour au moins un } \theta \in \Theta \ . \end{cases}$$

Une règle admissible est une règle qui n'est pas non-admissible. Elle est alors maximale pour l'ordre partiel induit par la fonction de risque.

La propriété d'admissibilité est un prérequis minimal sur une règle de décision, puisqu'elle ne peut alors être améliorée *uniformément*. En revanche, cette propriété seule n'est pas très intéressante, puisque les estimateurs constants sont en général admissibles. Dans ce chapitre, nous proposons deux critères d'optimalité d'usage courant en statistique. Ces deux critères sont basés sur le fait d'associer à la fonction de risque un réel. Une fois ce réel choisi, les règles de décision sont naturellement ordonnées (la meilleure est celle minimisant le réel associé).

5.2 Approche minimax

Dans l'approche minimax, le réel associé à la fonction est le sup de cette fonction.

Definition 74. *Le risque maximal d'une règle de décision a est par définition le réel*

$$R_{\max}(a) = \sup_{\theta \in \Theta} R_\ell(a, \theta) \ .$$

Le risque minimax pour la perte ℓ est l'infimum sur l'ensemble des règles de décision des risques maximaux :

$$\overline{R}_\ell = \inf_a R_{\max}(a) = \inf_a \sup_{\theta \in \Theta} R_\ell(a, \theta) \ .$$

Une règle a^* est dite minimax si elle est optimale du point de vue du risque maximal, i.e. si

$$R_{\max}(a^*) = \overline{R}_\ell \ .$$

Nous allons maintenant chercher des procédures minimax pour le problème de tests de deux hypothèses simples et faire le lien avec la théorie de Neyman-Pearson.

Soit $Z_n = (X_1, \dots, X_n)$ un échantillon aléatoire d'une loi $\mathbb{P}_\theta$, avec $\theta \in \Theta$. On suppose la loi $\mathbb{P}_\theta$ uniformément continue par rapport à une mesure μ sur $\mathcal{X}$ et on note $f(x, \theta)$ la densité de $\mathbb{P}_\theta$ par rapport à μ. La densité de l'échantillon Z_n est notée $\mathcal{L}(z_n, \theta)$ et la vraisemblance $L(\theta) = \mathcal{L}(Z_n, \theta)$.

Soient θ_0, θ_1 deux éléments de Θ. On veut bâtir une fonction de décision $a : \mathcal{X}^n \to \{\theta_0, \theta_1\}$ qui décide quelle valeur choisir au vue de l'échantillon. On s'autorise ici à considérer les règles de décision randomisées $a(Z_n, U)$ où U est une variable aléatoire indépendante de Z_n. Toute règle de décision a, randomisée ou non, est naturellement associée au test ϕ de $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ défini par

$$\phi(z_n, u) = \mathbf{1}_{\{a(z_n, u) = \theta_1\}} \ .$$

La perte d'une règle de décision a en θ est notée $\ell(a, \theta)$. On se donne ℓ_0 et ℓ_1 deux réels positifs et on prend la fonction de perte

$$\ell(a, \theta) = \ell_0 \mathbf{1}_{\{a = \theta_0, \theta = \theta_1\}} + \ell_1 \mathbf{1}_{\{a = \theta_1, \theta = \theta_0\}} \ . \tag{5.1}$$

Cette fonction vérifie $\ell(\theta, \theta) = 0$ pour tout $\theta \in \Theta$, et $\ell(\theta, \theta') > 0$ pour tout $\theta \neq \theta'$.

Definition 75. *Le risque de la règle a est défini par*

$$R(a,\theta) = \mathbb{E}_\theta[\ell(a,\theta)] = \int \ell(a(z_n),\theta)\mathcal{L}(z_n,\theta)\mu(\mathrm{d}z_n) \ .$$

En utilisant les définitions de a et ℓ, on obtient

$$R(a,\theta_0) = \ell_1 \mathbb{P}_{\theta_0}(a=\theta_1), \qquad R(a,\theta_1) = \ell_0 \mathbb{P}_{\theta_1}(a=\theta_0) \ .$$

Ainsi, si on note β_ϕ la fonction puissance du test associé à a, on a

$$R(a,\theta_0) = \ell_1 \beta_\phi(\theta_0), \qquad R(a,\theta_1) = \ell_0(1-\beta_\phi(\theta_1)) \ .$$

$R(a,\theta_0)$ est donc proportionnelle à l'erreur de première espèce du test ϕ et $R(a,\theta_1)$ est proportionnelle à l'erreur de seconde espèce.

On cherche ici une règle a^* minimisant le risque maximal. On introduit pour cela

$$a^* = \begin{cases} \theta_1 & \text{si } \frac{L(\theta_0)}{L(\theta_1)} < c \text{ ou si } \frac{L(\theta_0)}{L(\theta_1)} = c \text{ et } B = 1 \\ \theta_0 & \text{sinon} \end{cases} , \tag{5.2}$$

où B est une variable aléatoire de Bernoulli $B(\gamma)$ indépendante de Z_n et les paramètres $c \geqslant 0$ et $\gamma \in [0,1)$ sont choisis de façon à ce que

$$\mathbb{P}_{\theta_0}\left(\frac{L(\theta_0)}{L(\theta_1)} < c\right) + \gamma \mathbb{P}_{\theta_0}\left(\frac{L(\theta_0)}{L(\theta_1)} = c\right) = \alpha \ , \tag{5.3}$$

α étant lui même choisi de façon à ce que

$$R(a,\theta_0) = R(a,\theta_1) \ .$$

Autrement dit, c et γ sont choisis de façon à ce que

$$\frac{\ell_1}{\ell_0} = \frac{1 - \mathbb{P}_{\theta_1}\left(\frac{L(\theta_0)}{L(\theta_1)} < c\right) - \gamma \mathbb{P}_{\theta_1}\left(\frac{L(\theta_0)}{L(\theta_1)} = c\right)}{\mathbb{P}_{\theta_0}\left(\frac{L(\theta_0)}{L(\theta_1)} < c\right) + \gamma \mathbb{P}_{\theta_0}\left(\frac{L(\theta_0)}{L(\theta_1)} = c\right)} \ .$$

Comme pour les tests, la version randomisée n'est utile que lorsque les observations sont discrètes. Dans les cas continus, on se contente de choisir c de façon à ce que

$$\frac{\ell_1}{\ell_0} = \frac{1 - \mathbb{P}_{\theta_1}\left(\frac{L(\theta_0)}{L(\theta_1)} < c\right)}{\mathbb{P}_{\theta_0}\left(\frac{L(\theta_0)}{L(\theta_1)} < c\right)} \ .$$

Remarquons que, dans les deux cas, l'équation admet toujours une unique solution. Cela dit, même dans des cas simples, cette équation peut n'être résoluble que numériquement. Il transparait de ces équations que le test $\phi = \mathbf{1}_{\{a=\theta_1\}}$ est le test du rapport de vraisemblance de Neyman-Pearson de niveau α.

Théorème 76. *La règle de décision a^* définie à l'équation* (5.2) *est minimax, c'est à dire que $R_{\max}(a^*) \leqslant R_{\max}(a)$, pour toute règle a.*

Preuve. Remarquons d'abord que le choix de α dans l'équation (5.3) assure que

$$R_{\max}(a^*) = R(a^*, \theta_0) = R(a^*, \theta_1) \ .$$

Soit a une règle de décision. Si $R(a, \theta_0) > R(a^*, \theta_0)$, alors

$$R_{\max}(a) \geqslant R(a, \theta_0) > R(a^*, \theta_0) = R_{\max}(a^*) \ .$$

Supposons donc que $R(a, \theta_0) \leqslant R(a^*, \theta_0)$. Alors le test $\phi = \mathbf{1}_{\{a=\theta_1\}}$ est de niveau α donc, par le théorème de Neyman-Pearson, sa puissance $\beta_\phi(\theta_1) \leqslant \beta_{\phi^*}(\theta_1)$, où $\phi^* = \mathbf{1}_{a^*=\theta_1}$ est le test du rapport de vraisemblance. Comme $\ell_0 > 0$, on a donc

$$\begin{aligned} R_{\max}(a) \geqslant R(a, \theta_1) &= \ell_1(1 - \beta_\phi(\theta_1)) \\ &\geqslant \ell_1(1 - \beta_{\phi^*}(\theta_1)) = R(a^*, \theta_1) = R_{\max}(a^*) \ . \end{aligned}$$

□

5.3 Approche Bayésienne

L'approche Bayésienne est une alternative à l'approche minimax dans laquelle le nombre réel associé à la fonction de risque est l'intégrale du risque par rapport à une mesure de probabilité. On suppose dans cette section que Θ est muni d'une tribu $\mathcal{T}$ telle que les applications ℓ et, pour tout événement B, $\theta \mapsto \mathbb{P}_\theta(B)$ soient mesurables. On déduit de la première condition que, pour toute fonction mesurable positive $f(z, \theta)$, la fonction $\theta \mapsto \mathbb{E}_\theta[f(Z, \theta)]$ est aussi mesurable, et donc, d'après la première condition, que la fonction de risque $\theta \mapsto R_\ell(a(Z), \theta)$ est mesurable.

Definition 77. *Soit* Π *une mesure de probabilité sur* $(\Theta, \mathcal{T})$*, appelée loi a priori. Pour toute règle de décision* a*, le risque de Bayes de* a *par rapport à* Π *est défini par*

$$R_\ell(a, \Pi) = \int_\Theta R_\ell(a, \theta) \Pi(\mathrm{d}\theta) \ .$$

Le risque Bayésien selon Π *est défini comme l'infimum des risques de Bayes des règles de décision :*

$$\mathbb{R}_\ell(\Pi) = \inf_a R_\ell(a, \Pi) \ .$$

Une règle a_Π *est dite Bayésienne selon* Π *si elle atteint le risque Bayésien selon* Π*, i.e., si*

$$R_\ell(a_\Pi, \Pi) = \mathbb{R}_\ell(\Pi) = \inf_a R_\ell(a, \Pi) \ .$$

Comme pour l'approche minimax, on étudie maintenant la forme d'un test Bayésien pour le problème du test d'hypothèses simples. On reprend le cadre de la section précédente en gardant notamment la perte définie à l'équation (5.1).

Proposition 78. *Soit* Π *une loi de probabilité sur* Θ *telle que* $\Pi(\theta_0) = \pi_0 > 0$ *et* $\Pi(\theta_1) = \pi_1 > 0$*. Alors, toute règle* a_Π *satisfaisant les conditions suivantes est Bayésienne :*

$$a_\Pi(z_n) = \begin{cases} \theta_1 & si \ \frac{\mathcal{L}(z_n, \theta_1)}{\mathcal{L}(z_n, \theta_0)} > \frac{\pi_0 \ell_1}{\pi_1 \ell_0} \ , \\ \theta_0 & si \ \frac{\mathcal{L}(z_n, \theta_1)}{\mathcal{L}(z_n, \theta_0)} < \frac{\pi_0 \ell_1}{\pi_1 \ell_0} \ . \end{cases}$$

Remarque 79. *Comme pour les règles minimax, ces règles de décision sont associées à des tests du rapport de vraisemblance. En revanche, il est inutile ici d'introduire des tests randomisés, des tests purs sont suffisants.*

Preuve. Le risque Bayésien d'une règle a s'écrit, d'après le théorème de Fubbini-Tonelli

$$\begin{aligned} R_\ell(a,\Pi) &= \int_\Theta \int_{\mathcal{X}^n} \ell(a(z_n),\theta)\mathcal{L}(z_n,\theta)\mu(\mathrm{d}z_n)\Pi(\mathrm{d}\theta) \\ &= \int_{\mathcal{X}^n} \left(\int_\Theta \ell(a(z_n),\theta)\mathcal{L}(z_n,\theta)\Pi(\mathrm{d}\theta) \right) \mu(\mathrm{d}z_n) \\ &= \int_{\mathcal{X}^n} \left(\pi_0\ell_1\mathcal{L}(z_n,\theta_0)\mathbf{1}_{\{a(z_n)=\theta_1\}} + \pi_1\ell_0\mathcal{L}(z_n,\theta_1)\mathbf{1}_{\{a(z_n)=\theta_0\}} \right) \mu(\mathrm{d}z_n) \ . \end{aligned}$$

Pour minimiser ce risque intégré, il suffit de minimiser, pour presque tout $z_n \in \mathcal{X}^n$, le terme entre parenthèse. Ce terme est clairement toujours supérieur à

$$\min\left(\pi_0\ell_1\mathcal{L}(z_n,\theta_0), \pi_1\ell_0\mathcal{L}(z_n,\theta_1) \right) \ .$$

De plus, toute règle a_Π satisfaisant les conditions de l'énoncé atteint cette borne supérieure. Cela signifie donc d'une part que cette dernière borne est le risque Bayésien et d'autre part que les règles de décision satisfaisant les conditions de l'énoncé sont Bayésiennes. □

Chapitre 6

L'algorithme EM

Dans ce chapitre, on présente un algorithme d'approximation de l'estimateur du maximum de vraisemblance dans le cas où toutes les variables ne sont pas observées. Cette situation se rencontre en effet très fréquemment en pratique, comme par exemple dans l'exemple suivant.

Dans les jeux vidéos en ligne, il est utile d'évaluer le niveau des joueurs. Supposons que le jeu se joue en parties à 2 en face à face, et que l'issue d'une rencontre est la victoire d'un des joueurs. On observe donc typiquement des variables aléatoires de Bernoulli $X_{i,j,k}$ donnant le résultat de la k-ième partie entre les joueurs i et j, à valeur dans $\{0,1\}$, 0 si le joueur i a perdu et 1 s'il a gagné cette k-ième partie. Le résultat de cette partie dépend des niveaux V_i et V_j des joueurs i et j. Un modèle simple pour la loi conditionnelle de $X_{i,j,k}$ sachant V_i et V_j est celui de Bradley-Terry dans lequel $X_{i,j,k}$ sont, conditionnellement au vecteur $V = (V_1, \dots, V_N)$ des valeurs des joueurs, des variables aléatoires indépendantes de loi de Bernoulli de paramètres $V_i/(V_i + V_j)$. Ainsi, en notant $n_{i,j}$ le nombre de parties entre i et j,

$$(X_{i,j,k})_{1 \leqslant i < j < N,\ k \in \{1,\dots,n_{i,j}\}} \text{ sont indépendantes et } \mathbb{P}(X_{i,j,k} = 1 | V) = \frac{V_i}{V_i + V_j} \ .$$

Supposons que $V_1, \dots, V_N$ soient elles-mêmes indépendantes de loi $\mathbb{P}_\theta$, avec $\theta \in \Theta$. On peut alors écrire la densité jointe

$$\mathcal{L}(v, x, \theta) = \prod_{i=1}^{N} f(v_i, \theta) \prod_{1 \leqslant i < j \leqslant N} \prod_{k=1}^{n_{i,j}} \left(\frac{v_i}{v_i + v_j} \right)^{x_{i,j,k}} \left(\frac{v_j}{v_i + v_j} \right)^{1 - x_{i,j,k}} \ .$$

On en déduit la densité des observations

$$\mathcal{L}(x, \theta) = \int \mathcal{L}(v, x, \theta) \prod_{i=1}^{N} \mu(\mathrm{d}v_i) \ .$$

L'algorithme EM s'intéresse typiquement à ce genre de situations, cherchant à maximiser la vraisemblance observée $L(X, \theta)$ à partir de la vraisemblance complète $L(X, V, \theta)$ (plus simple).

6.1 Cadre général

Considérons un échantillon aléatoire de n variables aléatoires. Parmi ces variables n_1 sont observées et $n_2 = n - n_1$ ne sont pas observables. On note $X = (X_1, \dots, X_{n_1})$ les variables observées et $V = (V_1, \dots, V_{n_2})$ les variables non observées.

Supposons que les X_i sont i.i.d. de même densité $f(x, \theta)$, avec $\theta \in \Theta$. Supposons les X_i et V_j mutuellement indépendantes. On note $g(x, \theta)$ la loi jointe des observations X, $h(x, v, \theta)$ la loi jointe de (X, V) et $k(v, \theta|x)$ la loi des variables non observées V sachant les observations X.

