Statistiques
Master Statistique et Économétrie
Notes de cours

V. Monbet

Master 1 - 2012

Table des matières

Chapitre 1

Introduction

Généralités

En probabilité on supose (implicitement ou explicitement) que tous les paramètres nécessaires pour calculer des probabilités associées à un modèle particulier sont connus. Par exemple, on peut calculer la probabilité d'occurence d'un évènement donné de façon exacte ou approchée (à l'aide des théorèmes limite). En statistique, le rôle des paramètres (des modèles de probabilité) et des évènement/résultats (liés à une expérience) sont en quelques sortes inversés. Le résultat d'une expérience est observé par l'expérimentateur tandis que la vraie valeur du paramètre (et plus généralement le modèle de probabilité) est inconnue de l'expérimentateur. Ainsi, l'objectif de la statistique est d'utiliser les résultats d'une expérience (c'est à dire les données) pour inférer la valeur des paramètres inconnus du modèle de probabilité supposé être sous-jacent.

Le paragraphe précédent suggère qu'il n'y a pas d'ambiguité dans le choix du modèle sous-jacent à une expérience donnée. Cependant, en réalité dans les problèmes statistiques, on verra qu'il y a de nombreuses incertitudes quand au choix du modèle et ce choix est souvent fait essentiellement sur la base des observations. Et dans la grande majorité des cas, le modèle est seulement une approximation de la réalité ; et il est important pour un statisticien de vérifier que les modèles supposés sont plus ou moins proches de la réalité et d'être conscient des conséquences du choix d'un "mauvais" modèle.

Une philosophie reconnue en statistique est qu'un modèle doit être aussi simple que possible. On préférera toujours un modèle ayant peu de paramètres à un modèle caractérisé par un grand nombre de paramètres.

Notations

Dans la mesure du possible, nous utiliserons les notations suivantes dans ce cours :

- Variables aléatoires : lettres majuscules (ex : $X_1, \cdots, X_n$)
- Observations : lettres minuscules (ex : $x_1, \cdots, x_n$))
- Paramètres : lettres greques (ex : θ, μ, σ)

Plan du cours

Estimation ponctuelle

Dans la première partie du cours, nous aborderons le problème de l'estimation ponctuelle de paramètres.

Exemple : On suppose que l'on a deux candidats A et B lors d'une élection. On cherche à prédire la proportion de votes pour A à partir d'un échantillon représentatif de taille n sélectionné par un institut de sondage. Chaque individu de l'échantillon donne son intention de vote. On modélise le choix A pour l'individu i par une variable aléatoire

$$X_i = 1 \text{ si } i \text{ vote A}, X_i = 0 \text{ sinon}$$

Les X_i suivent une loi de Bernouilli de paramètre π inconnu. On dispose de n observations $x_1, \cdots, x_n$ de $X_1, \cdots, X_n$. On cherche à inférer π à partir de l'échantillon $x_1, \cdots, x_n$. Une estimation naturelle est

$$\hat{\pi}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

On se pose alors des questions : Cette quantité estime t'elle bien le paramètre π ? Peut-on lui associer une marge d'erreur ? Que se passe t'il si on dispose d'un échantillon plus grand (c'est à dire si n grandit) ? Existe t'il d'autres quantités qui donnerait une meilleure estimation de π ? En existe t'il une qui est optimale pour un critère bien choisi ?

Théorie de la décision : tests statistiques

Dans une seconde partie, nous introduirons le concept de test statistique.

Exemple : On teste l'efficacité d'un médicament contre le cholestérol. On dispose pour cela de n individus pour lesquels on a effectué deux mesures du taux de cholestérol, l'une avant et l'autre après le traitement. On note

- X : le taux de cholestérol avant le traitement,
- Y : le taux de cholestérol après le traitement.

On dispose donc des couples d'observations $(x_1, y_1), \cdots, (x_n, y_n)$ et on veut déterminer à l'aide de ces observations si

$$D = Y - X$$

est positif. Etant donnés les caractères aléatoires de D et de l'expérience, on décidera de l'efficacité du traitement si la moyenne observée

$$\bar{d}_n = \frac{1}{n}(y_i - x_i)$$

est plus grande qu'un seuil. Ce seuil est défini en fonction du risque de se tromper qui est fixé par l'expérimentateur et de la taille n de l'échantillon.

Chapitre 2

Rappels de probabilité

2.1 Généralités

2.1.1 Évènement aléatoires

Une expérience aléatoire est une expérience dont l'issue est incertaine. Exemple : partie de foot, A gagne dans 40% des cas, etc.

On note Ω l'ensemble des issues (ou évènements) possibles. Un sous ensemble de Ω est appelé un évènement.

2.1.2 Mesures de probabilité

Étant donné un évènement sur Ω, on définit une fonction ou mesure $P(.)$ sur les sous ensembles de Ω qui associe un nombre réel à chaque évènement ; ce nombre représente la probabilité qu'un certain évènement arrive. Cette fonction doit être consistente. Par exemple, si $A \subset B$ alors on doit avoir $P(A) \leq P(B)$.

Définition 1 *$P(.)$ est appelée mesure de probabilité si les axiomes suivants sont satisfaits :*

1. $P(A) \geq 0$ *quelque soit* $A \in \Omega$
2. $P(\Omega) = 1$
3. *Si* $A_1, A_2, \cdots$ *sont des évènements disjoints alors* $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

Si une expérience est infiniment répétable, on peut interpréter $P(A)$ comme la fréquence relative de l'occurence de A ; si on répète l'expérience N fois, on peut approcher $P(A)$ par k/N avec k le nombre de fois où A arrive.

Proposition 1 *La définition précédente a les conséquences suivantes.*

1. $P(A^c) = 1 - P(A)$.
2. $P(A \cap B) \leq \min(P(A), P(B))$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
4. *Soit* $\{A_n\}$ *une suite telle que* $A_n \subset A_{n+1}$ *avec* $A = \cup_{n=1}^{\infty} A_n$. *Alors,* $P(A_n) \to P(A)$ *quand* $n \to \infty$.
5. *Quelques soient les évènements* $A_1, A_2, \cdots$, $P(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n)$.

2.1.3 Probabilités conditionelles et indépendance

Probabilités conditionelles

Si on sait qu'un certain évènement est arrivé, on peut exploiter cette connaissance pour affiner la probabilité qu'un autre évènement arrive.

Définition 2 *Supposons que A et B sont des évènements de Ω. Si $P(B) > 0$ alors*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

est appelée probabilité conditionnelle de A sachant B.

Proposition 2 - *Loi des probabilités totales*
Si $B_1, B_2, \cdots$ sont des évènements disjoints avec $P(B_k) > 0$ pour tout k et $\cup_{k=1}^{\infty} B_k = \Omega$ alors

$$P(A) = \sum_{k=1}^{\infty} P(B_k)P(A|B_k)$$

Un corrolaire simple de la proposition est le théorème de Bayes.

Proposition 3 - *Théorème de Bayes*
Soient $B_1, B_2, \cdots$ ensembles disjoints tels que $P(B_k) > 0$ pour tout k et $\cup_{k=1}^{\infty} B_k = \Omega$. Alors pour tout évènement A,

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{\sum_{k=1}^{\infty} P(B_k)P(A|B_k)}.$$

Indépendance

Définition 3 *On dit que les évènements A et B sont indépendants si $P(A \cap B) = P(A)P(B)$.*

2.2 Variables aléatoires

Définition 4 *Une variable aléatoire X est une fonction de Ω dans $\mathbb{R}$; quelque soit $\omega \in \Omega$, $X(\omega)$ est un réel.*

Soit X une variable aléatoire définie sur un espace Ω. Si on définit l'évènement

$$[a \leq X \leq b] = \{\omega \in \Omega : a \leq X(\omega) \leq b\} = A$$

alors $P(a \leq X \leq b) = P(A)$.

Définition 5 *Soit X une variable aléatoire définie sur un espace Ω. La fonction de répartition de X est définie par*

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

La fonction de répartition [1] satisfait les propriétés suivantes

1. cumulative distribution function (cdf)

- Si $x \leq y$ alors $F(x) \leq F(y)$: fonction croissante.
- Si $y \downarrow x$, alors $F(y) \downarrow F(x)$: fonction continue à droite.
- $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$

Soit X une variable aléatoire de fonction de répartition F alors on a
- $P(a \leq X \leq B) = F(b) - F(a)$
- $P(X > a) = 1 - F(a)$

2.2.1 V.a. discrètes

Définition 6 *On dit qu'une variable aléatoire X est discrète si son ensemble de définition est fini ou dénombrable. C'est à dire qu'il existe une ensemble $S = \{s_1, s_2, \cdots\}$ tel que $P(X \in S) = 1$.*

On peut définir la fonction de fréquence (ou densité) d'une variable aléatoire discrète : $f(x) = P(X = x)$. On a alors pour la fonction de répartition

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t).$$

Exemples
- Loi de Bernoulli
  $\Omega = \{0, 1\}$, $P(X = 1) = \pi$
- Loi binomiale : n tirages d'une bernoulli
  $f(x) = P(X = x) = C_n^k \pi^x (1 - \pi)^{n-x}$ pour $x = 0, 1, \cdots, n$. On note $X \sim B(n, \pi)$.
- Loi de Poisson
  $f(x) = \frac{\exp(-\lambda)\lambda^x}{x!}$ pour $x = 0, 1, 2, \cdots$. On note $X \sim Pois(\lambda)$

2.2.2 V.a. continues

Définition 7 *On dit qu'une varaible aléatoire X est continue si sa fonction de répartition $F(x)$ est continue pour tout réel x.*

Si X est une variable aléatoire continue $P(X = x) = 0$. On ne peut donc pas définir de fonction de fréquence.

Définition 8 *Une variable aléatoire continue admet une densité $f(x) \geq 0$ si pour $-\infty < a < b < \infty$, on a*

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

Exemples
- Loi exponentielle de paramètre λ

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) \text{ si } x \geq 0 \\ 0 \text{ sinon} \end{cases}$$

- Loi uniforme sur $[0, \theta]$

$$f(x) = \begin{cases} \frac{1}{\theta} \text{ si } x \in [0, \theta] \\ 0 \text{ sinon} \end{cases}$$

Théorème 1 *Soit X une v.a. continue de fonction de répartition F et soit $U = F(X)$. Alors U suit une loi uniforme sur $[0, 1]$.*

– Loi de Gauss (ou loi normale) de paramètres μ et σ^2 ($X \sim \mathcal{N}(\mu, \sigma^2)$)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

– Loi du χ^2 à k degrés de liberté
Proposition 4 *Soient $X_1, \ldots, X_k$, k variables aléatoires indépendantes de même loi normale de moyennes respectives μ_i et d'écart-type σ_i ; $Y_i = \frac{X_i - \mu_i}{\sigma_i}$ leurs variables centrées et réduites, alors par définition la variable X, telle que*

$$X = \sum_{i=1}^{k} Y_i^2 = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

suit une loi du χ^2 à k degrés de liberté.

– Loi de Student à k degrés de liberté
Proposition 5 *Soit Z une variable aléatoire de loi normale centrée et réduite et soit U une variable indépendante de Z et distribuée suivant la loi du χ^2 à k degrés de liberté. Par définition la variable*

$$T = \frac{Z}{\sqrt{U/k}}$$

suit une loi de Student à k degrés de liberté.

– Loi de Fisher à (k_1, k_2) degrés de liberté
Proposition 6 *Soient U_1 et U_2 des variables indépendantes distribuées respectivement selon des lois du χ^2 à k_1 et k_2 degrés de liberté. Alors*

$$Z = \frac{U_1/k_1}{U_2/k_2}$$

suit une loi de Fisher à (k_1, k_2) degrés de liberté.

2.2.3 Espérances

Soit X une variable aléatoire. L'espérance de X permet de caractériser (au moins partiellement) la distribution de X. L'espérance est le "centre de masse" de la distribution.

Si X est une variable aléatoire discrête de fonction de fréquence f, l'espérance (ou moyenne) de X, notée $E(X)$ est donnée par

$$E(X) = \sum_x x f(x)$$

sous réserve qu'au moins l'une des deux quantités

$$E(X^+) = \sum_{x>0} x f(x) \text{ et } E(X^-) = -\sum_{x<0} x f(x)$$

est finie.

Si X est une variable aléatoire continue de densité $f(x)$,

$$E(X) = \int_{-\infty}^{\infty} tf(t)dt$$

et plus généralement, si g est une fonction intégrable

$$E(g(X)) = \int_{-\infty}^{\infty} g(t)f(t)dt$$

Définition 9 - *Soit X unde variable aléatoire telle que $\mu = E(X)$. Alors on définie la variance de X, notée $Var(X)$ par*

$$Var(X) = E\left[(X-\mu)^2\right].$$

On remarque que
- $Var(X) = E(X^2) - \mu^2$
- $Var(aX+b) = a^2 Var(X)$

En notant $\sigma^2 = Var(X)$, on définit l'écart-type de X par σ.