Par construction, on a donc

$$k(v, \theta|x) = \frac{h(x, v, \theta)}{g(x, \theta)} \ . \tag{6.1}$$

Definition 80. *La vraisemblance observée est par définition $L(\theta) = g(X, \theta)$, c'est la fonction qu'on cherche à maximiser.*

On appelle vraisemblance complète, et on note $L_c(\theta) = h(X, V, \theta)$.

Le but de l'algorithme EM est donc de maximiser la vraisemblance $L(\theta)$ en utilisant la vraisemblance complète $L_c(\theta)$.

Soit $\theta_0 \in \Theta$. En utilisant la formule de Bayes 6.1, on a

$$\begin{aligned}
\log L(\theta) &= \int (\log g(X, \theta)) k(v, \theta_0|X) \mu(\mathrm{d}v) \\
&= \int (\log h(X, v, \theta) - \log k(v, \theta|X)) k(v, \theta_0|X) \mu(\mathrm{d}v) \\
&= \int (\log h(X, v, \theta)) k(v, \theta_0|X) \mu(\mathrm{d}v) - \int (\log k(v, \theta|X)) k(v, \theta_0|X) \mu(\mathrm{d}v) \\
&= \mathbb{E}_{\theta_0}[L_c(\theta)|X] - \mathbb{E}_{\theta_0}[\log k(V, \theta|X)|X] \ .
\end{aligned} \tag{6.2}$$

6.2 L'algorithme

Définissons maintenant

$$Q(\theta_0, \theta) = Q(\theta_0, \theta, X) = \mathbb{E}_{\theta_0}[L_c(\theta)|X] \ . \tag{6.3}$$

L'algorithme EM est un algorithme itératif dont la mise à jour est définie ainsi.

Definition 81 (Algorithme EM)**.** *Etant donnée la valeur courante $\hat{\theta}_m$ de l'estimateur, on effectue successivement l'étape E de l'algorithme (pour espérance) consistant à calculer la fonction*

$$\forall \theta \in \Theta, \qquad Q(\hat{\theta}_m, \theta) = \mathbb{E}_{\hat{\theta}_m}[L_c(\theta)|X] \ , \tag{6.4}$$

puis l'étape M de l'algorithme (pour maximisation) consistant à mettre à jour l'estimateur en calculant

$$\hat{\theta}_{m+1} \in argmax_{\theta \in \Theta} Q(\hat{\theta}_m, \theta) \ . \tag{6.5}$$

Nous allons démontrer que cette mise à jour fait toujours augmenter la vraisemblance.

Théorème 82. *La suite $\hat{\theta}_m$ introduite à la définition 81 vérifie*

$$\forall m \geqslant 0, \qquad L(\hat{\theta}_{m+1}) \geqslant L(\hat{\theta}_m) \ .$$

Démonstration. Par construction, on a

$$Q(\hat{\theta}_m, \hat{\theta}_{m+1}) \geqslant Q(\hat{\theta}_m, \hat{\theta}_m) \ .$$

D'après l'équation (6.2), il suffit donc pour montrer le théorème de prouver que

$$\mathbb{E}_{\hat{\theta}_m}[\log k(V, \hat{\theta}_m|X)|X] \geqslant \mathbb{E}_{\hat{\theta}_m}[\log k(V, \hat{\theta}_{m+1}|X)|X] \ . \tag{6.6}$$

Or, on a, par concavité de la fonction log, d'après l'inégalité de Jensen, pour tout $\theta, \theta' \in \Theta$,

$$\int \log\left(\frac{k(v,\theta'|X)}{k(v,\theta|X)}\right) k(v,\theta|X)\mu(\mathrm{d}v) \leqslant \log\left(\int \frac{k(v,\theta'|X)}{k(v,\theta|X)} k(v,\theta|X)\mu(\mathrm{d}v)\right) = 0$$

Autrement dit

$$\mathbb{E}_\theta[\log k(V,\theta'|X)] \leqslant \mathbb{E}_\theta[\log k(V,\theta|X)] \ .$$

Ceci étant vrai pour tout θ, θ', on peut l'appliquer à $\theta = \hat{\theta}_m$ et $\theta' = \hat{\theta}_{m+1}$ pour obtenir (6.6) et terminer la preuve du théorème. □

6.3 Exemple 1

Illustrons ce résultat sur un exemple. Supposons qu'on observe des durées de vies de patients. Certains de ces patients étant encore en vie à la fin de l'étude, on sait seulement que cette durée de vie est supérieure à une valeur fixe a. Notons $X_1, \ldots, X_{n_1}$ les durées de vie observées complètement et $V_1, \ldots, V_{n_2}$ les valeurs censurées à la valeur a. Supposons que la densité de X_i est de la forme $f(x-\theta)$, de sorte que la fonction de répartition de cette loi vaut $F(x-\theta)$. Les vraisemblances observées et complètes sont donc respectivement

$$L(\theta) = (1 - F(a-\theta))^{n_2} \prod_{i=1}^{n_1} f(X_i - \theta) \ ,$$
$$L_c(\theta) = \prod_{i=1}^{n_1} f(X_i - \theta) \prod_{i=1}^{n_2} f(V_i - \theta)\mathbf{1}_{V_i > a}$$

L'expression de la densité conditionnelle est alors, d'après la formule de Bayes (6.1),

$$k(v,\theta|x) = (1 - F(a-\theta))^{-n_2} \prod_{i=1}^{n_2} f(v_i - \theta)\mathbf{1}_{V_i > a} \ .$$

On en déduit que X et V sont indépendantes et que $V_1, \ldots, V_{n_2}$ sont i.i.d. de densité

$$\forall v > a, \qquad \frac{f(v-\theta)}{1 - F(a-\theta)} \ .$$

Fixons alors $\theta_0 \in \Theta$, on a

$$Q(\theta_0, \theta) = \sum_{i=1}^{n_1} \log f(X_i - \theta) + n_2 \int_a^{+\infty} (\log f(v-\theta)) \frac{f(v-\theta_0)}{1 - F(a-\theta_0)} \mu(\mathrm{d}v) \ .$$

Pour effectuer l'étape E de l'algorithme, il suffit donc à chaque étape d'évaluer cette second intégrale en $\theta_0 = \hat{\theta}_m$.

Fixons ensuite $\theta_0 \in \Theta$. On cherche le maximum de $Q(\theta_0, \theta)$ en θ. On a

$$\frac{\partial}{\partial\theta} Q(\theta_0, \theta) = -\sum_{i=1}^{n_1} \frac{f'(X_i - \theta)}{f(X_i - \theta)} - n_2 \int_a^{+\infty} \frac{f'(v-\theta)}{f(v-\theta)} \frac{f(v-\theta_0)}{1 - F(a-\theta_0)} \mu(\mathrm{d}v) \ . \quad (6.7)$$

Pour déterminer l'étape M, on cherche alors, lorsque $\theta_0 = \hat{\theta}_m$, une solution à $\frac{\partial}{\partial\theta} Q(\hat{\theta}_m, \theta) = 0$.

Pour poursuivre ce calcul, on spécifie un modèle pour f en supposant que $f(x) = \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$ est la densité d'une loi normale centrée réduite. On a alors $f'(x) = -xf(x)$ donc, pour tout $\theta \in \Theta = \mathbb{R}$,

$$-\frac{f'(x-\theta)}{f(x-\theta)} = x - \theta$$

En insérant cette égalité dans l'équation (6.7), on voit que $\hat{\theta}_{m+1}$ est solution de

$$\sum_{i=1}^{n_1} (X_i - \theta) + n_2 \int_a^{+\infty} (v-\theta) \frac{f(v-\hat{\theta}_m)}{1 - F(a-\hat{\theta}_m)} \mu(\mathrm{d}v) = 0 \ .$$

Pour résoudre ce système, on remarque d'abord que

$$\sum_{i=1}^{n_1} (X_i - \theta) = n_1 \left(\frac{1}{n_1} \sum_{i=1}^{n_1} X_i - \theta \right) \ .$$

Ensuite, on a, pour tous réels a et b,

$$\int_b^{+\infty} (x-a) \mathrm{e}^{-\frac{1}{2}(x-a)^2} \mathrm{d}x = \mathrm{e}^{-\frac{1}{2}(b-a)^2} \ .$$

Donc

$$\int_a^{+\infty} (v-\theta) \frac{f(v-\hat{\theta}_m)}{1 - F(a-\hat{\theta}_m)} \mu(\mathrm{d}v) = (\hat{\theta}_m - \theta) + \frac{f(a-\hat{\theta}_m)}{1 - F(a-\hat{\theta}_m)}$$

Finalement,

$$\hat{\theta}_{m+1} = \frac{n_1}{n} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} X_i \right) + \frac{n_2}{n} \left(\hat{\theta}_m + \frac{f(a-\hat{\theta}_m)}{1 - F(a-\hat{\theta}_m)} \right)$$

6.4 Exemple 2

Soit Y_1 et Y_2 deux variables aléatoires Gaussiennes $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ avec $i \in \{1, 2\}$, soit B une variable aléatoire de Bernoulli $B \sim B(\epsilon)$ et soit $A = (1-B)Y_1 + BY_2$. Soit $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \epsilon)$. La loi de A a pour densité

$$f(x, \theta) = (1-\epsilon) f(x, \mu_1, \sigma_1^2) + \epsilon f(x, \mu_2, \sigma_2^2) \ ,$$

où

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}, \qquad f(x, \mu, \sigma^2) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma} \right) \ .$$

La loi de B sachant A est une loi de Bernoulli de paramètre donné grâce à la formule de Bayes par

$$\mathbb{P}(B=1|A=x) = \frac{\mathbb{P}(B=1\cap A=x)}{\mathbb{P}(A=x)} = \frac{\epsilon f(x,\mu_2,\sigma_2^2)}{(1-\epsilon)f(x,\mu_1,\sigma_1^2)+\epsilon f(x,\mu_2,\sigma_2^2)} \ .$$

Supposons qu'on observe $X_1,\dots,X_n$ i.i.d. de densité $f(x,\theta)$ et soit $V_1,\dots,V_n$ les variables non observées $B_1,\dots,B_n$.

La vraisemblance complète est donnée par

$$L_c(\theta) = \prod_{i=1}^n f(X_i,\mu_1,\sigma_1^2)^{1-V_i} f(X_i,\mu_2,\sigma_2^2)^{V_i}$$

La fonction $Q(\theta_0,\theta)$ est donnée par

$$Q(\theta_0,\theta) = \sum_{i=1}^n (1-\hat{\epsilon}_i)\log(f(X_i,\mu_1,\sigma_1^2)) + \hat{\epsilon}_i \log(f(X_i,\mu_2,\sigma_2^2)) \ ,$$

où

$$\hat{\epsilon}_i = \mathbb{E}_{\theta_0}[B_i|X] = \frac{\epsilon_0 f(X_i,\mu_{0,2},\sigma_{0,2}^2)}{(1-\epsilon_0)f(X_i,\mu_{0,1},\sigma_{0,1}^2)+\epsilon_0 f(X_i,\mu_{0,2},\sigma_{0,2}^2)}$$

On a alors, θ_0 étant fixe,

$$\frac{\partial}{\partial\mu_1}Q(\theta_0,\theta)=0 \quad \Leftrightarrow \quad \hat{\mu}_1 = \frac{\sum_{i=1}^n(1-\hat{\epsilon}_i)X_i}{\sum_{i=1}^n(1-\hat{\epsilon}_i)} \ ,$$

$$\frac{\partial}{\partial\mu_2}Q(\theta_0,\theta)=0 \quad \Leftrightarrow \quad \hat{\mu}_2 = \frac{\sum_{i=1}^n\hat{\epsilon}_i X_i}{\sum_{i=1}^n\hat{\epsilon}_i} \ ,$$

$$\frac{\partial}{\partial\sigma_1^2}Q(\theta_0,(\hat{\mu}_1,\sigma_1^2,\hat{\mu}_2,\sigma_2^2,\epsilon))=0 \quad \Leftrightarrow \quad \hat{\sigma}_1 = \frac{\sum_{i=1}^n(1-\hat{\epsilon}_i)(X_i-\hat{\mu}_1)^2}{\sum_{i=1}^n(1-\hat{\epsilon}_i)} \ ,$$

$$\frac{\partial}{\partial\sigma_2^2}Q(\theta_0,(\hat{\mu}_1,\hat{\sigma}_1^2,\hat{\mu}_2,\sigma_2^2,\epsilon))=0 \quad \Leftrightarrow \quad \hat{\sigma}_2 = \frac{\sum_{i=1}^n\hat{\epsilon}_i(X_i-\hat{\mu}_2)^2}{\sum_{i=1}^n\hat{\epsilon}_i} \ .$$

Finalement, $\hat{\epsilon}_i$ étant un estimateur de $\mathbb{P}_{\theta_0}[V_i=1|X]$, on estime $\epsilon=\mathbb{P}_{\theta_0}[V_i=1]$ par $\frac{1}{n}\sum_{i=1}^n\hat{\epsilon}_i$.

Chapitre 7

Médiane empirique et test du signe

7.1 Le modèle de translation

Soit X une variable aléatoire de fonction de répartition $F = F(\cdot, \theta)$ et de densité $f = f(\cdot, \theta)$, avec $\theta \in \Theta$. On voit ici le paramètre d'intérêt $g(\theta)$ comme une fonction de $F = F(\cdot, \theta)$, on l'écrit donc $g(\theta) = T(F)$. Ainsi, si on est intéressé par l'estimation de l'espérance μ de X, on peut écrire $\mu = T(F)$, avec

$$T(F) = \int_{-\infty}^{+\infty} x \mathrm{d}F(x) \ .$$

De même, la médiane de X est définie par $\text{mediane}(X) = F^{-1}(1/2)$. Les fonctions T ainsi introduites sont des fonctions de la fonction de répartition F (ou de la densité f), on les appelle des fonctionnelles.

Soit $Z_n = (X_1, \ldots, X_n)$ un échantillon aléatoire de la loi F et soit $T(F)$ une fonctionnelle. La fonction de répartition empirique de l'échantillon est définie par

$$\forall x \in \mathbb{R}, \qquad F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leqslant x\}} = \frac{1}{n} |\{i : X_i \leqslant x\}| \ .$$

Comme F_n est une fonction de répartition, $T(F_n)$ est également définie. Comme de plus $T(F_n)$ ne dépend que de l'échantillon, c'est une statistique qu'on appelle estimateur induit de $T(F)$.

On peut vérifier (faites le en exercice !) que, si $T(F) = \mathbb{E}[X]$, on a $T(F_n) = n^{-1} \sum_{i=1}^{n} X_i$ est la moyenne empirique et, si $T(F)$ est la médiane de X, $T(F_n)$ est la médiane empirique de l'échantillon.

Definition 83. *Soit X une variable aléatoire de densité f et de fonction de répartition F. Soit T une fonctionnelle. On dit que T est une fonctionnelle de translation si*

- *pour tout $a \in \mathbb{R}$, si $Y = X + a$, alors $T(F_Y) = T(F_X) + a$,*
- *pour tout $a \neq 0$, si $Y = aX$, alors $T(F_Y) = aT(F_X)$.*

Exercice : Vérifier que la moyenne et la médiane sont des fonctionnelles de translation. Pour que, si $\alpha \neq 1/2$, la fonctionnelle α-quantile n'est pas une fonctionnelle de translation.

Definition 84 (Modèle de translation)**.** *Les variables aléatoires $X_1, \dots, X_n$ suivent un modèle de translation s'il existe une fonctionnelle de translation T telle que*

$$X_i = \theta + \varepsilon_i, \qquad i = 1, \dots, n \ ,$$

avec $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. de densité f_ε, de fonction de répartition F_ε telles que $T(F_\varepsilon) = 0$. Dans ce cas, on dit que $X_1, \dots, X_n$ suivent le modèle de translation de fonctionelle T.