2.2.4 Vecteurs aléatoires et distributions jointes

Soient $X_1, \cdots, X_k$, k variables aléatoires. On dit que $\mathbf{X} = (X_1, \cdots, X_k)$ est un vecteur aléatoire.

Définition 10 - *La distribution jointe d'un vecteur aléatoire $(X_1, \cdots, X_k)$ est*

$$F(x_1, \cdots, x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_k \leq x_k)$$

où l'évènement $[X_1 \leq x_1, X_2 \leq x_2, \cdots, X_k \leq x_k]$ est l'intersection de $[X_1 \leq x_1]$, $[X_2 \leq x_2]$, $\cdots$, $[X_k \leq x_k]$.

Définition 11 - *Soient $X_1, \cdots, X_k$ des variables aléatoires. On dit que $X_1, \cdots, X_k$ sont indépendantes si les évènements $[a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \cdots, a_k \leq X_k \leq b_k]$ sont indépendants quelque soient $a_i < b_i$, $i = 1, \cdots, k$.*

Si $(X_1, \cdots, X_k)$ admet une densité jointe ou une fonction de fréquence jointe, alors on a une condition équivalente pour l'indépendance.

Théorème 2 *Si $(X_1, \cdots, X_k)$ sont indépendantes et admettent une densité (ou fonction de fréquence) jointe $f(x_1 \cdots, x_k)$ alors*

$$f(x_1 \cdots, x_k) = \prod_{i=1}^{k} f_i(x_i)$$

où $f_i(x_i)$ est la densité (ou fonction de fréqunce) marginale de X_i. Réciproquement si on a $f(x_1 \cdots, x_k) = \prod_{i=1}^{k} f_i(x_i)$, alors $(X_1, \cdots, X_k)$ sont indépendantes.

L'indépendance est une hypothèse importante dans de nombreux modèles statistique. implicitement, elle signifie qu'on peut se concentrer sur la connaissance des distributions marginales.

Définition 12 *Soient X et Y des variables aléatoires telles que $E(X^2) < \infty$ et $E(Y^2) < \infty$. Soient $\mu_X = E(X)$ et $\mu_Y = E(Y)$. La covariance entre X et Y est définie par*

$$Cov(X,Y) = E\left[(X-\mu_x)(Y-\mu_Y)\right] = E(XY) - \mu_X\mu_Y$$

et la corrélation par

$$Corr(X,Y) = \frac{Cov(X,Y)}{[Var(X)Var(Y)]^{1/2}}$$

On remarque facilement que

- Pour toutes constantes a, b, c et d,

$$Cov(aX+b, cY+d) = acCov(X,Y)$$

- Si X et Y sont des variables indépendantes d'espérance finie alors $Cov(X,Y) = 0$. La réciproque n'est pas vraie.
- $-1 \leq Corr(X,Y) \leq 1$

2.3 Théorèmes de convergence

2.3.1 Convergence en probabilité et en distribution

Nous considérons deux types de convergence.

Définition 13 - *Soient $\{X_n\}$, X des variables aléatoires. Alors $\{X_n\}$ converge en probabilité vers X quand n tend vers l'infini ($X_n \to_p X$) si pour tout $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0.$$

Définition 14 - *Soient $\{X_n\}$, X des variables aléatoires. Alors $\{X_n\}$ converge en distribution vers X quand n tend vers l'infini ($X_n \to_d X$) si pour tout $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(X_n \geq x\right) = P(X \geq x) = F(x)$$

en tout point de continuité de $F(x)$.

Convergence en probabilité : convergence des variables, convergence en distribution : convergence des lois.

Théorème 3 *Soient $\{X_n\}$, X des variables aléatoires.*

(a) Si $X_n \to_p X$ alors $X_n \to_d X$.

(b) Si $X_n \to_d \theta$ (θ une constante) alors $X_n \to_p \theta$.

2.3.2 Loi faible des grands nombres

La loi des grands nombres donne la convergence de la moyenne d'un échantillon vers la moyenne de la population quand la taille de l'échantillon augmente.

Théorème 4 - Loi faible des grands nombres
Soient X_1,X_2,$\cdots$, des variables aléatoires indépendantes et de même loi telles que $E(X_i) = \mu$ et $E(|X_i|) < \infty$. Alors

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow_p \mu$$

quand $n \rightarrow \infty$.

En statistique, on utilise aussi l'**inégalité de Markov**

$$\forall \lambda > 0 : P(X \geq \lambda) \leq \frac{E(X)}{\lambda}$$

avec $E(X) \leq \lambda$.

2.3.3 Théorème de Limite Centrale (TCL)

En probabilité, le TCL établit les conditions sous lesquelles la distribution d'une somme de variables aléatoire peut être approché par une loi normale.

Théorème 5 - TCL pour les variables aléatoire indépendantes et identiquement distribuées
Soient X_1,X_2,$\cdots$, des variables aléatoires indépendantes et de même loi de moyenne μ et de variance $\sigma^2 < \infty$. Définissons

$$S_n = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

Alors $S_n \rightarrow_d Z \sim \mathcal{N}(0,1)$ quand $n \rightarrow \infty$.

Théorème 6 - Inégalité de Chebychev
Soit X une variable aléatoire telle que $E(X^2) < \infty$. Alors, pour tout $\epsilon > 0$,

$$P\left(\|X - E(X)\| > \epsilon\right) \leq \frac{Var(X)}{\epsilon^2}.$$

Chapitre 3

Principes de l'estimation ponctuelle

3.1 Modèle statistique

3.1.1 Définitions

Soient $X_1, \cdots, X_n$ des variables aléatoires définies sur $(\Omega, \mathcal{A}, P)$ dans un espace mesurable $(E_n, \mathcal{E}_n)$. On suppose qu'on observe $x_1, \cdots, x_n$ qui sont telles que pour tout i, x_i est une *réalisation* (un tirage) de la variable aléatoire X_i. Supposons que la distribution jointe de $\mathbf{X} = (X_1, \cdots, X_n)$ est inconnue mais qu'elle appartient à une famille particulière de distributions. Le couple formé par l'espace d'observation E_n et cette famille de distributions $\mathcal{P}_n$ est appelé *modèle statistique*. On note $(E_n, \mathcal{P}_n)$.

Remarque - Bien qu'on suppose ici que $\mathbf{X}$ est observée, on peut parler de modèle statistique pour $\mathbf{X}$ même si certaines variables X_i ne sont pas observables.

En général, les distributions appartenant à un modèle statistique sont indéxées par un paramètre $\theta \in \Theta.$; θ représente typiquement la partie inconnue ou non spécifiée du modèle. On peut alors écrire

$$\mathbf{X} = (X_1, \cdots, X_n) \sim F_\theta \text{ pour } \theta \in \Theta$$

où F_θ est la distribution jointe de $\mathbf{X}$ et Θ l'ensemble des toutes les valeurs possibles pour θ. Le plus souvent, $\Theta \subset \mathbb{R}^p$. Si $p > 1$, $\theta = (\theta_1, \cdots, \theta_p)$ est un vecteur de paramètres. Le modèle statistique est ici $(E_n, F_\theta)_{\{\theta \in \Theta\}}$. On notera parfois abusivement $P_\theta(A)$, $E_\theta(X)$ et $Var_\theta(X)$ pour les probabilités, espérance et variance qui dépendent de θ.

Si les variables $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées (i.i.d.), on a

$$P_{(X_1, \cdots, X_n)} = P_{X_1} \times \cdots \times P_{X_n} = P_{X_1}^n$$

Et on note alors le modèle statistique $(E, P_{X_1})^n$ où E est l'espace de définition de X_1.

Quand les distributions d'un modèle peuvent être indéxées par un paramètre de dimension finie, on parlera de *modèle paramétrique*. Il existe des cas où le paramètre est de dimension infinie. On parle alors, à tort, de *modèle non paramétrique*. Dans la suite du cours non nous intéressons principalement aux modèles paramétriques. On notera $f(x, \theta)$ la densité de P_θ relativement à une mesure dominante et σ-finie, μ. On va se restreindre au cas où

– au cas où μ est la mesure de Lebesgue (variables aléatoires de loi absolument continue) et on retrouve la densitée $f_\theta(x)$ ou,
– au cas où μ est la mesure de comptage (variables aléatoires de loi discrète) et on retrouve le systême $P_\theta(X = x)$. On note $\mathbf{X}$ l'échantillon $(X_1, ..., X_n)$ issu du même modèle (E^n, P_θ).

Exemple - Dans l'exemple des votes, $X_1, \cdots, X_n$ sont des variables indépendantes de même loi de Bernouilli de paramètre $\pi \in [0, 1]$. Chaque variable X_i est définie sur $E = \{0, 1\}$.

Exemple - Supposons que $X_1, \cdots, X_n$ sont des variables aléatoires indépendantes et identiquement distribuées suivant une loi de Poisson de moyenne λ.
La probabilité jointe de $\mathbf{X} = (X_1, \cdots, X_n)$ a pour densité

$$f(\mathbf{x}; \lambda) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!}$$

pour une réalisation $\mathbf{x} = (x_1, \cdots, x_n)$ de $\mathbf{X}$. L'espace des paramètres pour ce modèle est l'ensemble $\{\lambda : \lambda > 0\}$.

3.1.2 Identifiabilité

Pour un modèle statistique donné, un paramètre donné θ correspond à une unique distribution F_θ. Cependant, il peut exister des valeurs distinctes du paramètre, θ_1 et θ_2 telles que $F_{\theta_1} = F_{\theta_2}$. Pour éviter cette difficulté, on requiert qu'un modèle, ou plus précisement sa paramétrisation, soit *identifiable*. En effet une paramétrisation non identifiable pose souvent des problèmes d'estimation.

On dit qu'un modèle a une paramétrisation identifiable si $F_{\theta_1} = F_{\theta_2}$ implique que $\theta_1 = \theta_2$.

Exemple - Supposons que $X_1, \cdots, X_n$ sont des variables aléatoires indépendantes gaussiennes avec

$$E(X_i) = \beta_0 + \beta_1 t_i + \beta_2 s_i$$

où $t_1, \cdots, t_n$ et $s_1, \cdots, s_n$ sont des constantes connues, et $Var(X_i) = \sigma^2$. L'espace des paramètres est

$$\{(\beta_0, \beta_1, \beta_2, \sigma) : -\infty < \beta_k < \infty, \sigma > 0\}$$

La paramétrisation de ce modèle est identifiabe si et seulement si les vecteurs

$$z_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \; z_1 = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} \text{ et } z_2 = \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix}$$

sont linéairement indépendants, c'est à dire que $a_0 z_0 + a_1 z_1 + a_1 z_2 = 0$ implique que $a_0 = a_1 = a_2 = 0$. Notons

$$\mu = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix}$$

et remarquons que la paramétrisation est identifiable si il y a une bijection entre les valeurs possibles de μ et les paramètres $\beta_0, \beta_1, \beta_2$. Suposons maintenant que z_0, z_1 et z_2 sont linéairement dépendants ; alors on peut avoir $a_0 z_0 + a_1 z_1 + a_2 z_2 = 0$ avec au moins un des coefficients a_0, a_1 ou a_2 non nul. Dans ce cas, nous aurons

$$\begin{aligned} \mu &= \beta_0 z_0 + \beta_1 z_1 + \beta_2 z_2 \\ &= (\beta_0 + a_0) z_0 + (\beta_1 + a_1) z_1 + (\beta_2 + a_2) z_2 \end{aligned}$$

et par conséquent il n'y a pas bijection entre μ et $(\beta_0, \beta_1, \beta_2)$. Cependant, quand z_0, z_1 et z_2 sont linéairement dépendants, il est possible d'obtenir une paramétrisation identifiable en restreignant l'espace des paramètres ; ceci est réalisé en contraignant les paramètres $\beta_0, \beta_1, \beta_2$.

Dans la suite, nous supposerons implicitement que les modèles statistiques sont identifiables ; sauf mention contraire.

3.1.3 Familles exponentielles

Une classe importante de modèles statistiques est la classe des modèles de la famile exponentielle.

Définition 15 *Supposons que $X_1, \cdots, X_n$ a une distribution jointe F_θ avec $\theta = (\theta_1, \cdots, \theta_p)$ un paramètre (inconnu). On dit que la famille de distribution $\{F_\theta\}$ est une famille exponentielle à k paramètres si la densité de probabilité jointe de $(X_1, \cdots, X_n)$ est de la forme*

$$f(\mathbf{x}; \theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta) T_i(\mathbf{x}) - d(\theta) + S(\mathbf{x})\right]$$

pour $\mathbf{x} \in A$ avec A un ensemble qui ne dépend pas de θ.