Exemple 12. *Soit ε une variable aléatoire de médiane 0. Si $\varepsilon_1, \dots, \varepsilon_n$ sont i.i.d. de même loi que ε et que, pour tout $i \in \{1, \dots, n\}$, $X_i = \theta + \varepsilon_i$, alors $X_1, \dots, X_n$ suivent un modèle de translation dont la fonctionnelle θ est la médiane de X_i.*

Théorème 85. *Supposons la loi de la variable aléatoire X symétrique par rapport à un point $\theta_0 \in \mathbb{R}$. Soit F la fonction de répartition de X. Alors, quelque soit la fonctionnelle de translation T, on a $T(F) = \theta_0$.*

Preuve. Soit T une fonctionnelle de translation et soit F_Y la fonction de répartition de la loi Y, quelque soit la variable aléatoire Y. D'après la première propriété des fonctionnelles de translation, on a $T(F_{X-\theta_0}) = T(F) - \theta_0$. De plus, l'hypothèse de symétrie signifie que $F_{X-\theta_0} = F_{\theta_0 - X}$, donc, comme T est une fonctionnelle de translation,

$$T(F) - \theta_0 = T(F_{X-\theta_0}) = T(F_{\theta_0 - X}) = -T(F_{X-\theta_0}) = \theta_0 - T(F) \ .$$

□

7.2 Test du signe

Soient $X_1, \dots, X_n$ un échantillon aléatoire satisfaisant un modèle de translation de fonctionnelle de translation la médiane, i.e.

$$X_i = \theta + \varepsilon_i, \qquad i \in \{1, \dots, n\} \ .$$

On note F la fonction de répartition de ε, f sa densité et on a que la médiane de F est nulle. On s'intéresse au test des hypothèses

$$H_0 : \theta = \theta_0 \quad \text{contre} \quad H_1 : \theta > \theta_0 \ .$$

On introduit pour cela la statistique du signe, définie par

$$S(\theta_0) = |\{i : X_i > \theta_0\}| = \sum_{i=1}^{n} \mathbf{1}_{\{X_i > \theta_0\}} \ .$$

Si H_0 est vraie, on s'attend à ce que $S(\theta_0)$ vaille environ $n/2$ alors que, si H_1 est vraie, $S(\theta_0)$ sera typiquement plus grande. Cela suggère de rejeter H_0 si $S(\theta_0) > c$, pour une valeur de seuil c à déterminer. Remarquons que, sous H_0, les variables $\mathbf{1}_{\{X_i > \theta_0\}}$ sont indépendantes et de loi de Bernoulli $B(1/2)$. Ainsi, $S(\theta_0)$ suit, sous H_0, une loi binomiale $B(n, 1/2)$. Cette loi étant indépendante de θ, on peut l'utiliser pour calibrer le seuil c.

Comme c'est une loi discrète, il est possible, selon les valeurs de α de devoir construire un test randomisé si on veut qu'il soit de taille α. Il est facile d'avoir accès aux valeurs de la fonction de répartition de la loi binomiale, par exemple, la

commande `pbinom(0:n,n,p)` en R donne les valeurs de la fonction de répartition de la loi binomiale $B(n,p)$, il est possible d'en déduire la valeur de c pour que le test soit d'un niveau α donné.

Introduisons la famille de tests $(\phi_\alpha)_{\alpha\in[0,1]}$ où, pour tout $\alpha \in [0,1]$, $\phi_\alpha = \mathbf{1}_{\{S(\theta_0)>c_\alpha\}}$, avec F la fonction de répartition de la loi $B(n,1/2)$ et

$$c_\alpha = F^{-1}(1-\alpha) = \inf\{c>0 : F(c) \geqslant 1-\alpha\} \ .$$

Comme F est continue à droite, on a $F(c_\alpha) \geqslant 1-\alpha$, et donc

$$\mathbb{P}_{\theta_0}(\phi_\alpha = 1) = \mathbb{P}_{\theta_0}(S(\theta_0) > c_\alpha) = 1 - F(c_\alpha) \leqslant \alpha \ .$$

Ainsi, pour tout $\alpha \in [0,1]$, ϕ_α est de niveau α. De plus, si $\alpha < \alpha'$, on a clairement $c_\alpha \geqslant c_{\alpha'}$, donc, si $S(\theta_0) > c_\alpha$, on a $S(\theta_0) > c_{\alpha'}$, et donc $\phi_\alpha \leqslant \phi_{\alpha'}$. Ainsi, on peut définir la p-valeur associée à cette famille de tests. On a

$$\hat{\alpha}(\phi) = \inf\{\alpha \in [0,1] : \phi_\alpha = 1\} = \inf\{\alpha \in [0,1] : S(\theta_0) > c_\alpha\} \ .$$

Supposons F continue, par construction, on a

$$\{S(\theta_0) > c_\alpha\} \quad \Longleftrightarrow \quad \{F(S(\theta_0)) > 1-\alpha\} \ .$$

Ainsi,

$$\hat{\alpha}(\phi) = \inf\{\alpha \in [0,1] : F(S(\theta_0)) > 1-\alpha\} = 1 - F(S(\theta_0)) \ .$$

Exercice : Etablir de même un test de niveau α basé sur la statistique $S(\theta_0)$ pour les hypothèses $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Calculer la p-valeur associée à la famille de tests ainsi construite.

On s'intéresse maintenant à la puissance du test du signe. Quitte à soustraire θ_0 à toutes les données, on peut supposer sans perte de généralité que $\theta_0 = 0$. On définit alors, pour tout $\theta \in \mathbb{R}$, la statistique

$$S(\theta) = \sum_{i=1}^{n} \mathbf{1}_{\{X_i>\theta\}} = |\{i : X_i > \theta\}| \ .$$

On a clairement que $S(\theta)$ est une fonction décroissante, qui prend des valeurs entières entre 0 et n et qui change de valeur en chaque statistique d'ordre $X_{i:n}$. On a également le résultat suivant.

Proposition 86. *Pour tout $\theta \in \mathbb{R}$ et tout $k \in \{1,\dots,n\}$,*

$$\mathbb{P}_\theta(S(0) > k) = \mathbb{P}_0(S(-\theta) > k) \ .$$

Preuve. Soit $X \sim \mathbb{P}_\theta$, on a $X = \theta + \varepsilon$, avec $\varepsilon \sim \mathbb{P}_0$. Soient $X_1,\dots,X_n$ i.i.d. de même loi que X et $\varepsilon_1,\dots,\varepsilon_n$ tels que $X_i = \theta+\varepsilon_i$, de sorte que $\varepsilon_1,\dots,\varepsilon_n$ sont i.i.d. de même loi que ε. La probabilité de droite est la probabilité de l'événement

$$\left\{ \sum_{i=1}^{n} \mathbf{1}_{\{X_i>0\}} > k \right\} \ .$$

Celle de gauche est celle de l'événement

$$\left\{ \sum_{i=1}^{n} \mathbf{1}_{\{\varepsilon_i>-\theta\}} > k \right\} \ .$$

Le résultat vient du fait que, par construction, $\{\varepsilon_i > -\theta\} = \{X_i > 0\}$, donc

$$\sum_{i=1}^{n} \mathbf{1}_{\{\varepsilon_i > -\theta\}} = \sum_{i=1}^{n} \mathbf{1}_{\{X_i > 0\}} \ .$$

□

Proposition 87. *La fonction puissance du test du signe est une fonction décroissante de θ.*

Preuve : La fonction puissance du test du signe vaut

$$\beta_\phi(\theta) = \mathbb{P}_\theta(\phi = 1) = \mathbb{P}_\theta(S(0) > c) \ .$$

D'après la proposition 86, on a donc

$$\beta_\phi(\theta) = \mathbb{P}_0(S(-\theta) > c) \ .$$

Enfin, pour toute valeur de $\varepsilon_1, \ldots, \varepsilon_n$, S est une fonction décroissante, donc, si $\theta > \theta'$, $S(-\theta) \geqslant S(-\theta')$, donc

$$\beta_\phi(\theta) = \mathbb{P}_0(S(-\theta) > c) \geqslant \mathbb{P}_0(S(-\theta') > c) = \beta_\phi(\theta') \ .$$

□

Remarquons que cette propriété est encore vraie si on considère le test du signe comme test des hypothèses $H_0 : \theta \leqslant \theta_0$ contre $H_1 : \theta > \theta_0$. D'autre part, pour calibrer ce nouveau test, on peut utiliser la même valeur de c puisque par définition, la taille du test vaut $\sup_{\theta \leqslant \theta_0} \beta_\phi(\theta)$ ce qui, par la propriété de croissance de la fonction β_ϕ, vaut simplement $\beta_\phi(\theta_0)$.

Sous l'alternative $\theta = \theta_1$, la statistique de test $S(0)$ suit aussi une loi binomiale $B(n, p_1)$, où

$$p_1 = \mathbb{P}_{\theta_1}(X > 0) = 1 - F(-\theta_1) \ .$$

Pour comparer cette puissance à celle d'autres tests, on introduit dans la section suivante un outil classique appelé efficacité asymptotique relative.

7.3 Efficacité asymptotique relative

Nous savons qu'il est commode d'utiliser le comportement asymptotique des statistiques pour comparer les procédures. Cette méthode est utile dans le problème d'estimation où on est capable de définir une borne inférieure sur la variance des M-estimateurs pour les modèles réguliers. Dans cette section, nous développons une approche due à Pitman permettant de comparer les procédures de tests à partir du comportement asymptotique de statistiques permettant de les définir. Rappelons qu'on s'intéresse de manière générale dans ce chapitre au problème du test des hypothèses

$$H_0 : \theta = 0, \quad \text{contre} \quad H_1 : \theta > 0 \ .$$

Le théorème de la limite centrale assure que

$$\sqrt{n}(n^{-1}S(0) - 1/2) \xrightarrow{\mathbb{P}_0} \mathrm{N}(0, 1/4) \ . \tag{7.1}$$

Il est donc aisé de construire un test de taille asymptotique α, il suffit de poser $\phi(Z_n) = \mathbf{1}_{\{2\sqrt{n}(n^{-1}S(0)-1/2)>z_\alpha\}}$, avec z_α, le $1-\alpha$ quantile de la loi $\mathrm{N}(0,1)$. Ce faisant, une idée naturelle est de calculer la puissance asymptotique en un point $\theta > 0$. On a, toujours par le théorème de la limite centrale,

$$2\sqrt{n}(n^{-1}S(0) - (1 - F(-\theta))) \overset{\mathbb{P}_\theta}{\Longrightarrow} \mathrm{N}(0,1) \ . \tag{7.2}$$

Or, la puissance en θ vaut

$$\begin{aligned}\beta_\phi(\theta) &= \mathbb{P}_\theta(2\sqrt{n}(n^{-1}S(0) - 1/2) > z_\alpha) \\ &= \mathbb{P}_\theta(2\sqrt{n}(n^{-1}S(0) - (1 - F(-\theta))) > z_\alpha - 2\sqrt{n}(1/2 - F(-\theta))) \ .\end{aligned}$$

Comme $\theta > 0$, $1/2 - F(-\theta) > 0$, donc, pour tout $M > 0$, il existe n_0 tel que, pour tout $n \geqslant n_0$, $z_\alpha - 2\sqrt{n}(1/2 - F(-\theta)) < -M$. On en déduit que, pour tout $n \geqslant n_0$,

$$\beta_\phi(\theta) \geqslant \mathbb{P}_\theta(2\sqrt{n}(n^{-1}S(0) - (1 - F(-\theta))) > -M) \to \Phi(M) \ ,$$

où Φ désigne la fonction de répartition de la Gaussienne standard. Ceci étant vrai pour tout M, on en déduit que $\lim_n \beta_\phi(\theta) = 1$. Il est clair que le même raisonnement aurait pu être mené avec tout test basé sur une statistique asymptotiquement normale. Ainsi, tous les tests ainsi construits sont consistants, c'est à dire que leur puissance asymptotique vaut 1 en tout point $\theta > 0$ de l'alternative. Pour pouvoir comparer les tests en utilisant le comportement asymptotique des statistiques, il est donc nécessaire d'introduire une autre idée.

L'idée de Pitman est de regarder la puissance asymptotique des tests lorsque l'alternative a la forme suivante : pour tout $h > 0$:

$$H_{1,n} : \theta = \frac{h}{\sqrt{n}} \ .$$

Les alternatives $H_{1,n}$ sont appelées alternatives locales car elles convergent vers H_0 lorsque $n \to \infty$.

Notons $\theta_n = h/\sqrt{n}$. Le comportement asymptotique des statistiques sous $\mathbb{P}_{\theta_n}$ est plus délicat et ne peut être analysé à partir du théorème de la limite centrale puisque, pour tout n, la loi de $X_i = \theta_n + \epsilon_i$ change. Une possibilité est d'utiliser le théorème de Lindeberg-Feller.

Théorème 88 (Lindeberg-Feller)**.** *Soit k_n une suite croissante d'entiers. Soit $(Y_{n,i})_{i=1,\dots,k_n}$ un tableau de vecteurs aléatoires définis sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, indépendants, centrés et tels que $\mathbb{E}[\|Y_{n,i}\|^2] < \infty$ pour $i = 1, \dots, k_n$. Supposons*

- *pour tout $\epsilon > 0$, $\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}[\|Y_{n,i}\|^2 \mathbf{1}_{\|Y_{n,i}\|>\epsilon}] = 0$,*
- *$\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathrm{Var}(Y_{n,i}) = \Sigma$.*

Alors, la suite $\sum_{i=1}^{k_n} Y_{n,i} \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0, \Sigma)$.

Nous ne démontrons pas ce résultat ici. On va se contenter de l'appliquer au problème du test du signe. On pose

$$Y_{n,i} = \frac{2(\mathbf{1}_{\{\epsilon_i > -\theta_n\}} - (1 - F(-\theta_n)))}{\sqrt{n}} \ .$$

Les variables $\mathbf{1}_{\{\epsilon_i > -\theta_n\}}$ suivent une loi de Bernoulli $B(p_n)$ avec $p_n = 1 - F(-\theta_n)$, donc les $Y_{n,i}$ sont bien centrées, indépendantes, et on a $\mathrm{Var}(Y_{n,i}) = 4p_n(1-p_n)/n$. Comme $\theta_n \to 0$, on a $p_n \to 1/2$ donc

$$\sum_{i=1}^{n} \mathrm{Var}(Y_{n,i}) \to 1 \ .$$

De plus, on a

$$\mathbb{E}[|Y_{n,i}|^2 \mathbf{1}_{|Y_{n,i}|>\epsilon}] \leqslant \frac{\mathbb{E}[|Y_{n,i}|^3]}{\epsilon} \ ,$$

et

$$\mathbb{E}[|Y_{n,i}|^3] = \frac{8}{n^{3/2}}\big(p_n(1-p_n)^3 + (1-p_n)p_n^3\big) \ ,$$

donc

$$\sum_{i=1}^{n} \mathbb{E}[|Y_{n,i}|^2 \mathbf{1}_{|Y_{n,i}|>\epsilon}] = \frac{8\big(p_n(1-p_n)^3 + (1-p_n)p_n^3\big)}{\sqrt{n}} \to 0 \ ,$$

car

$$8\big(p_n(1-p_n)^3 + (1-p_n)p_n^3\big) \to 1 \ .$$

Le théorème de Lindeberg-Feller assure alors que

$$\sum_{i=1}^{n} Y_{n,i} \stackrel{\mathbb{P}_0}{\Longrightarrow} \mathrm{N}(0,1) \ . \tag{7.3}$$

Si $X_i = \theta_n + \epsilon_i$, on a

$$2\sqrt{n}(n^{-1}S(0) - (1 - F(-\theta_n))) = \sum_{i=1}^{n} Y_{n,i} \ .$$

On en déduit

$$2\sqrt{n}(n^{-1}S(0) - (1 - F(-\theta_n))) \stackrel{\mathbb{P}_{\theta_n}}{\Longrightarrow} \mathrm{N}(0,1) \ .$$

Ainsi

$$2\sqrt{n}(n^{-1}S(0) - 1/2) = 2\sqrt{n}(n^{-1}S(0) - (1 - F(-\theta_n))) + \sqrt{n}(1 - 2F(-\theta_n)) \ .$$

On utilise alors le développement de Taylor de F en 0 pour obtenir

$$F(-\theta_n) = F(0) - \theta_n f(0) + o(\theta_n) = 1/2 - \frac{hf(0)}{\sqrt{n}} + o(n^{-1/2}) \ .$$

Ainsi

$$\lim_{n\to\infty} \sqrt{n}(1 - 2F(-\theta_n)) = 2hf(0) \ .$$

Finalement, par le lemme de Slutsky, on a donc

$$2\sqrt{n}(n^{-1}S(0) - 1/2) \stackrel{\mathbb{P}_{\theta_n}}{\Longrightarrow} \mathrm{N}(2hf(0), 1) \ . \tag{7.4}$$

Ce résultat est à la base du théorème suivant.

Théorème 89 (Lemme de puissance asymptotique)**.** *Le test ϕ du signe*

$$\phi(Z_n) = \mathbf{1}_{\{2(S(0)-n/2)/\sqrt{n}>z_\alpha\}}$$

est de niveau asymptotique α et sa fonction puissance vérifie

$$\forall h > 0, \qquad \lim_{n\to\infty} \beta_\phi(h/\sqrt{n}) = 1 - \Phi(z_\alpha - 2hf(0)) \ ,$$

où Φ est la fonction de répartition de la Gaussienne standard $\mathrm{N}(0,1)$ et $z_\alpha = \Phi^{-1}(1-\alpha)$ est le $(1-\alpha)$-quantile de cette loi.

Preuve. D'après l'équation (7.2), sous H_0, on a

$$\mathbb{P}_0(\phi(Z_n) = 1) = \mathbb{P}_0(2\sqrt{n}(n^{-1}S(0) - 1/2) > z_\alpha) \to \alpha \ .$$

Soit maintenant $h > 0$ et $\theta_n = h/\sqrt{n}$, on a

$$\begin{aligned}\beta_\phi(h/\sqrt{n}) &= \mathbb{P}_{\theta_n}(2\sqrt{n}(n^{-1}S(0) - 1/2) > z_\alpha) \\ &= \mathbb{P}_{\theta_n}(2\sqrt{n}(n^{-1}S(0) - 1/2) - 2hf(0) > z_\alpha - 2hf(0)) \ .\end{aligned}$$

Donc, d'après l'équation (7.4),

$$\beta_\phi(h/\sqrt{n}) \to 1 - \Phi(z_\alpha - 2hf(0)) \ .$$

□

La quantité $2f(0)$ est appelée la *pente* du test du signe. C'est une quantité non triviale qui peut être utilisée pour comparer le test du signe à d'autres tests. La méthodologie utilisée pour calculer la pente du test du signe peut être utilisée pour d'autres tests. Nous ne présentons pas la théorie générale ici, mais donnons un second exemple, toujours sur le modèle de translation $\{f(\cdot - \theta), \theta \in \Theta\}$. Le but est de se familiariser avec les outils probabilistes que nous avons introduits pour calculer la pente du test du signe. La théorie générale requiert des outils développés dans des cours de M2 de statistiques asymptotiques.

Supposons que f est symétrique, alors toute fonctionnelle de translation peut être utilisée pour estimer θ. En particulier, si $\int |x| f(x)\mu(\mathrm{d}x) < \infty$, on a $\mathbb{E}_\theta[X] = \theta$ pour tout $\theta \in \mathbb{R}$, et si $\sigma_f^2 = \int x^2 f(x)\mu(\mathrm{d}x) < \infty$, on a de plus, par le théorème de la limite centrale, pour tout θ,

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \overset{\mathbb{P}_\theta}{\Longrightarrow} \mathrm{N}(0, \sigma_f^2), \qquad \hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i \ .$$

Ce résultat montre que le test ϕ_2 suivant est de niveau α :

$$\phi_2(Z_n) = \mathbf{1}\big\{\sqrt{n}\hat{\theta}_n > \sigma_f z_\alpha\big\} \ .$$

Pour évaluer la pente de ce test, il s'agit maintenant d'étudier son comportement sous l'alternative locale $\theta_n = h/\sqrt{n}$. Les variables aléatoires $\epsilon_i = X_i - \theta_n$ sont i.i.d. de variance σ_f, donc d'après le théorème de la limite centrale centrale,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i \overset{\mathbb{P}_0}{\Longrightarrow} \mathrm{N}(0, \sigma_f^2) \ .$$

Si $X_i = \theta_n + \epsilon_i$,

$$\sqrt{n}(\hat{\theta}_n - \theta_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \ .$$

Donc

$$\sqrt{n}\frac{\hat{\theta}_n - \theta_n}{\sigma_f} \overset{\mathbb{P}_{\theta_n}}{\Longrightarrow} \mathrm{N}(0,1) \ .$$

Ainsi, comme $\sqrt{n}\theta_n/\sigma_f \to h/\sigma_f$, on a, par le lemme de Slutsky

$$\begin{aligned}\beta_{\phi_2}(\theta_n) &= \mathbb{P}_{\theta_n}\left(\sqrt{n}\frac{\hat{\theta}_n}{\sigma_f} > z_\alpha\right) \\ &= \mathbb{P}_{\theta_n}\left(\sqrt{n}\frac{\hat{\theta}_n - \theta_n}{\sigma_f} > z_\alpha - \sqrt{n}\frac{\theta_n}{\sigma_f}\right) \to 1 - \Phi(z_\alpha - h/\sigma_f) \ .\end{aligned}$$

La pente du test ϕ_2 est donc σ_f^{-1}. On peut comparer le test ϕ_2 avec le test du signe ϕ grâce à leurs pentes respectives. Si la pente de ϕ est supérieure à celle de ϕ_2, i.e., si $2f(0) > \sigma_f^{-1}$, alors $1 - \Phi(z_\alpha - 2f(0)h) > 1 - \Phi(z_\alpha - \sigma_f^{-1}h)$, donc la puissance (asymptotique) de ϕ est supérieure à celle de ϕ_2 et ϕ est donc préférable à ϕ_2.

Exemple 13. *Supposons que f soit la densité d'une Gaussienne standard, alors $2f(0) = \sqrt{2/\pi}$ et $\sigma_f = 1$, donc $2f(0) < \sigma_f^{-1}$, le test de la moyenne est plus puissant que le test du signe.*

Exemple 14. *Supposons que f soit la densité de la loi de Laplace $f(x) = e^{-|x|}/2$, alors, $2f(0) = 1$, et $\sigma_f^2 = \int_0^{+\infty} x^2 e^{-x} \mathrm{d}x = 2$, donc $\sigma_f^{-1} = 1/\sqrt{2}$. Ainsi, $\sigma_f^{-1} < 2f(0)$, donc le test du signe est plus puissant que celui de la moyenne.*

Exemple 15. *Supposons que*

$$f(x) = \frac{1-\epsilon}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} + \frac{\epsilon}{\sqrt{2\pi\sigma_c^2}} \mathrm{e}^{-\frac{x^2}{2\sigma_c^2}} \ .$$

L'idée de ce modèle est la suivante. L'essentiel des données est issu d'un modèle Gaussien standard mais une petite proportion ϵ de ces données a été "contaminée" et sont distribués selon une loi Gaussienne de variance $\sigma_c^2 > 1$ potentiellement très grande. Dans ce cadre, on a

$$2f(0) = \sqrt{\frac{2}{\pi}}\left(1 - \epsilon + \frac{\epsilon}{\sigma_c}\right), \quad \frac{1}{\sigma_f} = \frac{1}{\sqrt{1 - \epsilon + \epsilon\sigma_c^2}} \ .$$

Remarquons que, lorsque $\sigma_c = 1$, on retrouve le résultat du premier exemple et le test de la moyenne est préférable à celui du signe. A l'inverse, lorsque $\sigma_c \to \infty$, on a $2f(0) \sim \sqrt{\frac{2}{\pi}}(1-\epsilon)$ et $\sigma_f^{-1} \sim \sigma_c^{-1}\epsilon^{-1/2}$, la pente du test du signe reste supérieure à une constante absolue strictement positive tandis que celle du test de la moyenne tend vers 0. Ainsi, si la contamination est importante $\sigma_f \gg 0$, même si la proportion de données contaminée est petite, le test du signe devient préférable à celui de la moyenne. En pratique, si on suspecte que le jeu de données a pu être corrompu, on préfère le test du signe basé sur la médiane qui est un estimateur plus robuste que la moyenne.

Chapitre 8

Test du rang de Wilcoxon

8.1 Présentation et premiers résultats

Soient $X_1 \dots, X_n$ des variables i.i.d. tirées selon un modèle de translation, i.e. de densité $f(\cdot - \theta)$ avec $\theta \in \mathbb{R}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$. On suppose en outre dans ce chapitre que f est symétrique par rapport à l'origine, de sorte que, d'après le théorème 85, c'est un modèle de translation par rapport à n'importe quelle fonctionnelle de translation. On s'intéresse ici au test des hypothèses

$$H_0 : \theta = 0, \quad \text{contre} \quad H_1 : \theta > 0 \ .$$

On définit, pour tout réel x, le signe de x par $\mathrm{sgn}(x) = \mathbf{1}_{\{x \geqslant 0\}} - \mathbf{1}_{\{x<0\}}$. De plus, étant donné un vecteur $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$, on définit le rang de la coordonnée i par $\mathrm{rg}(x_i) = k$ si $x_i = x_{k:n}$. La statistique du test du rang de Wilcoxon est définie par

$$T = \sum_{i=1}^{n} \mathrm{sgn}(X_i)\, \mathrm{rg}(|X_i|) \ .$$

Sous H_0, la moitié des X_i sera positive (par symétrie de f) et les rangs sont distribués uniformément sur $\{1, \dots, n\}$ par échangeabilité de la loi du vecteur $(X_1, \dots, X_n)$. En revanche, sous H_1, les X_i seront majoritairement positives et les plus grandes valeurs de $|X_i|$ seront positives. Ceci suggère de rejeter H_0 lorsque $T > c$, où c est un seuil qu'il s'agit de déterminer. La loi de T est à support dans l'ensemble

$$\left\{ -\frac{n(n-1)}{2}, -\frac{n(n-1)}{2} + 2, \dots, \frac{n(n-1)}{2} \right\} \ .$$

On a, $\mathrm{sgn}(X_i)$ sont des variables i.i.d. à valeurs dans $\{-1, 1\}$ et, par symétrie de f,

$$\mathbb{P}_0(\mathrm{sgn}(X) = 1) = \mathbb{P}_0(X \geqslant 0) = \int_0^{+\infty} f(x)\mathrm{d}x = \frac{1}{2} \ .$$

Ainsi $\mathrm{sgn}(X_1), \dots, \mathrm{sgn}(X_n)$ sont i.i.d. de loi de Rademacher. De plus, on a le lemme suivant.

Lemme 90. *Si $\epsilon_1, \dots, \epsilon_n$ sont i.i.d. de densité f symétrique par rapport à 0, alors $\mathrm{sgn}(\epsilon_1), \dots, \mathrm{sgn}(\epsilon_n)$ sont indépendantes de $|\epsilon_1|, \dots, |\epsilon_n|$.*

Démonstration. Il suffit de montrer que $\mathrm{sgn}(\epsilon)$ est indépendante de $|\epsilon|$ et pour cela, que, pour tout $x > 0$,

$$\mathbb{P}(\mathrm{sgn}(\epsilon) = 1, |\epsilon| \leqslant x) = \mathbb{P}(\mathrm{sgn}(\epsilon) = 1)\mathbb{P}(|\epsilon| \leqslant x) \ .$$

Or, comme f est symétrique,

$$\mathbb{P}(|\epsilon| \leqslant x) = \int_{-x}^{x} f(t)\mathrm{d}t = 2\int_{0}^{x} f(t)\mathrm{d}t = 2\mathbb{P}(0 \leqslant \epsilon \leqslant x) = 2\mathbb{P}(\mathrm{sgn}(\epsilon) = 1, |\epsilon| \leqslant x) \ .$$

Ceci conclut la preuve car $\mathbb{P}(\mathrm{sgn}(\epsilon) = 1) = 1/2$. □

Notons maintenant $i_1, \ldots, i_n$ les indices tels que $|\epsilon_{i_j}|$ est le j-ème plus grand éléments des $|\epsilon_i|$. On a, si $X_1, \ldots, X_n$ sont i.i.d. $\mathbb{P}_0$,

$$T = \sum_{j=1}^{n} j\,\mathrm{sgn}(\epsilon_{i_j}) \ .$$

D'après le lemme 90, les variables aléatoires i_j, qui sont mesurables par rapport à $\sigma(|\epsilon_1|, \ldots, |\epsilon_n|)$ sont indépendantes de $\mathrm{sgn}(\epsilon_1), \ldots, \mathrm{sgn}(\epsilon_n)$, donc les variables aléatoires $\mathrm{sgn}(\epsilon_{i_1}), \ldots, \mathrm{sgn}(\epsilon_{i_1})$, conditionnellement à $i_1, \ldots, i_n$ sont des variables aléatoires i.i.d. de Rademacher. On peut déduire de ce résultat le théorème suivant.

Théorème 91. *Sous l'hypothèse H_0, la loi de T est la même que celle de*

$$T' = \sum_{j=1}^{n} j\nu_j \ ,$$

où $\nu_1, \ldots, \nu_n$ sont i.i.d. de loi de Rademacher. En particulier,

1. $\mathbb{E}_0[T] = 0$, $\mathrm{Var}_0(T) = \sum_{j=1}^{n} j^2 = n(n+1)(2n+1)/6$.
2. $T/\sqrt{\mathrm{Var}_0(T)} \overset{\mathbb{P}_0}{\Longrightarrow} \mathrm{N}(0,1)$.

Preuve. La première affirmation vient d'être démontrée et le point 1 en est une conséquence immédiate. Pour le point 2, on introduit

$$Y_{n,j} = \frac{j\nu_j}{\sqrt{\mathrm{Var}_0(T)}} \ .$$

On a clairement $\mathrm{Var}(\sum_{i=1}^{n} Y_{n,i}) = 1$. De plus, on a

$$\sum_{j=1}^{n} j^3 = O(n^4), \qquad \mathrm{Var}_0(T) = \Omega(n^3) \ ,$$

donc $\sum_{j=1}^{n} \mathbb{E}_0[|Y_{n,j}|^3] = O(n^{-1/2})$ et, par l'inégalité de Markov, pour tout $\epsilon > 0$,

$$\sum_{j=1}^{n} \mathbb{E}_0[|Y_{n,j}|^2 \mathbf{1}_{|Y_{n,j}|>\epsilon}] \leqslant \frac{\sum_{j=1}^{n} \mathbb{E}_0[|Y_{n,j}|^3]}{\epsilon} \to 0 \ .$$

Le point 2 est donc une conséquence du théorème de Lindeberg-Feller. □

Il est parfois utile d'utiliser une autre formulation de la statistique du test du rang. Introduisons T^+, la somme des rangs des $X_i > 0$, i.e.

$$T^+ = \sum_{i:X_i>0} \mathrm{rg}(|X_i|) \ .$$

Comme $\sum_{i=1}^n \mathrm{rg}(|X_i|) = n(n+1)/2$, on a

$$T = \sum_{i=1}^n \mathrm{sgn}(X_i)\,\mathrm{rg}(|X_i|) = T^+ - \sum_{i:X_i<0} \mathrm{rg}(|X_i|) = 2T^+ - \frac{n(n+1)}{2} \ .$$

On déduit de cette formule et du théorème 91 que

$$\mathbb{E}_0[T^+] = \frac{1}{2}\left(\mathbb{E}_0[T] + \frac{n(n+1)}{2}\right) = \frac{n(n+1)}{4}, \qquad \mathrm{Var}_0(T^+) = \frac{1}{4}\mathrm{Var}_0(T) = \frac{n(n+1)(2n+1)}{24} \ .$$

Pour obtenir $\mathbb{P}_0(T^+ \leqslant t)$ pour un échantillon de taille n sur R, on utilise la commande `psignrank(t,n)`.