Remarque - k n'est pas forcément égal à p, même si c'est souvent le cas.

Exemple - Loi binomiale.
Supposons que X suit une loi binomiale de paramètres n et θ avec θ inconnu. Alors

$$\begin{aligned} f(x, \theta) &= \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{1-x} \\ &= \exp\left[\ln\left(\frac{\theta}{1-\theta}\right) x + n \ln(1-\theta) + \ln\left(\begin{array}{c} n \\ x \end{array} \right)\right] \end{aligned}$$

pour $x \in A = \{0, 1, \cdots, n\}$ et ainsi la distribution de X est dans une famille exponentielle à 1 paramètre.

Exemple - Loi de Gauss.
Supposons que $X_1, \cdots, X_n$ sont des variables aléatoires normales i.i.d. de moyenne θ et de variance θ^2 avec $\theta > 0$. La densité de probabilité jointe de $(X_1, \cdots, X_n)$ est

$$\begin{aligned} f(\mathbf{x}; \theta) &= \prod_{i=1}^{n} \left[\frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{1}{2\theta^2}(x_i - \theta)^2\right)\right] \\ &= \exp\left[-\frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 + \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{n}{2}(1 + \ln(\theta^2)) + \ln(2\pi)\right] \end{aligned}$$

et on a $A = \mathbb{R}^n$. On conclut que c'est une loi de la famille exponentielle à 2 paramètres en dépit du fait que l'espace des paramètres est de dimension 1.

Proposition 7 *Si $X_1, \cdots, X_n$ sont i.i.d. de loi appartenant à la famlle exponentielle alors la loi jointe du vecteur $(X_1, \cdots, X_n)$ appartient à la famille exponentielle.*

Preuve - La densité de $(X_1, \cdots, X_n)$ est

$$\begin{aligned} f(\mathbf{x};\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \text{ car les } x_i \text{ sont i.i.d.} \\ &= \prod_{i=1}^{n} \exp\left[\sum_{j=1}^{k} c_j(\theta)T_j(x_i) - d(\theta) + S(x_i)\right] \\ &= \exp\left[\sum_{j=1}^{k} c_j(\theta)\sum_{i=1}^{n} T_j(x_i) - nd(\theta) + \sum_{i=1}^{n} S(x_i)\right] \\ &= \exp\left[\sum_{j=1}^{k} c_j(\theta)\tilde{T}_j(\mathbf{x}) - nd(\theta) + \tilde{S}(\mathbf{x})\right] \end{aligned}$$

□

Proposition 8 *Supposons que $\mathbf{X} = (X_1, \cdots, X_n)$ est distribué suivant une loi d'une famille exponentielle à un paramètre avec une densité de la forme*

$$f(\mathbf{x};\theta) = \exp\left[c(\theta)T(\mathbf{x}) - d(\theta) + S(\mathbf{x})\right]$$

pour $\mathbf{x} \in A$ où
(a) l'espace Θ du paramètre est ouvert,
(b) $c(\theta)$ est une fonction bijective sur Θ
(c) $c(\theta)$ et $d(\theta)$ sont deux fois différentiables sur Θ.
Alors

$$E_\theta[T(\mathbf{X})] = \frac{d'(\theta)}{c'(\theta)}$$

$$Var_\theta[T(\mathbf{X})] = \frac{d''(\theta)c'(\theta) - d'(\theta)c''(\theta)}{(c'(\theta)^3}$$

Preuve - Par définition f est une densité, on a donc

$$\int \exp\left[c(\theta)T(\mathbf{x}) - d(\theta) + S(\mathbf{x})\right] dx = 1$$

soit encore

$$\int \exp\left[c(\theta)T(\mathbf{x}) + S(\mathbf{x})\right] dx = exp(d(\theta))$$

en dérivant par θ les termes de gauche et de droite, on obtient

$$\int c'(\theta)T(\mathbf{x})\exp\left[c(\theta)T(\mathbf{x})+S(\mathbf{x})\right]dx = d'(\theta)exp(d(\theta))$$

On peut bien entrer le signe dérivée sous l'intégrale car $T(\mathbf{x})\exp\left[c(\theta)T(\mathbf{x})+S(\mathbf{x})\right]$ est intégrable par rapport à la mesure dominante. d'où

$$c'(\theta)\int T(\mathbf{x})\exp\left[c(\theta)T(\mathbf{x})-d(\theta)+S(\mathbf{x})\right]dx = d'(\theta)$$

soit

$$E_\theta[T(\mathbf{X})] = \frac{d'(\theta)}{c'(\theta)}$$

et qui conclut la preuve. □

3.1.4 Statistiques

Supposons que le modèle statistique pour $\mathbf{X} = (X_1, \cdots, X_n)$ a un espace de paramètres Θ. Comme le paramètre θ est inconnu, nous cherchons à extraire de l'information le concernant dans $\mathbf{X}$, sans perdre trop d'information.

Définition 16 *Une statistique est une fonction $T : E^n \to \mathbb{R}^p$ qui à $\mathbf{X}$ associe $T(\mathbf{X})$ et qui ne dépend d'aucun paramètre inconnu ; autrement dit, T ne dépend que de variables aléatoires observables et de constantes connues.*

Une statistique peut être scalaire ou vectorielle.

Exemple - Moyenne empirique :

$$T(\mathbf{X}) = \bar{X} = \frac{1}{n}X_i$$

La taille n est connue et T est bien une statistique.

Il est important de bien comprendre qu'une statistique est elle même une variable aléatoire et qu'elle a sa propre loi de probabilité ; cette distribution peut éventuellement dépendre de θ.

3.1.5 Exhaustivité

Afin de faire de l'inférence statistique, le statisticien va devoir extraire de l'information de la suite de variables aléatoires $X_1, ..., X_n$ dont il dispose. Lorsque la taille de l'échantillon n est grande, il est naturel de tenter de réduire l'échantillon et et de résumer l'information qui y est contenue. Lorsque il est possible de "remplacer" $(X_1, ..., X_n)$ par une statistique $T = T(X_1, ..., X_n)$, on optera bien sûr pour cette solution. Cependant, une question se pose : Comment savoir si la réduction des données opérée par la statistique T ne conduit pas à une perte d'information ? C'est ce type de problèmes que cherche à résoudre la notion d'exhaustivité.

L'idée est basée sur la remarque suivante : si la loi conditionnelle de $\mathbf{X}$ sachant T ne dépend pas de la loi P_θ de $\mathbf{X}$, alors T est suffisamment informative pour P_θ. En effet, dans cette situation la loi conditionnelle de $\mathbf{X}$ sachant S peut être spécifiée indépendamment de P_θ, lorsqu'on donne $S = s$, on peut générer une variable aléatoire X' de même loi que X. Donc les informations données par $\mathbf{X} = (X_1, \cdots, X_n)$ ne donnent pas plus sur P_θ que T ne le fait. T est dite alors statistique exhaustive (ou suffisante).

Définition 17 - *La statistique T sera dite exhaustive pour θ si la loi conditionnelle de $\mathbf{X}$ sachant $T(\mathbf{X}) = t$ n'est pas une fonction du paramètre θ :*

$$P_\theta(\mathbf{X}|T(\mathbf{X}) = t)$$

ne dépend pas de θ.

Exemple - Supposons que $X_1, \cdots, X_k$ soient des variables indépendantes de loi binomiale de paramètres n_i connus et θ (inconnu). Soit $T = X_1 + \cdots + X_k$; T a aussi une loi binomiale de paramètres $m = n_1 + \cdots + n_k$ et θ. Pour montrer que T est une statistique exhaustive pour θ nous devons montrer que

$$P_\theta[\mathbf{X} = \mathbf{x}|T = t]$$

est indépendant de θ pour tout t et tout $x_1, \cdots, x_k$. Tout d'abord notons que si $t \neq x_1 + \cdots + x_k$ alors sa probabilité conditionnelle est 0. Si $t = x_1 + \cdots + x_k$ alors

$$\begin{aligned} P_\theta[\mathbf{X} = \mathbf{x}|T = t] &= \frac{P_\theta[\mathbf{X} = \mathbf{x}]}{P_\theta[T = t]} \\ &= \frac{\prod_{i=1}^k \binom{n_i}{x_i} \theta^{x_i}(1-\theta)^{(n_i - x_i)}}{\binom{m}{t} \theta^t (1-\theta)^{(m-t)}} \\ &= \frac{\prod_{i=1}^k \binom{n_i}{x_i}}{\binom{m}{t}} \end{aligned}$$

ce qui est indépendant de θ. Ainsi T est une statistique exhaustive pour θ.

Le problème est que, dans la plupart des lois (en particulier les lois continues), il est difficile d'utiliser directement la définition pour montrer qu'une statistique est exhaustive. De plus, cette définition ne permet pas d'identifier ou de construire des statistiques exhaustives. Mais il existe un critère simple de Jerzy Neyman qui donne une condition nécessaire et suffisante pour que T soit une statistique exhaustive quand la loi de $\mathbf{X}$ admet une densité.

Théorème 7 *(Théorème de factorisation) - Supposons que $\mathbf{X} = (X_1, \cdots, X_n)$ admet une densité jointe $f(\mathbf{x}; \theta)$ pour $\theta \in \Theta$. Alors, $T = T(\mathbf{X})$ est une statistique exhaustive pour θ si et seulement*

s'il existe deux fonctions mesurables $g : \mathbb{R}^p \times \Theta \to \mathbb{R}^+$ *et* $h : E \to \mathbb{R}^+$ *telles que* $f(x,;\theta)$ *se met sous la forme*

$$f(\mathbf{x};\theta) = h(x)g(T(x),\theta)$$

(T *et* θ *peuvent être des vecteurs).*

La preuve rigoureuse de ce théorème est difficile dans le cas des variables continues.
Faire le schéma de la preuve ?.

Exemple - Supposons que $X_1, \cdots, X_n$ sont des variables i.i.d. avec pour densité

$$f(x;\theta) = \frac{1}{\theta} \text{ pour } 0 \leq x \leq \infty$$

avec $\theta > 0$. La densité jointe de $\mathbf{X} = (X_1, \cdots, X_n)$ est

$$\begin{aligned} f(\mathbf{x};\theta) &= \frac{1}{\theta^n} \text{ pour } 0 \leq x_1, \cdots, x_n \leq \infty \\ &= \frac{1}{\theta^n} I(0 \leq x_1, \cdots, x_n \leq \infty \\ &= \frac{1}{\theta^n} I\left(\max_{1\leq i\leq n} x_i \leq \theta\right) I\left(\min_{1\leq i\leq n} x_i \geq 0\right) \\ &= g(\max_i x_i;\theta)h(\mathbf{x}) \end{aligned}$$

et $X_{(n)} = \max_{i=1,\cdots,n}(X_i)$ est une statistique exhaustive pour θ.

Exemple - Supposons que $X_1, \cdots, X_n$ sont des variables i.i.d. issues d'une loi exponentielle à k paramètres de densité

$$f(\mathbf{x};\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta)T_i(\mathbf{x}) - d(\theta) + S(\mathbf{x})\right] I(\mathbf{x} \in A)$$

en prenant $h(\mathbf{x}) = \exp[S(\mathbf{x})]I(\mathbf{x} \in A)$, on obtient par le théorème de factorisation que, $T = (T_1(\mathbf{X}), \cdots, T_k(\mathbf{X}))$ est une statistique exhaustive pour θ.

3.2 Estimation ponctuelle

Un *estimateur* est une statistique qui a pour vocation d'estimer la vraie valeur d'un paramètre θ. Ainsi si

$$\mathbf{X} \sim F_\theta \text{ pour } \theta \in \Theta$$

alors un estimateur $\hat{\theta}$ est égal à une statistique $T(\mathbf{X})$.

Supposons que θ est un paramètre réel et que $\hat{\theta}$ est un estimateur de θ. La distribution de l'estimateur $\hat{\theta}$ est souvent appelée, distribution empirique de $\hat{\theta}$. Idéalement, nous souhaitons que la distribution empirique de $\hat{\theta}$ soit centrée sur θ et de variance faible. Plusieurs mesures de qualité d'un estimateur sont basées sur sa distribution.

3.2.1 Biais et erreurs en moyenne quadratique

Définition 18 - *Le biais*[1] *d'un estimateur $\hat{\theta}$ est défini par*

$$b_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$$

On dit qu'un estimateur est non biaisé ou sans biais si $b_\theta(\hat{\theta}) = 0$.