Soit $X_i > 0$ et soit j tel que $-X_i < X_j \leqslant X_i$. Le nombre d'entiers j vérifiant cette contrainte est $\mathrm{rg}(|X_i|)$ et, pour chacun d'eux, $(X_i + X_j)/2 > 0$. On en déduit

$$\begin{aligned} T^+ &= \sum_{i:X_i>0} \mathrm{rg}(|X_i|) = \sum_{i:X_i>0} \sum_{j=1}^n \mathbf{1}_{\{-X_i<X_j\leqslant X_i\}} \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbf{1}_{\{X_i+X_j>0\}} \mathbf{1}_{\{X_i\leqslant X_j\}} = \sum_{1\leqslant i\leqslant j\leqslant n} \mathbf{1}_{\{X_i+X_j>0\}} \ . \end{aligned}$$

Les moyennes $(X_i + X_j)/2$ sont appelées moyennes de Walsh. On a montré que T^+ était égal au nombre de moyennes de Walsh strictement positives.

8.2 Pente

Supposons $\theta_0 = 0$, soit $h > 0$ et $\theta_n = h/\sqrt{n}$. On étudie l'asymptotique du test du rang pour les hypothèses

$$H_0 : \theta = 0, \quad \text{contre} \quad H_1 : \theta = \theta_n \ .$$

Le second point du théorème 91 donne le comportement asymptotique de T sous H_0, ce qui permet de calibrer le test de façon à avoir un niveau asymptotique α. On pose $c_\alpha = \sqrt{\mathrm{Var}_0(T)}\, z_\alpha$, où $z_\alpha = \Phi^{-1}(1-\alpha)$ est le quantile d'ordre $1-\alpha$ de la loi $\mathrm{N}(0,1)$. Soit $\phi(Z_n) = \mathbf{1}_{\{T>c_\alpha\}}$. On a

$$\mathbb{P}_0(T > c_\alpha) = \mathbb{P}_0\left(\frac{T}{\sqrt{\mathrm{Var}_0(T)}} > z_\alpha\right) \to \alpha \ .$$

Pour évaluer la pente du test du signe, on utilise typiquement un argument de contigüité qui permet d'obtenir le résultat suivant.

Théorème 92. *La puissance du test du rang de Wilcoxon de niveau asymptotique α vérifie*

$$\beta_\phi(\theta_n) \to 1 - \Phi(z_\alpha - h\tau_W), \qquad \tau_W = \sqrt{12}\sigma_f^2 \ .$$

Rappelons la discussion de la fin du chapitre précédent. On a que le test du rang ϕ est préféré à un autre test ϕ' si sa pente τ_W est supérieure à celle de ϕ'. Dans ce chapitre, on avait calculé la pente du test du signe $2f(0)$ et celle du test de la moyenne σ_f^{-1}. On peut utiliser ces résultats pour comparer ces tests au test du rang.

Exemple 16. *Supposons f est une Gaussienne standard, alors $\sigma_f = 1$, donc le test de la moyenne a pour pente 1 et celui du rang $\sqrt{12}$, donc le test du rang est préférable à celui de la moyenne, qui était préférable à celui du signe.*

Exemple 17. *Supposons que $f(x) = e^{-|x|}/2$ est la densité de la loi de Laplace, de sorte que $\sigma_f^2 = 2$, et donc que la pente $\tau_W = 2\sqrt{12} > 1 = 2f(0)$. Le test du rang est de puissance supérieure à celui du signe, qui était supérieur à celui de la moyenne.*

Exemple 18. *Supposons qu'on se place dans le modèle Gaussien contaminé*

$$f(x) = \frac{1-\epsilon}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} + \frac{\epsilon}{\sqrt{2\pi\sigma_c^2}} \mathrm{e}^{-\frac{x^2}{2\sigma_c^2}} \quad .$$

On avait vu que l'idée du modèle était que ϵ était petit et σ_c grand. On avait de plus

$$\sigma_f^2 = 1 - \epsilon + \epsilon\sigma_c^2 \quad .$$

Ainsi,

$$\tau_W = \sqrt{12}(1 - \epsilon + \epsilon\sigma_c^2), \quad 2f(0) = \sqrt{\frac{2}{\pi}}\left(1 - \epsilon + \frac{\epsilon}{\sigma_c}\right), \quad \frac{1}{\sigma_f} = \frac{1}{\sqrt{1 - \epsilon + \epsilon\sigma_c^2}} \quad .$$

La conclusion du chapitre précédent était que le test de la moyenne était préférable dans le cas ou $\epsilon \to 0$ et σ_c fixe, mais, lorsque la proportion ϵ de contamination était fixe et $\sigma_c \to \infty$, le test du signe était plus "robuste". Dans ces deux asymptotiques, on voit ici que le test du rang est supérieur.

Chapitre 9

Modèles linéaires

Dans ce chapitre, on suppose que les observations sont des couples $Z_n = ((X_1, Y_1), \dots, (X_n, Y_n))$, où les X_i sont appelées variables d'entrée ou covariables et Y_i sont appelées variables de sorties, réponses ou étiquettes selon les applications. On suppose toujours les données mutuellement indépendantes, mais pas que X_i et Y_i sont indépendantes.

L'idée générale est que les variables X_i sont faciles à mesurer et permettent de mieux prédire la variable Y_i associée. Typiquement X_i est un vecteur et Y_i appartient à un sous-ensemble $\mathcal{Y}$ de $\mathbb{R}$.

On supposera toujours que $X_i \in \mathbb{R}^d$ et on considérera les cas où $\mathcal{Y} = \mathbb{R}$ et $\mathcal{Y} = \{-1, 1\}$. Dans le premier cas, on parle de problème de régression, dans le second du problème de classification.

9.1 Regression linéaire

On suppose ici que $\mathcal{Y} = \mathbb{R}$ et que, pour tout $i \in \{1, \dots, n\}$,

$$Y_i = X_i^T \mathbf{b} + \varepsilon_i \ ,$$

où $\mathbf{b} \in \mathbb{R}^d$ est inconnu et $\varepsilon_1, \dots, \varepsilon_n$ sont des variables aléatoires indépendantes et indépendantes de $X_1, \dots, X_n$, de même loi de cdf F et de densité f. On introduit le vecteur $\mathbf{Y} \in \mathbb{R}^n$ de coordonnées $Y_1, \dots, X_n$, la matrice $\mathbf{X} \in \mathbb{R}^{n \times d}$ dont les lignes sont les vecteurs $X_1^T, \dots, X_n^T$ et $\mathbf{e} \in \mathbb{R}^n$ le vecteur de coordonnées $\varepsilon_1, \dots, \varepsilon_n$.

Les équations définissant le modèle linéaire sont alors équivalentes à la relation matricielle.

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e} \ .$$

9.1.1 Estimation par les moindres carrés

Les estimateurs des moindres carrés de $\mathbf{b}$ sont les éléments

$$\widehat{\mathbf{b}} \in \operatorname{argmin}_{\mathbf{a} \in \mathbb{R}^d} \|\mathbf{Y} - \mathbf{X}\mathbf{a}\|^2 \ .$$

Clairement, les estimateurs des moindres vérifient que

$$\mathbf{X}\widehat{\mathbf{b}} = \Pi_{\mathbf{X}} \mathbf{Y} \ ,$$

où $\Pi_{\mathbf{X}}$ est la matrice de la projection orthogonale sur l'espace engendré par les colonnes de $\mathbf{X}$. Le résultat suivant est un résultat classique d'algèbre linéaire, que vous pourrez démontrer à titre d'exercice.

Proposition 93. *On a*

1. $\Pi_{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{X}^g$, *où* $\boldsymbol{X}^g$ *désigne n'importe quelle inverse généralisée de* $\boldsymbol{X}$. *Si le rang* r *de* $\boldsymbol{X}$ *vaut* $r = d$, *alors il n'y a qu'une inverse généralisée de* $\boldsymbol{X}$.
2. *Soit* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, *avec* $\boldsymbol{U} \in \mathcal{O}_n(\mathbb{R})$, $\boldsymbol{V} \in \mathcal{O}_d(\mathbb{R})$ *et* $\boldsymbol{D}$ *est une matrice de type* (n, d) *dont les seuls coefficients non nuls sont les coefficients diagonaux qui sont les racines carrées des valeurs propres* $\lambda_1, \dots, \lambda_r, 0, \dots, 0$ *de la matrice symétrique réelle* $\boldsymbol{X}^T\boldsymbol{X}$. *La décomposition* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ *est une décomposition de* $\boldsymbol{X}$ *en valeur singulière. Alors,* $\boldsymbol{X}^{MP} = \boldsymbol{V}\boldsymbol{D}^g\boldsymbol{U}^T$, *où* $\boldsymbol{D}^g$ *est la matrice de type* (d, n), *diagonale, de coefficient diagonaux* $\lambda_1^{-1}, \dots, \lambda_r^{-1}, 0, \dots, 0$ *est une inverse généralisée de* $\boldsymbol{X}$ *appelée inverse généralisée de Moore-Penrose.*
3. $\widehat{\boldsymbol{b}}^* = \boldsymbol{X}^{MP}\boldsymbol{Y}$ *est solution du problème des moindres carrés. L'ensemble des solutions de ce problème est l'ensemble*

$$\{\widehat{\boldsymbol{b}} = \widehat{\boldsymbol{b}}^* + \boldsymbol{x},\ \boldsymbol{x} \in \ker(\boldsymbol{X})\} \ .$$

Parmi ces solutions, $\widehat{\boldsymbol{b}}^*$ *est celle de norme Euclidienne minimale.*

Les estimateurs des moindres carrés vérifient toujours le résultat suivant.

Proposition 94. *Si* $\mathbb{E}[\varepsilon] = 0$, *alors tout estimateur des moindres carrés* $\widehat{\boldsymbol{b}}$ *vérifie* $\boldsymbol{X}\mathbb{E}[\widehat{\boldsymbol{b}}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{b}$.

Si de plus $r = d$, *alors l'estimateur des moindres carrés* $\widehat{\boldsymbol{b}}^*$ *est un estimateur sans biais de* $\boldsymbol{b}$.

Si de plus, $\mathrm{Var}(\varepsilon) = \sigma^2$, *on a*

$$\mathbb{E}[(\widehat{\boldsymbol{b}}^* - \boldsymbol{b})(\widehat{\boldsymbol{b}}^* - \boldsymbol{b})^T|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \ .$$

9.1.2 Régression linéaire Gaussienne

Dans cette section, on suppose en outre que $r < n$, $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ et quitte à conditionner tous les résultats par $\mathbf{X}$, on suppose la matrice $\mathbf{X}$ déterministe. On a alors $\mathbf{Y} = \mathrm{N}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}_d)$,

$$\mathbf{X}\widehat{\mathbf{b}}^* = \Pi_{\mathbf{X}}\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\mathbf{b}, \sigma^2\Pi_{\mathbf{X}}), \qquad \widehat{\mathbf{b}}^* = \mathbf{X}^{\mathrm{MP}}\mathbf{Y} \sim \mathrm{N}(\mathbf{X}^{\mathrm{MP}}\mathbf{X}\mathbf{b}, \sigma^2(\mathbf{X}^T\mathbf{X})^{\mathrm{MP}}) \ .$$

Supposons qu'on veuille tester $H_0 : \mathbf{X}\mathbf{b} = 0$ contre $H_1 : \mathbf{X}\mathbf{b} \neq 0$. Une idée naturelle est alors de rejeter H_0 si $\|\mathbf{X}\widehat{\mathbf{b}}^*\|^2 > c$, où c est un seuil à calibrer. On a d'après le théorème de Cochran (voir le théorème 23),

$$\frac{\|\mathbf{X}\widehat{\mathbf{b}}^* - \mathbf{X}\mathbf{b}\|^2}{\sigma^2} \sim \chi^2(r) \ .$$

La statistique $\sigma^{-2}\|\mathbf{X}\widehat{\mathbf{b}}^* - \mathbf{X}\mathbf{b}\|^2$ dépend du paramètre inconnu σ^2 et n'est donc pas un pivot.

Pour estimer σ^2, on remarque que $\mathbf{Y} - \mathbf{X}\widehat{\mathbf{b}}^* = (\mathbf{I} - \Pi_{\mathbf{X}})\mathbf{Y}$ et $\mathbf{X}\widehat{\mathbf{b}}^* = \Pi_{\mathbf{X}}\mathbf{Y}$ sont indépendants et

$$\frac{\|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{b}}^*\|^2}{\sigma^2} \sim \chi^2(n-r) \ .$$

On en déduit alors que le rapport de ces deux statistiques suit une loi de Fisher

$$\frac{\|\mathbf{X}\widehat{\mathbf{b}}^* - \mathbf{X}\mathbf{b}\|^2}{\|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{b}}^*\|^2} \sim \mathcal{F}(r, n-r) \ .$$

On peut utiliser ce résultat pour calibrer le test. On choisit

$$c = \|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{b}}^*\|^2 f_{1-\alpha}(r, n-r) \ ,$$

où $F_{a,b}$ désigne la fonction de répartition de la loi de Fisher $\mathcal{F}(a,b)$ et $f_{1-\alpha}(a,b) = F_{a,b}^{-1}(1-\alpha)$ son $1-\alpha$ quantile. on a le résultat suivant.

Proposition 95. *Le test* $\phi(Z_n) = \mathbf{1}_{\{\|\boldsymbol{X}\widehat{\boldsymbol{b}}^*\|^2 > \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{b}}^*\|^2 f_{1-\alpha}(r,n-r)\}}$ *est un test de niveau* α.

Chapitre 10

Tests multiples

Supposons que $\Theta \subset \mathbb{R}^d$ et qu'on s'intéresse à l'hypothèse

$$H_0 : \theta_1 = \ldots = \theta_s = 0, \quad \text{contre} \quad H_1 : \exists i \in \{1, \ldots, s\} : \theta_i \neq 0 \ .$$

Il est typiquement insuffisant dans de nombreux problèmes pratiques de simplement accepter ou rejeter H_0. Si on accepte H_0, aucun de ces paramètres n'est significatif, la conclusion est claire. En revanche, lorsqu'on rejette H_0, on va typiquement vouloir savoir quels paramètres sont significants. Si H_0 est testée contre des alternatives bilatères, on va également s'intéresser au signe des paramètres significatifs.

Exemple 19. *Supposons que $X_1, \ldots, X_n$ est un n-échantillon aléatoire de loi Gaussienne $\mathrm{N}(\mu, \sigma^2)$ et qu'on s'intéresse à l'hypothèse $H_0 : \mu \leqslant 0, \sigma \leqslant 1$. En cas de rejet, on va vouloir savoir si c'est l'hypothèse sur la moyenne ou celle sur la variance qui a été rejetée, ou les deux.*

Exemple 20. *Supposons qu'on teste plusieurs traitements contre un contrôle. L'hypothèse nulle est qu'aucun de ces traitement n'apporte une amélioration, ou qu'il diffère du contrôle. En cas de rejet, on voudra typiquement savoir quel traitement présente une différence significative.*

Exemple 21. *Plutôt que comparer plusieurs traitements avec un contrôle, on peut vouloir comparer plusieurs alternatives entre elles. Si la qualité de l'alternative i est mesurée par le paramètre θ_i, l'hypothèse nulle va s'écrire*

$$H_0 : \theta_1 = \ldots = \theta_s \ .$$

De nombreux problèmes pratiques peuvent se mettre sous la forme donnée dans les exemples 20 et 21. C'est pourquoi le problème de tests multiples est également souvent appelé problème de comparaisons multiples.

Lorsqu'on s'intéresse à comparer plusieurs traitements en médecine, agriculture, ou dans l'industrie, le nombre de traitements (donc d'hypothèses) est plutôt faible (typiquement inférieur à 10). Quelques études sur l'éducation font intervenir des nombres d'hypothèses supérieures, qui peuvent aller jusqu'à 100. L'application la plus récente des tests multiples est en génétique pour le séquençage de l'ADN. Dans cette application, chaque gène est testé, le nombre d'hypothèses est donc de l'ordre de la dizaine de milliers tandis que le nombre d'individus à disposition, de l'ordre de quelques dizaines, est typiquement largement inférieur.

10.1 Cadre général

Dans tout ce chapitre, on suppose donné un modèle statistique $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$, qu'on suppose identifiable. Une hypothèse H est un ensemble de lois de $\mathcal{P}$, ou, de manière équivalente, un sous ensemble de Θ. On se donne un nombre fini d'hypothèses nulles $\mathcal{H}_0 = \{H_{0,i}, i = 1, \dots, s\}$ et d'alternatives $\mathcal{H}_1 = \{H_{1,i}, i \in \{1, \dots, s\}$ telles que, pour tout $i \in \{1, \dots, s\}$, $H_{0,i} \cap H_{1,i} = \varnothing$. On considère le problème de tester simultanément les hypothèses

$$H_{0,i} \quad \text{contre} \quad H_{1,i} \ .$$

On suppose qu'on dispose, pour tout $\alpha \in [0,1]$ et tout $i \in \{1, \dots, s\}$, d'un test $\phi_{i,\alpha}$ de niveau α de $H_{0,i}$ contre $H_{1,i}$. Le but de ce chapitre est de donner des façons de combiner ou d'agréger ces différents tests de façon à tester toutes les hypothèses simultanément.

La première idée serait de faire abstraction de la multiplicité et faire chaque test au niveau α. Ce faisant, le risque de rejeter au moins une hypothèse $H_{0,i}$ alors qu'elle est vraie va rapidement augmenter avec s. Si s est grand, la probabilité de cet événement va s'approcher de 1. Supposons par exemple que les tests $\{\phi_{i,\alpha}, i = 1, \dots, s\}$ sont indépendants, la table suivante donne l'évolution de la probabilité de rejeter au moins une bonne hypothèse lorsque s augmente.

s	1	2	5	10	50
$\mathbb{P}$(1 bonne hyp rejetée)	0.05	0.10	0.23	0.40	0.92

Dire dans ces conditions que la procédure globale contrôle la probabilité de rejet à un niveau 5% est alors clairement trompeur.

Dans ce chapitre, on remplace le risque de première espèce pour les tests d'une seule hypothèse H_0 par la probabilité de rejeter au moins une hypothèse vraie.

Definition 96. *Une hypothèse H est vraie pour $\mathbb{P}_\theta$ si $\theta \in H$, elle est fausse pour $\mathbb{P}_\theta$ si $\theta \notin H$. On note $\mathcal{T}_\theta$ l'ensemble des hypothèses vraies pour $\mathbb{P}_\theta$ et $\mathcal{F}_\theta$ l'ensemble des hypothèses fausses pour $\mathbb{P}_\theta$, c'est à dire*

$$\mathcal{T}_\theta = \{i \in \{1, \dots, s\} : \theta \in H_{0,i}\}, \qquad \mathcal{F}_\theta = \{1, \dots, s\} \setminus \mathcal{T}_\theta = \{i \in \{1, \dots, s\} : \theta \in H_{1,i}\} \ .$$

Definition 97. *Un test multiple est un ensemble aléatoire $R(Z_n) \subset \{1, \dots, s\}$ d'hypothèses rejetées. L'erreur de première espèce d'un test multiple appelée* FWER *(pour* family-wise error rate*) est mesurée par*

$$\mathrm{FWER}(R) = \sup_{\theta \in \Theta} \mathbb{P}_\theta(R(Z_n) \cap \mathcal{T}_\theta \neq \varnothing) \ .$$

Le test R est dit FWER α *si* $\mathrm{FWER}(R) \leqslant \alpha$.

Le FWER est donc la probabilité maximale de rejeter à tort une bonne hypothèse. Cette notion généralise celle de niveau au sens ou, si $s = 1$, le FWER est égal au niveau.

En pratique, la "famille" de tests considérée dépend typiquement de la situation. En laboratoire, il peut s'agir de l'ensemble des tests effectués en une journée, ou de ceux réalisés par un même testeur.

Un contrôle du FWER est une garantie dite *forte* sur un test multiple, par opposition à un contrôle du FWER faible, ou le sup dans la définition du FWER est restreint au sous-ensemble Θ_0 de Θ des lois telles que $\mathcal{T}_\theta = \{1, \dots, s\}$.

10.2 La procédure de Bonferroni

Plusieurs procédures d'agrégation de tests simples utilisent la p-valeur associée aux tests individuels $\phi_{i,\alpha}$. Notons, pour chaque $i \in \{1,\dots,s\}$ par $\hat{\alpha}_i = \hat{\alpha}_i(Z_n)$ la p-valeur de la famille de tests $\{\phi_{i,\alpha}, \alpha \in [0,1]\}$. Rappelons que

$$\forall i \in \{1,\dots,s\}, \forall \theta \in H_{0,i}, \forall u \in [0,1], \qquad \mathbb{P}_\theta(\hat{\alpha}_i \leqslant u) \leqslant u \ . \tag{10.1}$$

La procédure de Bonferroni est la plus simple façon d'agréger des tests de façon à garantir un contrôle du FWER. Il s'agit simplement de corriger le niveau auquel on effectue chaque test et à utiliser une borne d'union.

Théorème 98. *Le test multiple de Bonferroni est défini par*

$$R_B(Z_n) = \{i \in \{1,\dots,s\} : \hat{\alpha}_i \leqslant \alpha/s\} \ .$$

Il vérifie $\mathrm{FWER}(R_B) \leqslant \alpha$.

Preuve. Soit $\theta \in \Theta$, on a en utilisant une borne d'union et (10.1)

$$\begin{aligned}\mathbb{P}_\theta(R_B \cap \mathcal{T}_\theta \neq \varnothing) &= \mathbb{P}_\theta(\cup_{i\in\mathcal{T}(\theta)}\{\hat{\alpha}_i \leqslant \alpha/s\}) \\ &\leqslant \sum_{i\in\mathcal{T}(\theta)} \mathbb{P}_\theta(\hat{\alpha}_i \leqslant \alpha/s) \leqslant \frac{|\mathcal{T}_\theta|}{s}\alpha \leqslant \alpha \ .\end{aligned}$$

□

La procédure de Bonferroni est parfaitement satisfaisante en théorie et il est assez difficile de l'améliorer de ce point de vue. En revanche, en pratique, elle est rapidement incapable de détecter les fausses hypothèses lorsque s augmente puisqu'elle revient à tester chaque hypothèse au niveau α/s.

10.3 La procédure de Holm

La recherche de procédures élevant le niveau auquel chaque hypothèse est testée sans déteriorer le FWER est un sujet de grande importance en pratique. La solution présentée dans cette section a été proposée par Holm (1979). Notons $\hat{\alpha}_{i:s}$ les statistiques d'ordre des p-valeurs, de sorte qu'on a en particulier

$$\hat{\alpha}_{1:n} \leqslant \ldots \leqslant \hat{\alpha}_{n:n} \ .$$

Notons également pour tout $i \in \{1,\dots,s\}$ $H_{0,i:n}$ l'hypothèse testée par la p-valeur $\hat{\alpha}_{i:s}$.

La procédure de Holm est une procédure récursive qui procède de la façon suivante.

Step 1 Si $\hat{\alpha}_{1:n} \geqslant \alpha/s$, on accepte toutes les hypothèses et on s'arrête. Si $\hat{\alpha}_{1:n} < \alpha/s$, on rejette $H_{0,1:n}$ et on passe à l'étape 2.

Step k Si $\hat{\alpha}_{k:n} \geqslant \alpha/(s-k+1)$, on accepte $H_{k:n},\dots,H_{n:n}$ et on s'arrête. Si $\hat{\alpha}_{k:n} < \alpha/(s-k+1)$, on rejette $H_{k:n}$ et on passe à l'étape $k+1$.

L'algorithme de Holm termine en s étape au maximum. Notons $R_H = R_H(Z_n)$ l'ensemble des hypothèses rejetées au terme de l'algorithme. On a le résultat suivant.

Théorème 99. *Le test multiple de Holm R_H vérifie*

$$\mathrm{FWER}(R_H) \leqslant \alpha \ .$$

Preuve. Soit $\theta \in \Theta$ et soit $i_0 \in \mathcal{T}_\theta$ tel que

$$\hat{\alpha}_{i_0:n} = \min_{i \in \mathcal{T}_\theta} \hat{\alpha}_i \ .$$

Si plusieurs indices i_0 peuvent convenir, prenons le plus petit d'entre eux. On a nécessairement par construction $i_0 \leqslant s - |\mathcal{T}_\theta| + 1$. Pour que la procédure de Holm rejette une vraie hypothèse nulle pour $\mathbb{P}_\theta$, il faut que

$$\hat{\alpha}_{1:n} \leqslant \frac{\alpha}{s}, \qquad \hat{\alpha}_{i_0:n} \leqslant \frac{\alpha}{s - i_0 + 1} \ .$$

Si cet événement est réalisé, on a donc en particulier

$$\min_{i \in \mathcal{T}_\theta} \hat{\alpha}_i \leqslant \frac{\alpha}{s - i_0 + 1} \leqslant \frac{\alpha}{|\mathcal{T}_\theta|} \ .$$

Ainsi,

$$\mathbb{P}_\theta(R_H \cap \mathcal{T}_\theta \neq \varnothing) \leqslant \mathbb{P}_\theta\left(\min_{i \in \mathcal{T}_\theta} \hat{\alpha}_i \leqslant \frac{\alpha}{|\mathcal{T}_\theta|} \right) \leqslant \sum_{i \in \mathcal{T}_\theta} \mathbb{P}_\theta\left(\hat{\alpha}_i \leqslant \frac{\alpha}{|\mathcal{T}_\theta|} \right) \leqslant \alpha \ .$$

□

10.4 Méthodes récursives “step-down"

La procédure de Bonferroni est un exemple de procédures “à un pas" (*single-step*) tandis que la procédure de Holm est une procédure récursive dite *step-down*. Nous décrivons dans cette section les procédures step-down en général et donnons des conditions suffisantes pour vérifier qu'elles contrôlent le FWER. Soit $2^{\{1,\dots,s\}}$ l'ensemble de tous les $R \subset \{1, \dots, s\}$. De manière général, une procédure séquentielle utilise une application $S : 2^{\{1,\dots,s\}} \to 2^{\{1,\dots,s\}}$ et définit ensuite récursivement

$$R_{S,0} = \varnothing, \qquad R_{S,i+1} = R_{S,i} \cup S(R_{S,i}) \ .$$

L'application S est donc utilisée pour définir les hypothèses suivantes à rejeter une fois qu'on a rejeté $R_{S,i}$ Le test séquentiel R_S est finalement défini par $R_S = R_{S,s}$. Le résultat général étudiant les méthodes de rejet séquentiel est le suivant.

Théorème 100. *Supposons que l'application S vérifie les propriétés suivantes :*

1. *Pour tous $R \subset R'$, $S(R) \subset S(R') \cup R'$.*
2. *Pour tout $\theta \in \Theta$, $\mathbb{P}_\theta(S(\mathcal{F}_\theta) \subset \mathcal{F}_\theta) \geqslant 1 - \alpha$.*

Alors, la procédure de rejet séquentiel R_S associée à S vérifie $\mathrm{FWER}(R_S) \leqslant \alpha$.

Preuve. Soit $\theta \in \Theta$ et considérons l'événement $S(\mathcal{F}_\theta) \subset \mathcal{F}_\theta$. Sur cet événement, on a $R_{S,0} = \varnothing \subset \mathcal{F}_\theta$ et, pour tout t tel que $R_{S,t} \subset \mathcal{F}(\theta)$, on a, par l'hypothèse 1, $S(R_{S,t}) \subset \mathcal{F}_\theta \cup S(\mathcal{F}_\theta) = \mathcal{F}_\theta$, donc $R_{S,t+1} \subset \mathcal{F}_\theta$ et donc $R_S \subset \mathcal{F}(\theta)$. Ainsi, par l'hypothèse 2,

$$\mathbb{P}_\theta(R_S \cap \mathcal{T}_\theta \neq \varnothing) = 1 - \mathbb{P}_\theta(R_S \subset \mathcal{F}_\theta) \leqslant 1 - \mathbb{P}_\theta(S(\mathcal{F}_\theta) \subset \mathcal{F}_\theta) \leqslant \alpha.$$

□

On peut utiliser ce résultat général pour retrouver celui de Holm. Définissons pour tout sous ensemble $R \subset \{1, \dots, s\}$,

$$S(R) = \{i \notin R : \hat{\alpha}_i \leqslant \alpha/(s - |R|)\} \ .$$

On vérifie (**Exercice**) que la procédure de rejet séquentiel R_S associée à cette application est égale à la procédure de Holm R_H en montrant que toutes deux valent

$$\{i \in \{1, \dots, s\} : \hat{\alpha}_i = \hat{\alpha}_{j_0:n}, \ \forall j \leqslant j_0, \hat{\alpha}_{j:n} \leqslant \alpha/(s - j + 1)\} \ .$$

On peut vérifier que cette fonction S vérifie les hypothèses du théorème 100. Si $R \subset R'$, on a $|R| \leqslant |R'|$ donc, pour tout $i \notin R$ tel que $\hat{\alpha}_i \leqslant \alpha/(s - |R|)$, on a soit $i \in R'$, soit $\hat{\alpha}_i \leqslant \alpha/(s - |R'|)$, donc $i \in S(R')$, autrement dit, la première hypothèse du théorème 100 est bien vérifiée. Pour la seconde hypothèse, soit $\theta \in \Theta$, on a

$$\mathbb{P}_\theta(S(\mathcal{F}_\theta) \neq \varnothing) = \mathbb{P}_\theta\left(\cup_{i \in \mathcal{T}_\theta} \left\{\hat{\alpha}_i \leqslant \frac{\alpha}{|\mathcal{T}_\theta|}\right\}\right) \leqslant \sum_{i \in \mathcal{T}_\theta} \mathbb{P}_\theta\left(\hat{\alpha}_i \leqslant \frac{\alpha}{|\mathcal{T}_\theta|}\right) \leqslant \sum_{i \in \mathcal{T}_\theta} \frac{\alpha}{|\mathcal{T}_\theta|} = \alpha \ .$$

10.5 FDR

Lorsque le nombre de tests est de l'ordre de la dizaine ou de la centaine de milliers, demander un contrôle du FWER devient tellement exigeant que de nombreuses hypothèses fausses ne sont pas détectées. Une alternative initialement proposée par Benjamini et Hochberg dans un des articles de mathématiques les plus cités au monde consiste à contrôler plutôt le taux de faux positifs. Introduisons la proportion de faux positifs :

$$\forall \theta \in \Theta, \qquad \mathrm{FDP}(R, \theta) = \frac{|R \cap \mathcal{T}_\theta|}{|R| \vee 1} \ .$$

Le taux de faux positifs est alors simplement l'espérance de la proportion de faux positifs :

$$\mathrm{FDR}(R, \theta) = \mathbb{E}_\theta[\mathrm{FDP}(R, \theta)] \ .$$

On dit que le taux de faux positifs est contrôlé par α si $\sup_{\theta \in \Theta} \mathrm{FDR}(R, \theta) \leqslant \alpha$.