Définition 19 - *L'erreur en moyenne absolue*[2] *(EMA) de θ est définie par*

$$EMA = E_\theta(|\hat{\theta} - \theta|)$$

Définition 20 - *L'erreur en moyenne quadratique*[3] *(EMQ) de θ est définie par*

$$EMQ = E_\theta((\hat{\theta} - \theta)^2)$$

Le biais indique si la distribution empirique de l'estimateur $\hat{\theta}$ est centrée ou non sur la vraie valeur du paramètre θ tandis que l'erreur en moyenne absolue et l'erreur en moyene quadratique donnent des informations sur la dispersion de la distribution autour de θ. Les erreurs EMA et EMQ sont de bonnes mesures pour comparer plusieurs estimateurs d'un paramètre θ. On utilise plus souvent l'EMQ que l'EMA en particulier parce que l'EMQ se décompose en une somme du biais au carré et de la variance de l'estimateur :

$$\text{EMQ} = Var_\theta(\hat{\theta}) + b_\theta(\hat{\theta})^2$$

Preuve à faire en TD.

Cette décomposition permet notamment de calculer ou d'approcher facilement l'EMQ alors que l'EMA est plus difficile à calculer. Notamment, si la variance de l'estimateur est bien plus grande que le biais alors $EQM \simeq Var_\theta(\theta)$.

Exemple - Supposons que $X_1, \cdots, X_n$ sont des variables i.i.d. de loi uniforme sur $[0, \theta]$. Et choisissons $\hat{\theta} = \max_{i=1,\cdots,n} X_i$. On montre (voir TD) que la densité de $\hat{\theta}$ est

$$f(x;\theta) = \frac{n}{\theta^n} x^{n-1} \text{ pour } 0 \leq x \leq \theta$$

On obtient alors

$$E_\theta(\hat{\theta}) = \frac{1}{n+1}\theta$$

D'où on déduit facilement un estimateur sans biais de θ :

$$\tilde{\theta} = \frac{n+1}{n} \max_{i=1,\cdots,n} X_i$$

1. bias
2. mean absolute error
3. mean square error

3.2.2 Consistance

Supposons que $\hat{\theta}_n$ soit un estimateur d'un paramètre θ construit à partir de n variables aléaloires $X_1, \cdots, X_n$. Quand n croit, il est raisonnable d'attendre qur la loi empirique de $\hat{\theta}$ soit de plus en plus concentrée autour de la vraie valeur du paramètre θ. Cette propriété de la suite $\{\hat{\theta}_n\}$ est la consistance.

Définition 21 - *Une suite d'estimateurs $\{\hat{\theta}_n\}$ est dite consistante pour θ si $\{\hat{\theta}_n\}$ converge en probabilité vers θ, c'est à dire si,*

$$\lim_{n \to +\infty} P_\theta[|\hat{\theta}_n - \theta| > \epsilon] = 0$$

pour tout $\epsilon > 0$ et pour toute valeur de θ.

On dit abusivement que *" $\hat{\theta}_n$ est un estimateur consistant de θ "*.

Exemple - Supposons que $X_1, \cdots, X_n$ sont des variables i.i.d. de moyenne μ, alors $\bar{X}_n$ est un estimateur consistant de μ. En effet par la loi forte des grands nombres, on a que $\bar{X}_n$ tend presque sûrement vers μ et on en déduit la consistance.

Chapitre 4

Estimation par maximum de vraisemblance

Un des estimateurs les plus utilisés en statistique est l'estimateur du maximum de vraisemblance. La vraisemblance est une fonction qui contient toute l'information des données sur un paramètre inconnu. Elle joue un rôle important dans de nombreuses méthodes statistiques. Et l'estimateur du maximum de vraisemblance a de très bonnes propriétés d'optimalité.

4.1 La vraisemblance

Soient $X_1, \cdots, X_n$ des variables aléatoires de densité $f(\mathbf{x}; \theta)$ avec $\theta \in \Theta$. Et considérons $\mathbf{x} = (x_1, \cdots, x_n)$ une réalisation du vecteur $\mathbf{X} = (X_1, \cdots, X_n)$. La fonction de vraisemblance [1] est définie par

$$\mathcal{L}(\theta) = f(\mathbf{x}; \theta)$$

C'est une fonction réelle définie sur l'espace des paramètres Θ. La vraisemblance donne, en quelque sorte, la probabilité que la réalisation $\mathbf{x} = (x_1, \cdots, x_n)$ soit émise par le modèle associé à la valeur θ du paramètre. Ainsi plus il est probable que le modèle émette cette réalisation pour la valeur θ, plus la vraisemblance sera grande.

Définition 22 - *Soit* $\mathbf{X} = (X_1, \cdots, X_n)$ *un vecteur aléatoire suivant le modèle* F_θ, $\theta \in \Theta$. *Pour une réalisation* $\mathbf{x} = (x_1, \cdots, x_n)$, *l'estimateur du maximum de vraisemblance (EMV) est l'estimateur* $\hat{\theta} = S((X))$ *avec* S *telle que*

$$\mathcal{L}(S(x)) \geq \mathcal{L}(\theta) \text{ pour tout } \theta \in \Theta$$

Exemple pour une loi discrête - Supposons que $X_1, \cdots, X_n$ sont des variables aléatoires i.i.d de loi de Bernouilli de paramètre inconnu $\theta \in]0, 1[$. Et notons $f(.; \theta)$ la densité de X_i pour tout $i = 1, \cdots, n$. On a

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}$$

1. likelihood function

Soit $(x_1, \cdots, x_n)$ une réalisation de $(X_1, \cdots, X_n)$. Par définition, la vraisemblance est

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

car les variables X_i sont indépendantes. D'où

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

On en déduit la log-vraisemblance :

$$\ln \mathcal{L}(\theta; \mathbf{x}) = \sum_{i=1}^{n} x_i \ln\theta + (n - \sum_{i=1}^{n} x_i)\ln(1-\theta)$$

Le maximum est la racine de la dérivée de cette expression dont la dérivée seconde est négative.

$$\frac{\partial \ln \mathcal{L}(\theta; \mathbf{x})}{\partial\theta} = \frac{\sum_{i=1}^n x_i}{\theta} + \frac{n - \sum_{i=1}^n x_i}{1-\theta}$$

s'annule en $\theta^* = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}_n$. On vérifie que

$$\frac{\partial^2 \ln \mathcal{L}(\theta; \mathbf{x})}{\partial\theta^2} < 0$$

au voisinage de $\theta^* = \bar{x}_n$. Et on conclut que l'estimateur du maximum de vraisemblance pour le paramètre de la loi de Bernouilli est $\hat{\theta}_n = \bar{X}_n$.

Propriétés 1 - *Soit ϕ telle que $\phi = g(\theta)$ pour une fonction bijective g. Si $\hat{\theta}$ est un EMV pour θ alors $\hat{\phi} = g(\hat{\theta})$ est un EMV pour ϕ.*

4.2 Propriétés asymptotiques de l'EMV

Nous allons voir dans cette section, que sous des conditions faibles de régularité, on peut montrer que l'EMV est consistant et asymptotiquement normal. Dans la suite de cette section, nous supposons que $X_1, \cdots, X_n$ sont des v.a.i.i.d. de densité $f(x;\theta)$ avec $\theta \in \Theta \subset \mathbb{R}$ avec $\ell(x;\theta) = \ln f(x;\theta)$ trois fois différentiable par rapport à θ. Et nous ajoutons les hypothèses suivantes :

(A1) Θ est un sous ensemble ouvert de $\mathbb{R}$.
(A2) L'ensemble $A = \{x : f(x;\theta) > 0\}$ ne dépend pas de θ.
(A3) $f(x;\theta)$ est trois fois continûment différentiable par rapport à θ sur A.
(A4) $E_\theta[\ell'(X_i;\theta)] = 0$ pour tout θ et $Var_\theta[\ell'(X_i;\theta)] = I(\theta)$ où $0 < I(\theta) < \infty$ pour tout θ.
(A5) $E_\theta[\ell''(X_i;\theta)] = -J(\theta)$ où $0 < J(\theta) < \infty$ pour tout θ.
(A6) Pour θ et $\delta > 0$, $|\ell'''(x;t)| < M$ pour $|\theta - t| \leq \delta$ où $E_\theta[M(X_i)] < \infty$.

Définition 23 *Si le modèle (E, P_θ) vérifie les hypothèses (A1) à (A3) et que l'intégrale $\int_B f(x;\theta)d\mu(x)$ est au moins deux fois dérivable sous le signe d'intégration pour tout borélien B, alors on dit que le modèle est un modèle régulier.*

Exemples - Les modèles basés sur la loi exponentielle et sur la loi de Gauss sont réguliers mais pas celui basé sur la loi uniforme car dans ce dernier cas, le domaine de définition du modèle dépend du paramètre de la loi.

4.2.1 Convergence en loi

Proposition 9 - *Sous l'hypothèse (A2), on a* $I(\theta) = J(\theta)$ *soit* $Var_\theta[\ell'(X_i;\theta)] = -E_\theta[\ell''(X_i;\theta)]$. $I(\theta)$ *est appelée information de Fisher.*

Preuve - La condition (A2) implique

$$\int_A f(x;\theta)dx = 1 \text{ pour tout } \theta \in \Theta$$

Si on peut échanger le signe somme et la dérivée, on a

$$\begin{aligned} 0 &= \int_A \frac{\partial}{\partial\theta} f(x;\theta)dx \\ &= \int_A \ell'(x;\theta)f(x;\theta)dx \\ &= E_\theta[\ell'(X_i;\theta)] \end{aligned}$$

De plus, si on dérive deux fois sous le signe $\int$, on a

$$\begin{aligned} 0 &= \int_A \frac{\partial}{\partial\theta}(\ell'(x;\theta)f(x;\theta))dx \\ &= \int_A \ell''(x;\theta)f(x;\theta)dx + \int_A (\ell'(x;\theta))^2 f(x;\theta)dx \\ &= -J(\theta) + I(\theta) \end{aligned}$$

$\diamond$

Théorème 8 - *Sous les hypothèses (A1) à (A5), on a un théorème de limite centrale pour l'EMV* $\hat{\theta}_n$ *de* θ *:*

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_d \frac{Z}{J(\theta)} \sim \mathcal{N}(0, I(\theta)/J^2(\theta))$$

avec $\to_d$ *la convergence en loi.*

Lemme 1 - lemme de Slutsky - *Soit* $\{X_n\}_{n=0,\cdots,\infty}$ *une suite de variables aléatoires qui tend en loi vers* X *et soit* $\{Y_n\}_{n=0,\cdots,\infty}$ *une suite de variables aléatoires qui tend en probabilité vers une constante* $c \in \mathbb{R}$*, alors* $X_n + Y_n$ *tend en loi vers* $X + c$ *et* X_nY_n *tend en loi vers* cX*.*

Preuve du théorème - Sous les conditions (A1) à (A3), si $\hat{\theta}_n$ maximise la vraisemblance, on a

$$\sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = 0$$

et en écrivant un développement de Taylor de cette expression, on obtient

$$0 = \sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = \sum_{i=1}^{n} \ell'(X_i; \theta)$$
$$+(\hat{\theta}_n - \theta) \sum_{i=1}^{n} \ell''(X_i; \theta) + \frac{1}{2}(\hat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*)$$

où θ_n^* appartient à l'interval $[\min(\hat{\theta}_n, \theta), \max(\hat{\theta}_n, \theta)]$. En divisant les termes de gauche et de droite par $\sqrt{n}$, on peut encore écrire

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{-(1/\sqrt{n}) \sum_{i=1}^{n} \ell'(X_i; \theta)}{n^{-1} \sum_{i=1}^{n} \ell''(X_i; \theta) + (\hat{\theta}_n - \theta)(2n^{-1}) \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*)}$$

Par le théorème de limite centrale et la condition (A4), on obtient

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i; \theta) \to_d Z \sim \mathcal{N}(0, I(\theta)$$

et par la loi faible des grands nombres et la condition (A5), on a

$$\frac{1}{n} \sum_{i=1}^{n} \ell''(X_i; \theta) \to_P -J(\theta)$$

On conclut la preuve en utilisant le lemme de Slutsky. $\diamond$

Théorème 9 - Delta méthode - *Supposons que that*

$$a_n(X_n - \theta) \to_d Z$$

où θ est une contante et $\{a_n\}$ une suite de constantes telle que $a_n \to +\infty$. Si $g(x)$ est une fonction dérivable de dérivée $g'(\theta)$ en θ alors

$$a_n(g(X_n) - g(\theta)) \to_d g'(\theta)Z.$$

Preuve de la Delta méthode - Développement de Taylor + lemme de Slutsky. Notons d'abord que $X_n \to_P \theta$ (d'après le lemme de Slutsky). Si g est continuement différentiable en θ,

$$g(X_n) = g(\theta) + g(\theta_n^*)(X_n - \theta)$$

avec θ_n^* entre X_n et θ ; ainsi $|\theta_n^* - \theta| \leq |X_n - \theta|$ et $\theta_n^* \to_P \theta$. Comme $g'(x)$ est continue en θ, il suit que $g'(\theta_n^*) \to_P g(\theta)$. Maintenant,

$$a_n(g(X_n) - g(\theta)) = g'(\theta_n^*)a_n(X_n - \theta) \to_d g'(\theta)Z$$

par le lemme de Slutsky. $\diamond$

Estimation de l'écart-type des estimateurs
Dans les cas où $I(\theta) = J(\theta)$, le résultat du théorème 8 suggère que pour n assez grand, l'EMV $\hat{\theta}_n$ suit approximativement une loi de Gauss de moyenne θ et de variance $1/(nI(\theta))$. Ce résultat peut être utilisé pour approcher l'écart-type de $\hat{\theta}_n$ par $[nI(\theta)]^{-1/2}$. Comme $I(\theta)$ dépend de θ il est nécessaire de l'estimer pour obtenir une approximation de l'écart-type de $\hat{\theta}$. Il y a deux approches pour faire ceci :

- Si $I(\theta)$ une expression analytique, on peut subtituer $\hat{\theta}_n$ à θ dans cette expression. On obtient alors
$$\widehat{se}(\hat{\theta}_n) = \frac{1}{\sqrt{nI(\hat{\theta}_n)}}$$
$nI(\hat{\theta}_n)$ est appelée *information de Fisher attendue* pour θ.
- Comme $I(\theta) = -E[\ell''(X_i;\theta))]$, on peut estimer $I(\theta)$ par
$$\widehat{I(\theta)} = -\frac{1}{n}\sum_{i=1}^{n} \ell''(X_i;\hat{\theta}_n)$$
ce qui conduit à l'écart-type
$$\hat{se}(\hat{\theta}_n) = \frac{1}{\sqrt{n\widehat{I(\theta)}}} = \left(-\frac{1}{n}\sum_{i=1}^{n} \ell''(X_i;\hat{\theta}_n)\right)^{-1/2}$$
$n\widehat{I(\theta)}$ est appelée *information de Fisher observée* pour θ.

Exemple - Supposons que $X_1, \cdots, X_n$ sont des variables aléatoires i.i.d. de loi géométrique de densité de probabilité
$$f(x;\theta) = \theta(1-\theta)^x \text{ pour } \theta = 0, 1, \cdots$$
L'estimateur du maximum de vraisemblance de θ basé sur $X_1, \cdots, X_n$ est
$$\hat{\theta}_n = \frac{1}{\bar{X}_n + 1}$$
Par le théorème de limite centrale, on a $\sqrt{\bar{X}_n - (\theta^{-1} - 1)}$ tend en loi vers une v.a. Z de loi de Gauss de moyenne nulle et de variance $\theta^{-2}(1-\theta)$. Ainsi nous on obtient que
$$\begin{array}{rcl} \sqrt{n}(\hat{\theta}_n - \theta & = & \sqrt{n}(g(\bar{X}_n) - g(\theta^{-1}) - 1)) \\ & \to_d & (0, \theta^2(1-\theta)) \end{array}$$
en appliquant la Delta-méthode avec $g(s) = 1/(1+x)$ et $g'(x) = -1/(1+x)^2$.

4.2.2 Consistance

Il existe des résultats de consistance asymptotique pour le maximum de vraisemblance global, mais sous des hypothèses restrictives et difficilement vérifiables. On s'intéresse plutôt aux maxima locaux, plus précisément aux solutions de

$$\frac{\partial \ln \mathcal{L}(\theta; x_1, \cdots, x_n)}{\partial \theta} = 0$$

Théorème 10 - *Soient $X_1, \cdots, X_n$ des variables aléatoires i.i.d. suivant le modèle paramétrique (E^n, P_θ) avec $\theta \in \Theta$ ouvert. On suppose que le modèle est identifiable (i.e. $\theta \mapsto P_\theta$ est injective). Alors il existe une suite $\{\hat{\theta}_n\}_{n=1}^{\infty}$ solution de*

$$\frac{\partial \ln \mathcal{L}(\theta; x_1, \cdots, x_n)}{\partial \theta_k} = 0 \quad k = 1, \cdots, p$$

telle que $\hat{\theta}_n \to \theta$

Remarque - Si le système d'équations

$$\frac{\partial \ln \mathcal{L}(\theta; x_1, \cdots, x_n)}{\partial \theta_k} = 0 \quad k = 1, \cdots, p$$

admet une unique solution alors c'est forcément l'EMV. S'il y a plusieurs solutions, le théorème ne précise pas laquelle choisir ; la suite de solutions qui converge peut ne pas correspondre au maximum global mais peut être un maximum local.

Preuve - Nous proposons une preuve pour $p = 1$ c'est à dire $\theta \in \mathbb{R}$. On s'intéresse aux solutions de

$$\frac{\partial \ln \mathcal{L}(\theta; x_1, \cdots, x_n)}{\partial \theta} = 0$$

La suite $X_1, \cdots, X_n$ étant constituée de variables aléatoires i.i.d. ceci revient à étudier

$$\frac{\sum_{i=1}^{n} \partial \ln \mathcal{L}(\theta; x_i)}{\partial \theta} = 0$$

On note θ_0 la vraie valeur du paramètre θ (inconnu). Résoudre l'équation précédente revient à résoudre

$$\frac{\partial}{\partial \theta} \frac{1}{n} \ln \left(\frac{\mathcal{L}(\theta; X_i)}{\mathcal{L}(\theta_0; X_i)} \right) = 0$$

On note

$$S_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln \left(\frac{\mathcal{L}(\theta; X_i)}{\mathcal{L}(\theta_0; X_i)} \right)$$

On va montrer qu'il existe θ_0^- et θ_0^+ autour de θ distants de $\epsilon > 0$ tels que

$$S_n(\theta_0^-) < 0, \; S_n(\theta_0^+) < 0, \; S_n(\theta_0) = 0$$

donc il existe un maximum $\hat{\theta}_n$ dans $]\theta_0^-, \theta_0^+[$ verifiant $|\hat{\theta}_n - \theta_0| < \epsilon$.
Si

$$\ln\left(\frac{\mathcal{L}(\theta; X_i)}{\mathcal{L}(\theta_0; X_i)}\right) \in L^1(P_{\theta_0})$$

alors par la loi forte des grands nombres $S_n(\theta)$ tend presque sûrement vers

$$E_{\theta_0}\left[\ln\left(\frac{\mathcal{L}(\theta; X)}{\mathcal{L}(\theta_0; X)}\right)\right].$$

On montre que la limite est strictement négative si $\theta \neq \theta_0$.

Inégalité de Jensen - Si ψ est stritement convexe et $Y \in L^1(P)$, $\psi(E(Y)) \leq E(\psi(Y))$. Et si on a égalité, $Y =$ cte p.s..

La fonction ln est strictement concave, donc $-\ln$ est strictement convexe. Donc par Jensen

$$E_{\theta_0}\left[\ln\left(\frac{\mathcal{L}(\theta; X)}{\mathcal{L}(\theta_0; X)}\right)\right] \leq \ln\left(E_{\theta_0}\left[\left(\frac{\mathcal{L}(\theta; X)}{\mathcal{L}(\theta_0; X)}\right)\right]\right)$$

or

$$E_{\theta_0}\left[\left(\frac{\mathcal{L}(\theta; X)}{\mathcal{L}(\theta_0; X)}\right)\right] = \int \frac{\mathcal{L}(\theta; x)}{\mathcal{L}(\theta_0; x)}\mathcal{L}(\theta_0; x)d\mu(x) = 1$$

car $x \mapsto \mathcal{L}(\theta; x)$ est une densité de probabilité par rapport à la mesure μ. Donc

$$E_{\theta_0}\left[\ln\left(\frac{\mathcal{L}(\theta; X)}{\mathcal{L}(\theta_0; X)}\right)\right] \leq 0$$

Quand on a égalité, c'est équivalent à

$$\ln\left(\frac{\mathcal{L}(\theta; x)}{\mathcal{L}(\theta_0; x)}\right) = \text{cte p.s.}$$

ou encore à

$$\frac{\mathcal{L}(\theta; x)}{\mathcal{L}(\theta_0; x)} = \text{cte p.s.}$$

On en déduit que $\mathcal{L}(\theta; x) = \mathcal{L}(\theta_0; x)$ car $\mathcal{L}(\theta; x)$ et $\mathcal{L}(\theta_0; x)$ sont des densités de probabilité. On a alors $\theta = \theta_0$ par l'hypothèse d'identifiabilité.
Si $\theta \neq \theta_0$,

$$E_{\theta_0}\left[\ln\left(\frac{\mathcal{L}(\theta; X)}{\mathcal{L}(\theta_0; X)}\right)\right] < 0$$

donc pour tout $\theta \neq \theta_0$ et pour tout $\omega \in \Omega_\theta$ avec $P(\Omega_\theta) = 1$, il existe un n_0 tel que pour $n \geq n_0$ on a $S_n(\theta, \omega) < 0$ p.s.
On considère les ensembles $\Theta_0 = \{\theta \in \Theta, \theta = \theta_0 \pm \frac{1}{k}, k \in \mathbb{N}^*\}$ et $\Omega = \cap_{\theta \in \Theta_0}\Omega_\theta$ qui est mesurable car Θ_0 est dénombrable. On remarque que Ω est fixé quelque soit θ. Soit $\omega \in \Omega$, soit $\epsilon > 0$ on choisit $\theta_0^- = \theta_0 - \frac{1}{k}$ et $\theta_0^+ = \theta_0 - \frac{1}{k}$ tels que $|\theta_0^+ - \theta_0^-| < \epsilon$. On a $S_n(\theta_0^-, \omega) < 0$, $S_n(\theta_0^+, \omega) < 0$ et $S_n(\theta_0, \omega) = 0$. Donc il existe $\hat{\theta}_n \in]\theta_0^-, \theta_0^+[$ tel que

$$\frac{\partial S_n(\theta, \omega)}{\partial \theta}\Big|_{\theta=\hat{\theta}_n} = 0 \tag{4.1}$$

Finalement, $\forall \omega \in \Omega$, avec $P(\Omega) = 1$, pour tout $\epsilon > 0$, $\exists n_0$ tel que $\forall n \geq n_0$, on peut trouver $\hat{\theta}_n$ vérifiant l'équation (4.1), c'est à dire $\hat{\theta}_n$ tend presque sûrement vers θ_0. $\diamond$

4.3 Etimateur efficace

On étudie l'optimalité des estimateurs sans biais. Parmi les estimateurs sans biais, on cherche ceux dont la variance est minimale (ou de manière équivalente l'EQM est minimale). On suppose que l'on est dans un modèle paramétrique (E, P_θ) et qu'on veut estimer une fonction quelconque $g(\theta)$ de θ. Dans un premier temps, on se restreint au cas $\theta \in \mathbb{R}$.

Théorème 11 - *Soit* $\mathbf{X} = (X_1, \cdots, X_n)$ *défini sur* $(E, P_\theta)_{\theta\in\Theta}$ *de densité jointe* $f(\mathbf{x}; \theta$ *et vérifiant les hypothèses (1) L'ensemble* $A = \{x \in E : f(\mathbf{x}; \theta) > 0\}$ *ne dépend pas de* θ
(2) Pour tout $\mathbf{x} \in A$, $f(\mathbf{x}; \theta$ *est différentiable par rapport à* θ.
(3) $E_\theta[\frac{\partial \ln f(\mathbf{X};\theta}{\partial\theta}] = 0$
Soit $T \in L^1(P_\theta)$ *une statistique telle que* $g(\theta) = E_\theta[T(\mathbf{X})]$ *différentiable,*

$$g'(\theta) = E_\theta\left[T(\mathbf{X})\frac{\partial \ln f(\mathbf{X};\theta}{\partial\theta}\right]$$

Alors,

$$Var_\theta(T(\mathbf{X})) \geq \frac{g'(\theta)^2}{I_n(\theta)}$$

Le minorant est appelé *borne de Cramer-Rao*. La borne de Cramer-Rao est atteinte si

$$Var_\theta(T) = \frac{(g'(\theta))^2}{I(\theta)}$$

donc si

$$Var_\theta(T) = \frac{Cov^2_\theta(T, U_\theta)}{I(\theta)}$$

avec $U_\theta = \frac{\partial \ln f(\mathbf{X};\theta}{\partial\theta}$. Or ceci est vérifié si et seulement si U_θ est une fonction affine de T ; c'est à dire si, avec probabilité 1,

$$U_\theta = C(\theta)T + \delta(\theta)$$

pour tout $\mathbf{x} \in A$. Ainsi

$$\ln f(\mathbf{x}; \theta) = C^T(\theta)T(\mathbf{x}) + \delta^T(\theta) + S((x))$$

Autrement dit, $Var_\theta(T)$ n'atteint la borne de Cramar-Rao si et seulement si la fonction de densité de $(X_1, \cdots, X_n)$ appartient à la famille exponentielle à un paramètre. En particulier, T doit être une statistique exhaustive.