10.5.1 Heuristique

Essayons d'abord de déterminer quelle pourrait être la forme d'un test multiple ayant le FDR contrôlé. On veut bien sûr avec un nombre de faux positifs aussi grand que possible et un nombre de rejets aussi grand que possible. On se concentre sur les procédures de la forme

$$R(Z_n) = \{i \in \{1, \dots, s\} : \hat{\alpha}_i \leqslant t\} \ ,$$

où t est un seuil possiblement aléatoire. Si $t \in (\hat{\alpha}_{i:s}, \hat{\alpha}_{i+1:s})$, on a

$$\{i : \hat{\alpha}_i \leqslant t\} = \{i : \hat{\alpha}_i \leqslant \hat{\alpha}_{i:s}\} \ .$$

Ainsi, on peut se restreindre, sans perte de généralité, aux seuils de la forme $t = \hat{\alpha}_{i:s}$, et il suffit donc de se donner un choix de $i \in \{1, \dots, s\}$. Notons que, si $t = \hat{\alpha}_{i:s}$, on a $|R| = i$. De plus, pour tout t est déterministe, on a

$$\mathbb{E}_\theta[|R \cap \mathcal{T}_\theta|] = \sum_{i \in \mathcal{T}_\theta} \mathbb{E}_\theta[\mathbf{1}_{\{\hat{\alpha}_i \leqslant t\}}] \leqslant |\mathcal{T}_\theta| t \leqslant st \ .$$

En supposant que le taux de faux positifs est bien représenté par son espérance, on en déduit que le FDP vaut environ, si $t = \hat{\alpha}_{i:s}$,

$$\mathrm{FDP}(R, \theta) \approx \frac{s\hat{\alpha}_{i:s}}{i} \ .$$

Ceci suggère de prendre pour i le plus grand indice tel que $\hat{\alpha}_{i:s} \leqslant \alpha i/s$. La procédure de Benjamini et Hochberg fait précisément cela, en définissant

$$\hat{k} = \max\{k \in \{1, \dots, s\} : \hat{\alpha}_{k:s} \leqslant \alpha k/s\}, \qquad R_{\mathrm{BH}} = \{i \in \{1, \dots, s\} : \hat{\alpha}_i \leqslant \hat{\alpha}_{\hat{k}:s}\} \ .$$

10.5.2 Résultat général

On montre un résultat pour une procédure légèrement plus générale que celle présentée en heuristique.

Théorème 101. *Soit $\beta : \{1, \dots, s\} \to \mathbb{R}$ une fonction croissante et*

$$\hat{k} = \max\{k \in \{1, \dots, s\} : \hat{\alpha}_{k:s} \leqslant \alpha\beta(k)/s\}, \qquad R_{BH} = \{i \in \{1, \dots, s\} : \hat{\alpha}_i \leqslant \hat{\alpha}_{\hat{k}:s}\} \ .$$

On prend la convention $\hat{k} = 0$ si $\hat{\alpha}_{1:s} > \alpha\beta(1)/s$ et on pose alors $R_{BH} = \varnothing$. Alors, pour tout $\theta \in \Theta$,

$$\mathrm{FDR}(R_{BH}, \theta) \leqslant \alpha \frac{|\mathcal{T}_\theta|}{s} \sum_{j \geqslant 1} \frac{\beta(j \wedge s)}{j(j+1)} \ .$$

Preuve. Soit $\theta \in \Theta$, par définition de R_{BH}, on a

$$\begin{aligned} \mathrm{FDR}(R_{\mathrm{BH}}, \theta) &= \mathbb{E}_\theta\left[\frac{|R_{\mathrm{BH}} \cap \mathcal{T}_\theta|}{\hat{k} \vee 1}\right] = \sum_{i \in \mathcal{T}_\theta} \mathbb{E}_\theta\left[\frac{\mathbf{1}_{\{i \in R_{\mathrm{BH}}\}}}{\hat{k}} \mathbf{1}_{\{\hat{k} \geqslant 1\}}\right] \\ &\leqslant \sum_{i \in \mathcal{T}_\theta} \mathbb{E}_\theta\left[\mathbf{1}_{\{\hat{\alpha}_i \leqslant \alpha\beta(\hat{k})/s\}} \frac{\mathbf{1}_{\{\hat{k} \geqslant 1\}}}{\hat{k}}\right] \ . \end{aligned}$$

Pour tout $\hat{k} \geqslant 1$, on a

$$\frac{1}{\hat{k}} = \sum_{j \geqslant \hat{k}} \frac{1}{j(j+1)} = \sum_{j \geqslant 1} \frac{\mathbf{1}_{\{j \geqslant \hat{k}\}}}{j(j+1)} \ .$$

On a donc

$$\mathrm{FDR}(R_{\mathrm{BH}}, \theta) = \sum_{i \in \mathcal{T}_\theta} \sum_{j \geqslant 1} \frac{1}{j(j+1)} \mathbb{E}_\theta\left[\mathbf{1}_{\{\hat{\alpha}_i \leqslant \alpha\beta(\hat{k})/s\}} \mathbf{1}_{\{j \geqslant \hat{k}\}} \mathbf{1}_{\{\hat{k} \geqslant 1\}}\right] \ .$$

Si $j \geqslant \hat{k}$, alors $\beta(\hat{k}) \leqslant \beta(j \wedge s)$. Donc

$$\begin{aligned} \mathrm{FDR}(R_{\mathrm{BH}}, \theta) &\leqslant \sum_{i \in \mathcal{T}_\theta} \sum_{j \geqslant 1} \frac{1}{j(j+1)} \mathbb{E}_\theta\left[\mathbf{1}_{\{\hat{\alpha}_i \leqslant \alpha\beta(j \wedge s)/s\}} \mathbf{1}_{\{j \geqslant \hat{k}\}} \mathbf{1}_{\{\hat{k} \geqslant 1\}}\right] \\ &\leqslant \sum_{i \in \mathcal{T}_\theta} \sum_{j \geqslant 1} \frac{1}{j(j+1)} \mathbb{E}_\theta\left[\mathbf{1}_{\{\hat{\alpha}_i \leqslant \alpha\beta(j \wedge s)/s\}}\right] \leqslant \sum_{i \in \mathcal{T}_\theta} \frac{\alpha}{s} \sum_{j \geqslant 1} \frac{\beta(j \wedge s)}{j(j+1)} \ . \end{aligned}$$

□

Remarquons que le théorème général ne permet de contrôler le FDR au niveau α que si β est choisie de façon à ce que $\sum_{j\geqslant 1} \frac{\beta(j\wedge s)}{j(j+1)} \leqslant 1$. Or, le choix de Benjamini et Hochberg de $\beta(j) = j$ conduit à

$$\sum_{j\geqslant 1} \frac{\beta(j\wedge s)}{j(j+1)} = \sum_{j=2}^{s+1} \frac{1}{j} + s \sum_{j\geqslant s} \frac{1}{j(j+1)} = \sum_{j=1}^{s+1} \frac{1}{j} > 1 \quad .$$

Il est possible de montrer que la procédure de Benjamini et Hochberg contrôle le FDR au niveau α sous des hypothèses supplémentaires. C'est le cas par exemple si les p-valeurs sont indépendantes.

En revanche, il est aussi possible de montrer que la borne générale est précise au sens ou il existe des distributions de p-valeurs pour lesquelles la borne sup est atteinte, donc pour lesquelles le test de Benjamini et Hochberg ne contrôle pas le FDR au niveau α, mais seulement au niveau $\alpha \sum_{j=1}^{s+1} \frac{1}{j} \sim \alpha \log(s)$ lorsque $s \to \infty$.

Chapitre 11

Statistiques robustes

On dit qu'un estimateur est robuste s'il est peu sensible à la présence d'outliers dans le jeu de données. On se place dans tout ce chapitre dans le cas où l'observation $Z_n = (X_1, \dots, X_n)$ est un n-échantillon d'un modèle de translation

$$X_i = \theta + \varepsilon_i, \qquad i = 1, \dots, n \ ,$$

où $\theta \in \Theta$ est un paramètre inconnu et ε_i sont des variables aléatoires i.i.d. dont on note F la fonction de répartition et f la densité.

11.1 Courbe de sensibilité

Soit $\hat{\theta}$ un estimateur de θ. On voit ici $\hat{\theta}$ comme un algorithme $\hat{\theta} : \cup_{n\geqslant 1} \mathcal{X}^n \to \Theta$. Pour mesurer la robustesse d'un estimateur, on introduit la notion suivante.

Definition 102. *Soit $x \in \mathcal{X}$ et $Z_n(x) = (X_1, \dots, X_n, x)$ l'échantillon observé augmenté d'une donnée x. La courbe de sensibilité de l'estimateur $\hat{\theta}$ est définie par*

$$S(x, \hat{\theta}) = (n+1)(\hat{\theta}(Z_n(x)) - \hat{\theta}(Z_n)) \ .$$

Dans cette définition, x représente un outlier, une donnée venue corrompre les observations qui va donc typiquement être "loin" du nuage de points Z_n. La courbe de sensibilité mesure l'impact que peut avoir un seul outlier situé à la position x. Si cette fonction est bornée, l'estimateur est donc plus "robuste".

Supposons d'abord que $\Theta = \mathbb{R}$ et que $\hat{\theta}(Z_n) = \frac{1}{n}\sum_{i=1}^n X_i$. Alors $\hat{\theta}(Z_n(x)) = \frac{1}{n+1}x + \frac{n}{n+1}\hat{\theta}(Z_n)$, donc

$$S(x, \hat{\theta}) = x + \hat{\theta}(Z_n) \ .$$

La fonction d'influence est une fonction linéaire, donc non bornée de x, ce qui indique, comme on s'y attendait, que la moyenne empirique est sensible à la taille des outliers, et qu'un seul outlier est suffisant pour détériorer les performances de cet estimateur.

En revanche, la médiane empirique ne peut varier, si par exemple n est impaire, que de $X_{(n+1)/2:n}$ à un réel de $[X_{(n-1)/2:n}, X_{(n+3)/2:n}]$. Autrement dit, peu importe la taille x de l'outlier, la déviation de la médiane est bornée indépendamment de x. La médiane est donc plus robuste en ce sens que la moyenne à la présence d'un outlier dans le jeu de données.

11.2 Fonction d'influence

La courbe de sensibilité est une fonction aléatoire, dépendant de l'échantillon d'observation Z_n. On présente dans cette section une mesure alternative de robustesse basée sur la loi des données plutôt que sur leur réalisation. On utilise ici la représentation du paramètre comme fonction de la loi des observations et on écrit $\theta = T(F_X)$. Ainsi, on avait vu que, si θ est l'espérance de X,

$$T(F_X) = \int x \mathrm{d}F_X(x) \ ,$$

et si θ est la médiane $T(F_X) = F_X^{-1}(1/2)$. De plus, comme T est une fonctionnelle de translation, on a $T(F_X) = T(F) + \theta$.

On estime $\theta = T(F_X)$ par son estimateur plug-in obtenu en estimant la loi F_X inconnue par la fonction de répartition empirique F_n et en injectant cet estimateur dans la fonctionnelle T pour obtenir l'estimateur $\widehat{\theta} = T(F_n)$.

Une façon alternative et qui est sera utilisée dans ce chapitre est de voir θ comme solution d'une équation intégrale et $\widehat{\theta}$ comme la solution de la version empirique de cette équation. Ainsi, la moyenne empirique $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$ est solution de l'équation

$$\sum_{i=1}^n (X_i - \widehat{\theta})\frac{1}{n} = 0 \ .$$

L'équation $\sum_{i=1}^n (X_i - \theta')\frac{1}{n}$ converge pour tout θ' vers $\mathbb{E}[X - \theta']$, et l'équation $\mathbb{E}[X - \theta'] = 0$ a clairement pour solution le paramètre $\theta = \mathbb{E}[X]$ estimé par $\widehat{\theta}$.

Comme pour la courbe de sensibilité, on va perturber la loi F_X par une masse ponctuelle en un point x déterministe et considérer la loi

$$F_{x,\epsilon} = (1-\epsilon)F_X + \epsilon\Delta_x \ ,$$

où

$$\Delta_x(t) = \begin{cases} 0 & \text{si } t < x \ , \\ 1 & \text{si } t \geqslant x \ . \end{cases}$$

La distribution correspondant à la fonction de répartition $F_{x,\epsilon}$ est un mélange. Une proportion $1-\epsilon$ des données a été générée par la loi F_X et une proportion ϵ des données est située au point x.

Pour tout $t \in \mathbb{R}$, on a $|F_X(t) - F_{x,\epsilon}(t)| \leqslant \epsilon$, donc si T est une fonctionnelle robuste, on doit avoir $T(F_{x,\epsilon})$ proche de $T(F_X)$.

Definition 103. *La fonction d'influence de la fonctionnelle T est la fonction définie en tout point x par*

$$\mathrm{IF}(x,T) = \lim_{\epsilon \to 0} \frac{T(F_{x,\epsilon}) - T(F_X)}{\epsilon} \ ,$$

lorsque cette limite existe.

L'estimateur $\widehat{\theta} = T(F_n)$ est dit robuste *si* $\sup_x |\mathrm{IF}(x,T)| < \infty$.

Exemple 22. *Pour la fonctionnelle espérance, on a, si $U \sim F_{x,\epsilon}$,*

$$\mathbb{E}[U] = (1-\epsilon)\mathbb{E}[X] + \epsilon x \ .$$

Donc $\mathrm{IF}(x,\mathbb{E}) = x - \mathbb{E}[X]$*, donc la moyenne empirique n'est robuste que si $\mathcal{X}$ est borné.*

Exemple 23. *Pour la fonctionnelle médiane, on a, pour tout t,*

$$\mathbb{P}(U \leqslant t) = (1-\epsilon)F_X(t) + \epsilon \mathbf{1}_{t \geqslant x} \ .$$

Donc,

$$\mathbb{P}(U \leqslant t) = u \quad \Longleftrightarrow \quad t = \begin{cases} F_X^{-1}\left(\frac{u}{1-\epsilon}\right) & si\ u < F_X(x) \ , \\ F_X^{-1}\left(\frac{u-\epsilon}{1-\epsilon}\right) & si\ u \geqslant F_X(x) \ . \end{cases}$$

Ainsi, on a

$$mediane(F_U) = F_U^{-1}(1/2) = \begin{cases} F_X^{-1}\left(\frac{1/2}{1-\epsilon}\right) & si\ 1/2 < F_X(x) \ , \\ F_X^{-1}\left(\frac{1/2-\epsilon}{1-\epsilon}\right) & si\ 1/2 \geqslant F_X(x) \ . \end{cases}$$

La fonction d'influence est alors la dérivée en 0 de cette fonction, qui existe si $f_X(\theta) \neq 0$ et vaut alors

$$\begin{aligned} \mathrm{IF}(x, mediane) &= \begin{cases} \frac{1}{2f_X(\theta)} & si\ x > \theta \\ \frac{-1}{2f_X(\theta)} & si\ x \leqslant \theta \end{cases} \\ &= \frac{\mathrm{sgn}(x-\theta)}{2f_X(\theta)} \ . \end{aligned}$$

La fonction d'influence a été très étudiée, on pourra se référer au livre de Huber et Ronchetti "robust statistics" pour lire de nombreux résultats. On présente ici quelques propriétés utiles dont les preuves peuvent être trouvées dans le livre.

Théorème 104. *Soit T une fonctionnelle telle que $\mathrm{IF}(x,T)$ est bien définie. On a les propriétés suivantes*

1. *Si elle existe $\mathbb{E}[\mathrm{IF}(X,T)] = 0$.*
2. *Si elle existe $\mathrm{Var}(\mathrm{IF}(X,T)) = \mathbb{E}[\mathrm{IF}(X,T)^2]$.*
3. *$\sqrt{n}(T(F_n) - T(F_X)) \overset{\mathbb{P}}{\Longrightarrow} \mathrm{N}(0, \mathbb{E}[\mathrm{IF}(X,T)^2])$.*

Remarque 105. *Le point 2 est une conséquence immédiate du point 1. Le point 3 peut être vu comme une "méthode Delta fonctionnelle". On peut vérifier qu'il permet de retrouver le théorème de la limite centrale dans le cas ou T est la fonctionnelle $\mathbb{E}$ et un théorème de normalité asymptotique de la médiane cohérent avec le résultat qu'on déduit du théorème 27.*

11.3 Breakdown point d'un estimateur

La courbe de sensibilité et la fonction d'influence mesurent la sensibilité d'un estimateur à une petite fraction d'outliers. Dans plusieurs applications, la présence d'outliers dans un jeu de données est due à un problème matériel qui peut se reproduire et conduire à une proportion ϵ d'outliers dans le jeu de données. Pour étudier la résistance d'estimateurs dans ce genre de situations, on étudie dans cette section la proportion maximale d'outliers que peut contenir un échantillon avant que cet estimateur ne s'"effondre", c'est à dire qu'il n'estime plus efficacement le paramètre d'intérêt. Soit $Z_n = (X_1, \ldots, X_n)$ un échantillon aléatoire et soit $Z_n^* = (x_1^*, \ldots, x_m^*, X_{m+1}, \ldots, X_n)$ un échantillon corrompu, dans

lequel les m premières données sont des outliers qui ont pu être générées après l'observation de Z_n, de façon à rendre cette corruption aussi néfaste que possible (de tels outliers sont dits "adversariaux"). Le biais de l'estimateur $\widehat{\theta}$ est défini par

$$\text{biais}(m, Z_n, \widehat{\theta}) = \sup_{x_1^*,\ldots,x_m^*} |\widehat{\theta}(Z_n) - \widehat{\theta}(Z_n^*)| \ .$$

Definition 106. *Si* $\text{biais}(m, Z_n, \widehat{\theta}) = +\infty$*, on dit que l'estimateur* $\widehat{\theta}$ *s'est effondré (break down). Soit*

$$\epsilon_n^* = \min\{m/n : \text{biais}(m, Z_n, \widehat{\theta}) = +\infty\} \ .$$

ϵ_n^* *est appelé le breakdown point non asymptotique de* $\widehat{\theta}$ *et, si* $\epsilon_n^* \to \epsilon^*$, ϵ^* *est appelé le breakdown point de* $\widehat{\theta}$*.*

Exemple 24. *Soit* $\widehat{\theta} = n^{-1}\sum_{i=1}^n X_i$ *l'estimateur de* $\mathbb{E}[X]$*. On a, pour tout* $m \geqslant 1$*,*

$$\text{biais}(m, Z_n, \widehat{\theta}) = \sup_{x_1^*,\ldots,x_m^*} |\frac{1}{n}\sum_{i=1}^m (x_i^* - X_i)| = +\infty \ .$$

Ainsi, $\epsilon_n^* = 1/n$ *et le breakdown point de la moyenne empirique est nul.*

Exemple 25. *Soit* $\widehat{\theta}(Z_n) = \text{mediane}\{X_i, i \in \{1,\ldots,n\}\}$ *l'estimateur de la médiane* $\theta = F^{-1}(1/2)$*. Il est clair que* $\widehat{\theta}(Z_n^*)$ *n'appartient pas à l'enveloppe convexe des données* Z_n *seulement si* m *est supérieur à* $n/2$*, de sorte que* $\epsilon_n^* \geqslant 1/2$*. Comme on a aussi* $\epsilon_n^* \leqslant 1/2 + 1/(2n)$ *(selon la parité de* n*), on en déduit que* $\epsilon^* = 1/2$*.*

11.4 Estimateurs de Huber

L'estimateur robuste probablement le plus célèbre dans la littérature et en pratique est l'estimateur de Huber. On présente dans la fin de ce chapitre cet estimateur et on en donne quelques propriétés élémentaires.

Definition 107. *Pour tout* $c > 0$*, on définit*

$$\psi_c(x) = \begin{cases} \frac{x^2}{2} & \text{si } |x| \leqslant c \\ c|x| & \text{si } |x| > c \end{cases} \ .$$

La fonction ψ_c *est de classe* $\mathcal{C}^1$ *sur* $\mathbb{R}$*, de dérivée*

$$\psi_c'(x) = \begin{cases} x & \text{si } |x| \leqslant c \\ c\,\text{sgn}(x) & \text{si } |x| > c \end{cases} \ .$$

En particulier, ψ_c' *est uniformément majorée par* c*.*

L'estimateur de Huber $\widehat{\theta}_c$ *au niveau* c *est défini par*

$$\widehat{\theta}_c \in \text{argmin}_{\theta\in\Theta} \sum_{i=1}^n \psi_c(X_i - \theta)$$

Il est aussi solution de l'équation $\sum_{i=1}^n \psi_c'(X_i - \theta) = 0$*.*

11.4.1 Fonctionnelle de Huber

Supposons que $Z_n = (X_1, \dots, X_n)$ est un échantillon aléatoire d'un modèle de translation, i.e.

$$X_i = \theta + \varepsilon_i \ ,$$

avec $\varepsilon_1, \dots, \varepsilon_n$ n échantillon de densité f. On suppose en outre dans toute la suite que f *est une fonction paire.*

Soit $T_c(F_X)$ tel que $0 = \mathbb{E}[\psi_c'(X - T_c(F_X))]$. On a, comme f est paire et ψ_c' est impaire pour tout x,

$$\mathbb{E}[\psi_c'(X-\theta)] = \int_{-\infty}^{+\infty} \psi_c'(x-\theta) f(x-\theta)\mathrm{d}x = \int_{-\infty}^{+\infty} \psi_c'(x) f(x)\mathrm{d}x = 0 \ .$$

Donc T_c est une fonctionnelle de translation.

Soit $h \in (0, c)$, on a aussi

$$\begin{aligned}\mathbb{E}[\psi_c'(X-\theta-h)] &= \int_{-\infty}^{+\infty} (\psi_c'(x-h) - \psi_c'(x)) f(x)\mathrm{d}x \\ &= \int_{-c}^{h+c} (\psi_c'(x-h) - \psi_c'(x)) f(x)\mathrm{d}x \ .\end{aligned}$$

On en déduit alors d'un part que

$$\mathbb{E}[\psi_c'(X-\theta-h)] \leqslant \int_{h-c}^{c} (\psi_c'(x-h) - \psi_c'(x)) f(x)\mathrm{d}x = -h\mathbb{P}(\varepsilon \in [h-c, c]) \ .$$

D'autre part, comme ψ_c' est 1-Lipschitz, que

$$\mathbb{E}[\psi_c'(X-\theta-h)] \geqslant -h\mathbb{P}(\varepsilon \in [-c, c+h]) \geqslant -h \ .$$

11.4.2 Asymptotique de l'estimateur de Huber

Evaluons la fonction d'influence de T. Soit U de c.d.f. $(1-\epsilon)F_X + \epsilon\Delta_x$. On a, pour tout $\theta' = \theta + h$,

$$\mathbb{E}[\psi_c'(U-\theta')] = (1-\epsilon)\mathbb{E}[\psi_c'(X-\theta-h)] + \epsilon\psi_c'(x-\theta-h) \ ,$$

donc, en raisonnant de façon informelle (on pourra en exercice chercher à rendre rigoureux ce raisonnement),

$$\mathbb{E}[\psi_c'(U-\theta')] \approx -h(1-\epsilon)\mathbb{E}[\psi_c''(X-\theta)] + \epsilon\psi_c'(x-\theta) + h\epsilon\psi_c''(x-\theta) \ .$$

Donc $\mathbb{E}[\psi_c'(U-\theta')] = 0$ si $\theta' = \theta + h$, avec

$$h \approx \frac{\epsilon\psi_c'(x-\theta)}{(1-\epsilon)\mathbb{E}[\psi_c''(X-\theta)] + \epsilon\psi_c''(x-\theta)} \ .$$

On en déduit que

$$\mathrm{IF}(x, T_c) = \frac{\psi_c'(x-\theta)}{\mathbb{E}[\psi_c''(X-\theta)]} = \frac{\psi_c'(x-\theta)}{\mathbb{P}(\varepsilon \in [-c, c])} \ .$$

D'après le point 3 du théorème 104, on en déduit que l'estimateur de Huber $\widehat{\theta}_c$ est asymptotiquement normal et vérifie

$$\sqrt{n}(\widehat{\theta}_c - \theta) \xRightarrow{\mathbb{P}} \mathrm{N}\left(0, \frac{\mathbb{E}[\psi_c'(X-\theta)^2]}{\mathbb{P}(\varepsilon \in [-c, c])^2}\right) \ . \tag{11.1}$$

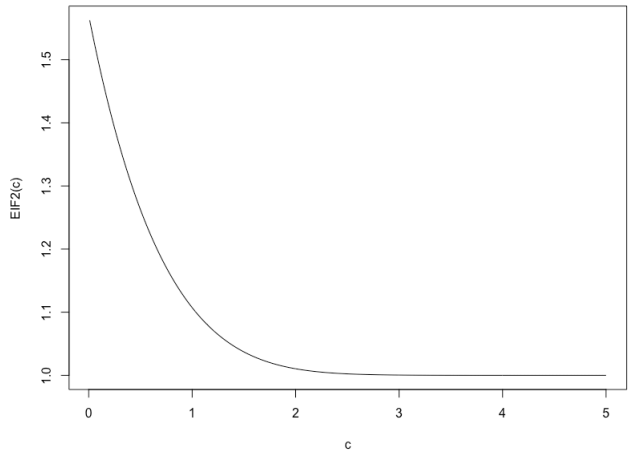

FIGURE 11.1 – Variance limite de l'estimateur de Huber

On peut évaluer cette variance asymptotique dans le cas où $f(x) = \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$, on a

$$\begin{aligned}\mathbb{E}[\psi_c'(X-\theta)^2] &= \mathbb{E}[(X-\theta)^2] + \mathbb{E}[(c^2-(X-\theta)^2)\mathbf{1}_{\{|X-\theta|>c\}}] \\ &= 1 - 2\int_c^{+\infty}(x^2-c^2)\mathrm{e}^{-x^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}} \\ &= 1 + 2[(c^2-1)(1-\Phi(c)) - \frac{c\mathrm{e}^{-c^2/2}}{\sqrt{2\pi}}] \ .\end{aligned}$$

La variance asymptotique de l'estimateur de Huber est donc

$$\frac{1 + 2[(c^2-1)(1-\Phi(c)) - \frac{c\mathrm{e}^{-c^2/2}}{\sqrt{2\pi}}]}{(2\Phi(c)-1)^2} \ .$$

On peut visualiser cette fonction en fonction de c à la figure 11.1. On voit en particulier que cette variance converge rapidement vers la variance optimale 1 lorsque $c \to \infty$.

11.4.3 Concentration de l'estimateur de Huber

Pour finir notre exploration de l'estimateur de Huber, on donne un résultat de concentration de $\widehat{\theta}_c$ autour de θ.

Théorème 108. *Soit $\sigma^2 = \mathbb{E}[\psi_c'(X-\theta)^2]$ la variance asymptotique de l'estimateur de Huber. Pour tout c, t et n tels que, en notant $p_c = \mathbb{P}(\varepsilon \in [0,c])$,*

$$c > \frac{24\sigma}{23}, \qquad t < \frac{np_c^2}{16} \ .$$

on a

$$\mathbb{P}\left(|\widehat{\theta}_c - \theta| > 4\left(\sqrt{\frac{\sigma^2 t}{np_c^2}} + \frac{ct}{6np_c}\right)\right) \geqslant 1 - 2\mathrm{e}^{-t} \ .$$

Remarque 109. *La constante* 16 *devant le terme* $\sqrt{\sigma^2 t/n}$ *n'est pas la constante optimale* $\sqrt{2}$ *qu'on pourrait obtenir avec une version précise non asymptotique du résultat* (11.1). *C'est un choix permettant de donner un résultat à peu près "digeste". Toutefois, on pourra montrer en utilisant le même type d'arguments que, pour tout* $\epsilon > 0$, *il existe des constantes* c_ϵ, c'_ϵ *telles que, si* $c > c_\epsilon(\sqrt{\mathbb{E}[\varepsilon^2]} \vee \sigma)$, $t < c'_\epsilon n$, *on peut montrer le même résultat avec une constante* $\sqrt{2} + \epsilon$ *devant le terme principal* $\sqrt{\sigma^2 t/n}$.

Démonstration. Avant d'entrer dans le détail de la preuve, donnons en les idées principales qui sont dues à Catoni dans son article "Challenging the empirical mean, a deviation study".

1. On va montrer par des arguments vus au chapitre 3 que, pour tout θ' et $t > 0$, $\mathbb{P}(\Omega(\theta')) \geqslant 1 - \mathrm{e}^{-t}$, où
$$\Omega(\theta') = \left\{ \frac{1}{n} \sum_{i=1}^{n} \psi'_c(X_i - \theta') - \mathbb{E}[\psi'_c(X - \theta')] \leqslant v_n(t) \right\} .$$
2. On va alors déterminer $\theta_+(t,n)$ déterministe tel que,
$$\forall \theta' \geqslant \theta_+(t,n), \qquad \mathbb{E}[\psi'_c(X - \theta')] + v_n(t) < 0 .$$
3. Sur l'événement $\Omega(\theta_+(t,n))$, on a alors, pour tout $\theta' \geqslant \theta_+(t,n)$,
$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \psi'_c(X_i - \theta') &\leqslant \frac{1}{n} \sum_{i=1}^{n} \psi'_c(X_i - \theta_+(t,n)) \text{ (car } \psi'_c \text{ est croissante)} \\ &\leqslant \mathbb{E}[\psi'_c(X - \theta_+(t,n))] + v_n(t) < 0 . \end{aligned}$$
On a donc nécessairement $\widehat{\theta}_c \leqslant \theta_+(n,t)$ sur $\Omega(\theta_+(t,n))$ et ainsi
$$\mathbb{P}(\widehat{\theta}_c \leqslant \theta_+(n,t)) \geqslant \mathbb{P}(\Omega(\theta_+(t,n))) \geqslant 1 - \mathrm{e}^{-t} .$$
4. On raisonne de manière symétrique pour obtenir la borne inférieure.

Pour le point 1, comme ψ'_c est majorée par c, on a, pour tout $k \geqslant 3$, pour tout θ',
$$(\psi'_c(X - \theta') - \mathbb{E}[\psi'_c(X - \theta')]_+^k \leqslant c^{k-2} (\psi'_c(X - \theta') - \mathbb{E}[\psi'_c(X - \theta')])^2 .$$
Notons $\sigma^2_{\theta'} = \mathrm{Var}(\psi'_c(X - \theta'))$. D'après la proposition 51, $\psi'_c(X - \theta)$ appartient à $\mathrm{sPoi}(\sigma^2_{\theta'}, c)$, donc, pour tout $\theta' \in \Theta$ et tout $t > 0$, avec probabilité $1 - \mathrm{e}^{-t}$,
$$\frac{1}{n} \sum_{i=1}^{n} \psi'_c(X_i - \theta') \leqslant \mathbb{E}[\psi'_c(X - \theta')] + \sqrt{\frac{2\sigma^2_{\theta'} t}{n}} + \frac{ct}{3n} .$$
Passons au point 2. Pour tout $\theta' \in (\theta, \theta + c)$, on a, en utilisant les inégalités de la fin de la section 11.4.1 et le fait que ψ'_c est 1-Lipschitz,
$$\begin{aligned} \mathbb{E}[\psi'_c(X - \theta')] &\leqslant -(\theta' - \theta) p_c , \\ \sigma^2_{\theta'} \leqslant \mathbb{E}[\psi'_c(X - \theta')^2] &\leqslant 2\sigma^2 + 2(\theta - \theta')^2 . \end{aligned}$$

Soit $v_n(t) = \sqrt{\frac{4\sigma^2 t}{n}} + \frac{ct}{3n}$. On a donc

$$\mathbb{E}[\psi_c'(X-\theta')] + \sqrt{\frac{2\sigma_{\theta'}^2 t}{n}} + \frac{ct}{3n} \leqslant -(\theta'-\theta)\left(p_c - 2\sqrt{\frac{t}{n}}\right) + v_n(t) < 0 \ ,$$

si $\theta' > \theta + v_n(t)/(p_c - 2\sqrt{t/n})$. On vérifiera que, sous les conditions du théorème, on a d'une part

$$\frac{v_n(t)}{p_c - 2\sqrt{t/n}} < \frac{2v_n(t)}{p_c} \ .$$

De plus,

$$\frac{ct}{3p_c n} < \frac{c}{48}, \qquad \sqrt{\frac{4\sigma^2 t}{np_c^2}} < \frac{\sigma}{2} < \frac{23c}{48}, \quad \text{donc} \quad \frac{2v_n(t)}{p_c} < c \ .$$

On conclut avec les arguments du point 3 de l'heuristique de la preuve.

La preuve de la borne inférieure est laissée à titre d'exercice. □