Preuve du théorème 11 - Par l'inégalité de Cauchy-Shwarz,

$$Var_\theta(T) \geq \frac{Cov^(T, U_\theta)}{Var_\theta(U_\theta)}$$

Comme $E_\theta[U_\theta] = 0$, on a $Var_\theta(U_\theta) = I(\theta)$. De plus,

$$\begin{aligned} Cov_\theta(T, U_\theta) &= E_\theta(TU_\theta) - E_\theta(T)U_\theta(U_\theta) \\ &= E_\theta(TU_\theta) \\ &= \frac{d}{d\theta}E_\theta(T) = g'(\theta) \end{aligned}$$

ce qui conlut la preuve. $\diamond$

Définition 24 - *On dit qu'un estimateur est efficace s'il est sans biais et que sa variance atteint la borne de Cramer-Rao.*

Définition 25 - *Soient T_n et T'_n deux estimateurs sans biais de $g(\theta)$. T'_n est dit plus efficace que T_n s'il est préférable au sens de la variance :*

$$Var_\theta(T'_n) \leq Var_\theta(T_n) \text{ pour tout } \theta \in \Theta$$

On dit que l'estimateur sans biais T'_n est uniformément plus efficace si il est plus efficace que tous les estimateurs sans biais. On dit aussi qu'il est de variance minimale.

Définition 26 *Une statistique complète est une statistique exhaustive minimale ne contient plus que de l'information utile à l'estimation de du paramètre θ. Ceci est formalisé ainsi : T est une statistique complète si pour toute fonction intégrale g :*

$$E_\theta(g(T)) = 0 \ \forall \theta \ \Theta \text{ implique que } g(T) = 0 p.s.$$

On rappelle que pour deux matrices A et B on a $A \leq B$ si et seulement si $B - A$ est une matrice symétrique positive.
Le critère d'efficacité n'a de sens que pour discriminer les estimateurs sans biais.

Théorème 12 *(Théorème de Lehmann-Scheffé). Si T_n est un estimateur sans biais de $g(\theta)$ et si S_n est une statistique exhaustive et complète, alors l'unique estimateur de $g(\theta)$ sans biais uniformément de variance minimale est $T'_n = E_\theta(T_n|S_n)$.*

Théorème 13 *Pour toute statistique T, on a*

$$I_T(\theta) \leq I_n(\theta)$$

et $I_T(\theta) = I_n(\theta)$ si et seulement si T est exhaustive,
$I_T(\theta) = 0$ si et seulement T est libre (c'est à dire que sa loi ne dépend pas de θ.

Remarques
- Un estimateur efficace est de variance minimale.
- Un estimateur peut être sans biais, de variance minimale, mais ne pas atteindre la borne de Cramer-Rao, donc ne pas être efficace. Dans ce cas-là, la borne Cramer-Rao est "trop petite" pour être atteinte.

Exemple - modèle de Poisson (voir TD).

Chapitre 5

Tests d'hypothèses

5.1 Principe et définitions

On se place dans un modèle paramétrique $\mathbf{X} = (X_1, \cdots, X_n)$ i.i.d suivant $(P_\theta, \theta \in \Theta)$. Supposons que $\Theta = \Theta_0 \cup \Theta_1$ où Θ_0 et Θ_1 sont deux ensembles disjoints. Connaissant une réalisation $(x_1, \cdots, x_n)$ de $\mathbf{X}$, on voudrait décider si θ est dans Θ_0 ou Θ_1. En pratique, on choisira toujours pour Θ_0 le plus petit des deux sous-espaces Θ_0, Θ_1. Ainsi, $\theta \in \Theta_0$ correspond à la version la plus simple du modèle.

5.1.1 Hypothèses de test

On pose une *hypothèse nulle* notée

$$H_0 : \theta \in \Theta_0$$

contre une *hypothèse alternative* notée

$$H_1 : \theta \in \Theta_1$$

Exemple - Supposons que $X_1, \cdots, X_m$ et $Y_1, \cdots, Y_n$ sont des variables indépendantes telles que $X_i \sim \mathcal{N}(\mu_1, \sigma_2)$ et $Y_i \sim \mathcal{N}(\mu_2, \sigma_2)$. L'espace des paramètres est

$$\Theta = \{(\mu_1, \mu_2, \sigma) : -\infty < \mu_1, \mu_2 < \infty, \sigma > 0\}$$

Dans les applications, on cherche à déterminer si les X_i et les Y_i ont la même distribution (c'est à dire si $\mu_1 = \mu_2$). Par exemple, on administre 2 somminifères différents à 2 groupes de patients et on cherche à savoir si la durée moyenne du sommeil est la même dans les deux groupes. On définit alors Θ_0 par

$$\Theta_0 = \{(\mu_1, \mu_2, \sigma) : -\infty < \mu_1 = \mu_2 < \infty, \sigma > 0\}$$

et Θ_1 est le complémentaire de Θ_0. On remarque ici de Θ et de dimension 3 alors que Θ_0 n'est plus que de dimension 2.

5.1.2 Principe général

On décide que θ est dans Θ_0 ou Θ_1 à l'aide des observations $(x_1, \cdots, x_n)$. Pour cela on cherche une règle de décision qui prend la forme suivante.

- Si $S(x_1, \cdots, x_n) \in$ RC alors on rejette H_0
- Si $S(x_1, \cdots, x_n) \notin$ RC alors on ne rejette pas H_0

où S est une *statistique* (ou fonction) *de test* et RC est une *région critique*. Le plus souvent, S est une fonction d'un estimateur de θ.

Exemple - Supposons que $X_1, \cdots, X_n$ sont des variables i.i.d telles que $E(X_i) = \theta$. On souhaite tester

$$H_0 : \theta = 0 \text{ contre } H_1 : \theta \neq 0$$

On peut alors choisir $S(X_1, \cdots, X_n) = \bar{X}$. Et on rejette H_0 si $\bar{x}$ est éloigné de 0 c'est à dire si RC $=]\infty, -c] \cup [c, \infty[$ avec c une constante positive. On verra plus loin comment on détermine c.

5.1.3 Erreurs et puissance

Il est peu probable qu'une règle de décision soit parfaite. Ainsi, quand on définit une règle de décision, on doit regarder qu'elle est la probabilité de faire des erreurs quand on prend l'une ou l'autre décision en fonction de la valeur de $\theta \in \Theta$. On peut faire deux types d'erreur

- On fait une *erreur de première espèce*[1] quand on rejette H_0 à tort c'est à dire alors que $\theta \in \Theta_0$. On peut associer une probabilité à cette erreur :

  $$P(S(X_1, \cdots, X_n) \in \text{RC} | \theta \in \Theta_0)$$

  On parle aussi parfois de *risque de première espèce*.
- On fait une *erreur de seconde espèce*[2] quand on accepte H_0 à tort c'est à dire alors que $\theta \in \Theta_1$. La probabilité associée est alors

  $$P(S(X_1, \cdots, X_n) \notin \text{RC} | \theta \in \Theta_1)$$

On pourrait être tenté de chercher à définir des statistiques de tests telles que ces deux erreurs soient uniformément petites pour tout $\theta \in \Theta$. On verra que c'est généralement impossible.

Définition 27 - *On appelle fonction de risque de première espèce la fonction*

$$\begin{aligned} \alpha \quad : \quad & \Theta_0 \to [0,1] \\ & \theta \mapsto P_\theta(S(X_1, \cdots, X_n) \in RC) \end{aligned}$$

On appelle fonction de risque de seconde espèce la fonction

$$\begin{aligned} \beta \quad : \quad & \Theta_1 \to [0,1] \\ & \theta \mapsto P_\theta(S(X_1, \cdots, X_n) \notin RC) \end{aligned}$$

1. *en anglais :* type I error
2. *en anglais :* type II error

Ainsi, α représente la probabilité de se tromper quand on est sous H_0. On note parfois abusivement $\alpha(\theta) = P_\theta(H_1|H_0)$ et $\beta(\theta) = P_\theta(H_0|H_1)$.

Définition 28 - *On appelle fonction puissance la fonction*

$$\begin{aligned} \Pi \quad : \quad & \Theta_1 \to [0,1] \\ & \theta \mapsto P_\theta(S(X_1, \cdots, X_n) \in RC) \end{aligned}$$

Propriété - On a pour tout $\theta \in \Theta_1$, $\Pi(\theta) = 1 - \beta(\theta)$

Définition 29 - $\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)$ *est appelé niveau du test. C'est le risque de première espèce maximal.*

5.1.4 Principe de Neyman

On voudrait trouver des procédures de test qui minimisent les 2 erreurs. Or il est facile de voir, que le plus souvent, si α diminue alors β augmente (voir par exemple le graphique de la puissance du test du signe ci-dessous).

Le principe de Neyman consiste à fixer le niveau α à une valeur petite (typiquement 5% ou 1 %) et à chercher une région critique qui minimise $\beta(\theta)$ à α fixé.

En pratique :

1. On fixe le niveau α.
2. On en déduit une région critique : RC(α). Si plusieurs régions sont possibles, on choisit celle qui minimise $\beta(\theta)$.
3. On conclut : si $S(x_1, \cdots, x_n) \in$ RC(α), on rejette H_0.

   On utilise parfois une alternative pour conclure. Au lieu de fixer α et de comparer la valeur de la statistique de test observée à la région critique RC(α), on estime un *degré de significativité* ou (p-value) :

   $$\hat{\alpha}(X_1, \cdots, X_n) = \inf\{\alpha \text{ tel que } S(X_1, \cdots, X_n) \in \text{RC}(\alpha)\}$$

   Ainsi le degré de significativité est le niveau le plus faible qu'on peut choisir pour conclure au rejet de H_0. On dit parfois que c'est l'erreur que l'on fait quand on rejette H_0. Concrètement, on compare $\hat{\alpha}$ au niveau α fixé.

5.2 Exemples

5.2.1 Test du signe

Un aquaculteur a 114 poissons d'une certaine espèce dans un de ses bassins. On extrait un échantillon de 12 poissons afin de vérifier l'hypothèse selon laquelle la médiane de la longueur des poissons est de 220 mm. On observe les longueurs suivantes :

126 142 156 228 245 246 370 419 433 454 478 503

1. Hypothèses de test : en notant μ la médiane, $H_0 : \mu = 220$ contre $H_1 : \mu \neq 220$.

2. Risque de première espèce : on fixe, arbitrairement, $\alpha = 0.05$.

3. Statistique de test. On choisit ici un test naïf appelé *test du signe*. La statistique de test est construite de la façon suivante. Si la médiane est 220, il est également probable pour chaque poisson sélectionné d'être plus ou moins long que 220 mm. Puis on calcule S égale aux nombre d'individus plus long que 220. (Implicitement, on associe à chaque individu un signe - si sa longueur est inférieure à 220 et un signe + sinon ; S est alors la somme des signes +).

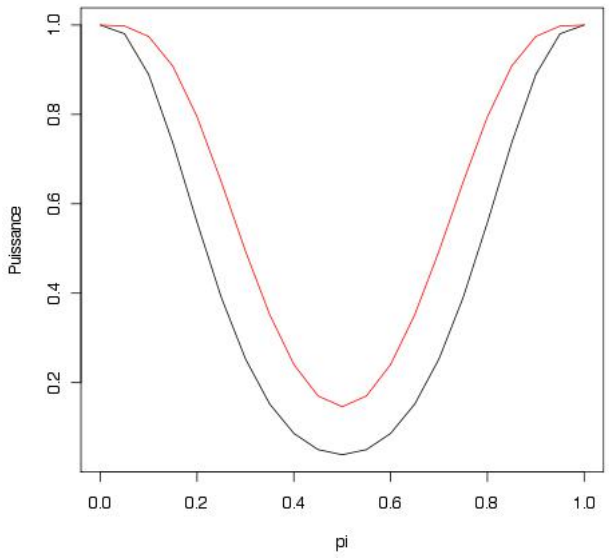
4. Loi de la statistique de test. Il est facile de voir que la loi de la statistique de test est une loi binomiale de paramètres $n = 12$ et $\pi = 1/2$.

5. Région critique. On remarque aisément que l'on va rejetter H_0 si S est trop grande ou trop petite (dominance de signes +, ou de signes -) ; la région critique est donc de la forme $\{s \text{ tels que } s \notin [s_{\text{inf}}, s_{\text{sup}}]\}$. Les bornes s_{inf} et s_{sup} sont déterminées à l'aide de la loi binomiale de telle sorte que

$$P_{H_0}(S < s_{\text{inf}}) = \frac{\alpha}{2} \text{ et } P_{H_0}(S > s_{\text{sup}}) = \frac{\alpha}{2}$$

   En s'aidant de la table 5.1, on trouve que $s_{\text{inf}} = 2$ et $s_{\text{inf}} = 10$ (car $P(x = 0) + P(x = 1) + P(x = 2) < .025$).

6. Puissance du test. On ne peut calculer $P_{H_1}(S \notin [s_{\text{inf}}, s_{\text{sup}}])$ que si on choisit une alternative ponctuelle pour H_1 c'est à dire si on fixe la valeur de la médiane sous H_1. En pratique, on dispose généralement pas d'une telle information. On peut alors regarder comment varie la puissance pour différentes valeurs de la médiane. Par exemple, le graphique ci-dessous montre la puissance du test du signe quand π varie. La courbe en noir correspond à un risque de première espèce de 5% et la courbe en rouge à un risque de première espèce de 20%.



r	0	1	2	3	4	5	6
P	0.000	0.003	0.016	0.054	0.121	0.193	0.226
		7	8	9	10	11	12
		0.193	0.121	0.054	0.016	0.003	0.000

TABLE 5.1 – Probabilités binomiales $P(S = k)$, $n = 12$, $\pi = \frac{1}{2}$

5.2.2 Test pour la moyenne d'une loi de Gauss

Un contrôle anti-dopage a été effectué sur 16 sportifs. On a mesuré la variable X de moyenne m, qui est le taux (dans le sang) d'une certaine substance interdite. Voici les données obtenues :

0.35	0.4	0.65	0.27	0.14	0.59	0.73	0.13
0.24	0.48	0.12	0.70	0.21	0.13	0.74	0.18

La variable X est supposée gausienne et de variance $\sigma^2 = 0.04$. On veut tester, au niveau 5% l'hypothèse selon laquelle le taux moyen dans le sang de la population des sportifs est égal à 0.4.

On pose des hypothèses de test unilatérales :

$$H_0 : m = m_0 = 0.4 \text{ contre } H_1 : m > 0.4$$

La statistique de test est la moyenne empirique. Si on note $X_1, \cdots, X_n$ l'échantillon de variables aléatoires de même loi que X, la moyenne empirique est donnée par

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Intuitivement, on comprend bien qu'on va rejeter H_0 si $\bar{X}_n - m_0$ est trop grande en valeur absolue c'est à dire si la moyenne empirique est trop éloignée de la moyenne sous H_0.

Sous H_0, $Z = \frac{\bar{X}_n - m_0}{\sigma/\sqrt{n}}$ suit une loi de Gauss de moyenne 0 et de variance 1. D'autre part, d'après la remarque faite plus haut on comprend qu'on rejette H_0 si $|Z| > z_0$. Pour construire la région critique, on cherche donc z_0 tel que

$$P(|Z| > z_0) = \alpha$$

soit encore

$$P(Z > z_0 \text{ou} Z < z_0) = P(Z > z_0) + P(Z < -z_0) = \alpha$$

or on a par symétrie de la loi de Gauss de moyenne 0 et de variance 1

$$P(Z > z_0) = P(Z < -z_0) = \Phi(-z_0) = 1 - \Phi(z_0)$$

où on note Φ la fonction de répartition de la loi Gauss de moyenne 0 et de variance 1. Ainsi z_0 est tel que

$$1 - \Phi(z_0) = \alpha/2$$

ce qui s'écrit encore

$$z_0 = \Phi^{-1}(1 - \alpha_2)$$

D'après la table de la fonction de répartition inverse de la loi normale, on en déduit que $z_0 = 1.96$ car $\alpha = 0.05$.

Finalement, on rejette donc H_0 si

$$|\bar{X}_n - m_0| > 1.96 \frac{\sigma}{\sqrt{n}}$$

.

Remarques

- Lorsque le nombre d'observations n est grand (supérieur à 30), d'après le théorème de limite centrale on a que la statistique de test

$$Z = \frac{\bar{X} - m_0}{\sigma/\sqrt{n}}$$

suit approximativement une loi de Gauss quelque soit la loi de la variable X considérée.

Si la variance est inconnue

Dans le cas où la variance n'est pas connue, on doit l'estimer en utilisant les observations. Et la statistique de test du test de la moyenne donnée par

$$Z = \frac{\bar{X} - m_0}{s/\sqrt{n}}$$

Elle ne suit plus une loi de Gauss car le dénominateur n'est plus une constante mais une réalisation de l'estimateur de la variance de la variable X. L'écart-type s est obtenu par

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

Par construction, S^2 suit une loi du χ^2. Y est donc une v.a. suivant une de Student à $(n-1)$ degrés de libertés. Et on utilise alors une table de la loi de Student pour conclure le test.

Remarque : Lorsque le nombre d'observations n est grand (supérieur à 30), on peut utiliser le théorème de limite centrale pour approcher la loi de la statistique Z.

Calcul de la puissance du test

Dans le cas d'un test de Student, on peut calculer la puissance du test si on on peut donner une valeur de la moyenne sous l'hypothèse alternative.

$$H_0 : m = m_0 \text{ contre } H_1 : m = m_1$$

La puissance est définie par

$$\mathcal{P} = P(\text{rejeter } H_0 | H_0 \textbf{estfausse})$$

Ainsi la puissance est la probabilité de la la région de rejet de H_0 sous la loi de H_1.

$$\begin{aligned}
\mathcal{P} &= P\left(Z > z_0 | \frac{Z - m_1}{\sigma/\sqrt{n}} \text{ suit une loi } \mathcal{N}(0,1)\right) \\
&= P\left(\tilde{Z} > \frac{z_0 - m_1}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\frac{z_0 - m_1}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

5.3 Principe de Neyman et optimalité

5.3.1 Test randomisé

Affinons le test du signe réalisé plus haut. En effet, nous avons remarqué qu'on ne peut pas toujours atteindre exactement le niveau α fixé.

Dans l'exemple des poissons, si on pose des hypothèses de test unilatérales, correspondants à la question : "les poissons péchés sont-il trop petits (de longueur inférieure à 22 mm) ?",

$$H_0 : \mu = 22 \text{ contre } H_1 : \mu < 22$$

alors la région critique est de la forme RC $= \{S \leq c\}$ où c doit vérifier $P_{H_0}(S \leq c) = \alpha$. Or sous H_0, la statistique de test suit une loi binomiale (définie sur un ensemble discret) et il n'existe pas de c qui permette d'obtenir l'égalité : $P(S \leq 2) = 0.019$ et $P(S \leq 3) = 0.073$.

Une solution consiste à randomiser le test, c'est à dire qu'on tire au hasard la solution du test avec une certaine probabilité. Dans l'exemple des poissons,

– si $s_{obs} \geq 4$ on ne refuse pas H_0 ;
– si $s_{obs} \leq 2$ on refuse H_0 ;
– si $s_{obs} = 3$, on tire au sort H_1 avec une probabilité $\gamma \in [0,1]$. On choisit γ telle que

$$\begin{aligned} \alpha = P_{H_0}(\text{RC}) &= 0 \times P_{H_0}(S \geq 4) + \gamma \times P_{H_0}(S = 3) + 1 \times P_{H_0}(S \leq 2) \\ &= \gamma P_{H_0}(S = 3) + P_{H_0}(S \leq 2) \end{aligned}$$

d'où $\gamma = (0.05 - 0.019)/0.054 = 0.57$.
Donc, si $s_{obs} = 3$ on décide H_1 avec une probabilité de 57%.

Définition 30 *Soit $(X_1, \cdots, X_n)$ à valeur dans E^n. Un test est une fonction aléatoire Ψ de $E^n \to [0,1]$.*

Interprétation - La fonction Ψ représente la probabilité de décider H_1.
Si $\Psi(X_1, \cdots, X_n) = 0$ on conclut à H_0.
Si $\Psi(X_1, \cdots, X_n) = 1$ on conclut à H_1.
Si $\Psi(X_1, \cdots, X_n) \in]0,1[$ on tire au hasard la décision H_1 avec la probabilité $\Psi(X_1, \cdots, X_n)$.

Lorsque Ψ est à valeurs dans $\{0,1\}$, on parle de test pur, c'est à dire non randomisé. C'est le cas de la plupart des tests classiques. Par exemple quand on teste, pour une variable aléatoire de loi de Gauss de moyenne μ,

$$H_0 : \mu = \mu_0 \text{ contre } H_1 : \mu \neq \mu_0$$

La région critique est

$$\text{RC} = \left\{ \sqrt{n} \left| \frac{\bar{x}_n - \mu_0}{s_n} \right| > F_{n-1}^{-1}(1 - \alpha/2) \right\}$$

au niveau α avec F_{n-1} la fonction de répartition de la loi de Student à $n-1$ degrés de liberté. C'est un test pur pour lequel $\Psi(x_1, \cdots, x_n) = \mathbb{I}_{\{\text{RC}\}}$.

Définition 31 *Pour le test $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$,*
– *le risque de 1ère espèce est la fonction $\alpha(\theta) = E_\theta(\Psi(X_1, \cdots, X_n))$, $\forall \theta \in \Theta_0$;*
– *le risque de 2nde espèce est la fonction $\beta(\theta) = E_\theta(1 - \Psi(X_1, \cdots, X_n))$, $\forall \theta \in \Theta_1$;*
– *le niveau est $\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)$;*
– *la puissance du test est la fonction $\Pi(\theta) = 1 - \beta(\theta)$.*

L'utilisation de tests randomisés permet de considérer des tests de niveau α pour tout $\alpha \in [0,1]$. Ils existent d'après le lemme de Neyman-Pearson donné ci-dessous. On peut donc définir la notion de test le plus puissant parmi les tests de niveau α.

Définition 32 *Un test associé à la fonction Ψ est un test uniformément plus puissant (UPP) au niveau α, si son niveau est inférieur ou égal à α et si pour tout test Ψ^* de niveau inférieur ou égal à α,*

$$\Pi_\theta(\Psi(X_1, \cdots, X_n)) \geq \Pi_\theta(\Psi^*(X_1, \cdots, X_n))$$

pour tout $\theta \in \Theta_1$.

5.3.2 Tests uniformément plus puissants

Le lemme de Neyman-Pearson est important car il suggère un principe pour trouver de "bons" tests au sens du compromis entre une puissance forte et une erreur de première espèce faible.

Théorème 14 - Lemme de Neyman-Pearson. *Soit $(X_1, \cdots, X_n)$ un échantillon de vraisemblance $\mathcal{L}(\theta; X_1, \cdots, X_n)$. Pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, $\theta_0 \neq \theta_1$, pour tout $\alpha \in]0,1[$ fixé, il existe $c > 0$ et $\gamma \in [0,1[$ tels que le test*

$$\Psi(X_1, \cdots, X_n) = \begin{cases} 1 \ si \ \frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} > c \\ \\ \gamma \ si \ \frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} = c \\ \\ 0 \ si \ \frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} < c \end{cases}$$

De plus ce test est uniformément plus puissant parmi les tests de niveau au plus α et c'est le seul. Ce test est appelé test de Neyman-Pearson associé à c et γ est déterminé par l'équation de test $E_{\theta_0}(\Psi(X_1, \cdots, X_n)) = \alpha$. ($\gamma$ n'est pas forcément unique.)

Preuve - 1. Nous montrons tout d'abord que le niveau est bien α.

$$\begin{aligned} E_{\theta_0}(\Psi(X_1, \cdots, X_n)) &= 1 \times P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} > c\right) \\ &+ \gamma \times P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} = c\right) \\ &+ 0 \times P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} < c\right) \end{aligned}$$

On veut trouver c et γ tels que $E_{\theta_0}(\Psi(X_1, \cdots, X_n)) = \alpha$ pour tout $\alpha \in]0,1[$. Soit F la fonction de répartition de $\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)}$

$$F(t) = P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} \leq t\right).$$

• Si F est continue, alors il existe c tel que $F(c) = 1 - \alpha$. En prenant ce c et $\gamma = 0$, on a

$$\begin{aligned} &P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} > c\right) + \gamma P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = c\right) \\ &\qquad = 1 - F(c) = 1 - (1 - \alpha) = \alpha \end{aligned}$$

• Si F n'est pas continue (cas discret), on note $c^+ = \min_t(F(t) \geq 1 - \alpha)$ et on a

$$F(c^+) = \lim_{x \to c^+, x > c^+} F(t), F(c^-) = \lim_{x \to c^+, x < c^+} F(t)$$

ainsi

$$F(c^+) - F(c^-) = \lim_{\substack{x \to c^+, x > c^+ \\ y \to c^+, y < c^+}} P\left(y < \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} \leq x\right) = P\left(\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = c^+\right)$$

Or on cherche c et γ tels que $E_{\theta_0}(\Psi(X_1,\cdots,X_n)) = \alpha$. En choisissant $c = c^+$, on obtient

$$1 - F(c^+) + \gamma(F(c^+) - F(c^-)) = \alpha$$

ce qui est équivalent à

$$\gamma = \frac{\alpha + F(c^+) - 1}{F(c^+) - F(c^-)}$$

Il reste à verifier que γ appartient à $[0,1[$.

$$\gamma \geq 0 \Leftrightarrow \alpha - 1 + F(c^+) \geq 0 \Leftrightarrow F(c^+) \geq 1 - \alpha$$

ce qui est vrai par définition de c^+. Par ailleurs,

$$\gamma < 1 \Leftrightarrow \alpha - 1 + F(c^+) < F(c^+) - F(c^-) \Leftrightarrow F(c^-) < 1 - \alpha$$

ce qui est vrai par définition.

2. Montrons maintenant que le test est UPP.
Soit Ψ^* un test de niveau au plus α. n considère l'intégrale

$$\int_{E^n} (\Psi(x_1,\cdots,x_n) - \Psi^*(x_1,\cdots,x_n)(\mathcal{L}(\theta_1) - c\mathcal{L}(\theta_0)))\, d\mu(x_1,\cdots,x_n) \tag{5.1}$$

où μ est la mesure de référence par rapport à laquelle $\mathcal{L}(\theta_0)$ et $\mathcal{L}(\theta_1)$ sont définies. Cette intégrale est toujours positive car
- si $\mathcal{L}(\theta_1) - c\mathcal{L}(\theta_0) > 0$ alors $\Psi(X_1,\cdots,X_n) = 1$ par définition de Ψ et, $\Psi(X_1,\cdots,X_n) \geq \Psi^*(X_1,\cdots,X_n)$;
- si $\mathcal{L}(\theta_1) - c\mathcal{L}(\theta_0) < 0$ alors $\Psi(X_1,\cdots,X_n) = 0$ par définition de Ψ et, $\Psi(X_1,\cdots,X_n) \leq \Psi^*(X_1,\cdots,X_n)$;
- si $\mathcal{L}(\theta_1) - c\mathcal{L}(\theta_0) = 0$ alors l'intégrale est nulle.

On a donc

$$\int_{E^n} \Psi(x_1,\cdots,x_n)\mathcal{L}(\theta_1)d\mu(x_1,\cdots,x_n) - c\int_{E^n} \Psi(x_1,\cdots,x_n)\mathcal{L}(\theta_0)d\mu(x_1,\cdots,x_n)$$
$$- \int_{E^n} \Psi^*(x_1,\cdots,x_n)\mathcal{L}(\theta_1)d\mu(x_1,\cdots,x_n) + c\int_{E^n} \Psi^*(x_1,\cdots,x_n)\mathcal{L}(\theta_0)d\mu(x_1,\cdots,x_n) \geq 0$$

Ce qui s'écrit aussi

$$E_{\theta_1}(\Psi) - cE_{\theta_0}(\Psi) - E_{\theta_1}(\Psi^*) + cE_{\theta_0}(\Psi^*) \geq 0$$

ou encore

$$E_{\theta_1}(\Psi) - E_{\theta_1}(\Psi^*) \geq c(E_{\theta_0}(\Psi) - E_{\theta_0}(\Psi^*))$$

et on reconnait que
- $E_{\theta_1}(\Psi)$: puissance de Ψ
- $E_{\theta_1}(\Psi^*)$: puissance de Ψ^*
- $E_{\theta_0}(\Psi)$: niveau de Ψ qui est égal à α
- $E_{\theta_0}(\Psi)$: niveau de Ψ^* qui est inférieur ou égal à α

On en déduit donc que $E_{\theta_1}(\Psi) - E_{\theta_1}(\Psi^*) \geq 0$ et que la puissance Ψ est supérieure à celle de tout test Ψ^* de niveau au plus α.

3. On montre enfin que c'est le seul test UPP, à γ près.
Soit Ψ^* un test UPP au niveau au plus α. On a donc
$E_{\theta_1}(\Psi^*(X_1, \cdots, X_n)) \geq E_{\theta_1}(\Psi(X_1, \cdots, X_n))$ car Ψ^* est UPP,
et $E_{\theta_1}(\Psi(X_1, \cdots, X_n)) \geq E_{\theta_1}(\Psi^*(X_1, \cdots, X_n))$ car Ψ est UPP.
Donc, ces deux tests sont de même puissance

$$E_{\theta_1}(\Psi(X_1, \cdots, X_n)) = E_{\theta_1}(\Psi^*(X_1, \cdots, X_n))$$

Reprenons l'intégrale (5.1) :

$$\begin{aligned}\int_{E^n} (\Psi(x_1, \cdots, x_n) - \Psi^*(x_1, \cdots, x_n)(\mathcal{L}(\theta_1) - c\mathcal{L}(\theta_0)))\, d\mu(x_1, \cdots, x_n) \\ = E_{\theta_1}(\Psi) - cE_{\theta_0}(\Psi) - E_{\theta_1}(\Psi^*) + cE_{\theta_0}(\Psi^*)\end{aligned}$$

On a noté que cette intégrale est positive donc $E_{\theta_0}(\Psi) - E_{\theta_0}(\Psi^*) \leq 0$. Or on a vu que $E_{\theta_0}(\Psi) = \alpha$ et $E_{\theta_0}(\Psi^*) \leq \alpha$. Donc $E_{\theta_0}(\Psi) = E_{\theta_0}(\Psi^*)$ et l'intégrale est nulle. Comme $\mathcal{L}(\theta_1) - c\mathcal{L}(\theta_0)$ est différent de 0, cela implique $Psi = Psi^*$, μ presque sûrement. Donc les tests coïncident (à γ près qui reste à déterminer). $\diamond$

Remarque - Dans le cas continu (μ est la mesure de Lebesgue), on retrouve un test pur de région critique

$$\text{RC} = \left\{ \frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} > c \right\}$$

En effet, dans ce cas,

$$E_{\theta_0}(\Psi(X)) = 1 \times P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1; X)}{\mathcal{L}(\theta_0; X)} > c \right) + \gamma \times P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1; X)}{\mathcal{L}(\theta_0; X)} = c \right)$$

or

$$P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1; X)}{\mathcal{L}(\theta_0; X)} = c \right) = 0$$

car la loi est continue.
Pour résoudre $E_{\theta_0}(\Psi(X)) = \alpha$, on peut choisir γ quelconque : on prend $\gamma = 0$ et le test devient

$$\Psi(X_1, \cdots, X_n) = \mathbb{I}_{\left\{ \frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0;; X_1, \cdots, X_n)} > c \right\}}$$

autrement dit, un test pur de région critique

$$\text{RC} = \left\{ \frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} > c \right\}$$

Pour déterminer c, on résoud $E_{\theta_0}(\Psi(X_1, \cdots, X_n)) = \alpha$ ce qui est équivalent à résoudre

$$P_{\theta_0}\left(\frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} > c \right) = \alpha$$

Exemple 1 - Soient $X_1, \cdots, X_n$ des v.a.i.i.d de loi de Gauss de moyenne θ et de variance un. On teste

$$H_0 : \theta = \theta_0 \text{ contre } H_1 : \theta = \theta_1$$

à partir de $(x_1, \cdots, x_n)$ une réalisation de $(X_1, \cdots, X_n)$. La vraisemblance s'écrit

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \frac{1}{\left(\sqrt{2\pi}\right)^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

Ainsi, le rapport des vraisemblances est

$$\frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} = \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_1)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_0)^2\right)$$

en passant au log, on obtient

$$\begin{aligned}\frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} > c \Leftrightarrow -\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_1)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_0)^2 > \log(c)\\ \Leftrightarrow (\theta_1 - \theta_0)\sum_{i=1}^{n} x_i - \frac{n\theta_1^2}{2} + \frac{n\theta_0^2}{2} > \log(c)\\ \Leftrightarrow (\theta_1 - \theta_0)\sum_{i=1}^{n} x_i > \log(c) + \frac{n\theta_1^2}{2} - \frac{n\theta_0^2}{2}\end{aligned}$$

Si $\theta_1 > \theta_0$, l'inégalité devient

$$\sum_{i=1}^{n} x_i > \frac{1}{\theta_1 - \theta_0}\left(\log(c) + \frac{n\theta_1^2}{2} - \frac{n\theta_0^2}{2}\right)$$

et si au contraire, $\theta_1 < \theta_0$, elle s'écrit

$$\sum_{i=1}^{n} x_i < \frac{1}{\theta_1 - \theta_0}\left(\log(c) + \frac{n\theta_1^2}{2} - \frac{n\theta_0^2}{2}\right)$$

Notons

$$c' = \frac{1}{\theta_1 - \theta_0}\left(\log(c) + \frac{n\theta_1^2}{2} - \frac{n\theta_0^2}{2}\right)$$

Le test de Neyman-Pearson se ramène, dans le cas $\theta_1 > \theta_0$ à

$$\Psi(x_1, \cdots, x_n) = \begin{cases} 1 \text{ si } \sum_{i=1}^{n} x_i > c' \\ \gamma \text{ si } \sum_{i=1}^{n} x_i = c' \\ 0 \text{ si } \sum_{i=1}^{n} x_i < c' \end{cases}$$

où c' et γ sont déterminés par $E_{\theta_0}(\Psi(X_1, \cdots, X_n)) = \alpha$ ce qui est équivalent à

$$1 \times P_{\theta_0}(\sum_{i=1}^{n} X_i > c') + \gamma \times P_{\theta_0}(\sum_{i=1}^{n} X_i = c') = \alpha$$

Ici $P_{\theta_0}(\sum_{i=1}^n X_i = c') = 0$ car la loi est continue. On choisit $\gamma = 0$ et c' est obtenu en résolvant l'équation

$$P_{\theta_0}(\sum_{i=1}^n X_i > c') = \alpha$$

Or, sous H_0, $\sum_{i=1}^n X_i$ suit une loi de Gauss de moyenne $n\theta_0$ et de variance n. On peut donc écrire,

$$P_{\theta_0}\left(\frac{\sum_{i=1}^n X_i - n\theta_0}{\sqrt{(n)}} > \frac{c' - n\theta_0}{\sqrt{(n)}}\right) = \alpha$$

et on en déduit que

$$\frac{c' - n\theta_0}{\sqrt{(n)}} = \Phi^{-1}(1-\alpha)$$

avec Φ la fonction de répartition de la loi de Gauss centrée et réduite. Autrement dit $\frac{c'-n\theta_0}{\sqrt{(n)}} = \Phi^{-1}(1-\alpha)$ est le quantile d'ordre $1-\alpha$ de la loi de Gauss de moyenne 0 et de variance 1 ; on note parfois $q(1-\alpha)$. Finalement, on conclut que pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ le test de Neyman-Pearson est un test pur ($\gamma = 0$) de région critique :

$$\text{RC} = \left\{\sum_{i=1}^n x_i > \sqrt{n}\Phi^{-1}(1-\alpha) + n\theta_0\right\}$$

Faut-il ajouter l'exemple sur le paramètre d'une loi binomiale ?

5.4 Tests UPP pour les hypothèses composites

Dans la partie précédente, nous avons montré des résultats (constructifs) d'existence et d'unicité pour des tests dont les hypothèses sont des singletons.

5.4.1 Test unilatéral $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_1$

Pour le test unilatéral $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$, on peut construire un test UPP mais en se restreignant à certaines familles de lois.

Définition 33 *La famille $(P_\theta)_{\theta\in\Theta}$ est à rapport de vraisemblance monotone s'il existe une statistique $U(X_1, \cdots, X_n)$ telle que pour tout $\theta < \theta_1$, le rapport*

$$\frac{\mathcal{L}(\theta_1; x_1, \cdots, x_n)}{\mathcal{L}(\theta_0; x_1, \cdots, x_n)} = h(U(x_1, \cdots, x_n))$$

avec une fonction h une fonction strictement monotone.

Remarque - On peut toujours supposer que h est strictement croissante quite à considérer $-U(x_1, \cdots, x_n)$ à la place de $U(x_1, \cdots, x_n)$.

Exemple - la famille des lois exponentielles.
Considérons le cas particulier

$$\mathcal{L}(x_1, \cdots, x_n) = K(\theta) \exp(c(\theta) T(x_1, \cdots, x_n))$$

ainsi pour tout $\theta_1 > \theta_0$,

$$\frac{\mathcal{L}(\theta_1; x_1, \cdots, x_n)}{\mathcal{L}(\theta_0; x_1, \cdots, x_n)} = \frac{K(\theta_1)}{K(\theta_0)} \exp((\theta_1 - \theta_0) T(x_1, \cdots, x_n))$$

est une fonction croissante de $T(x_1, \cdots, x_n)$. C'est donc une famille à rapport de vraisembalnce monotone avec $U(x_1, \cdots, x_n) = T(x_1, \cdots, x_n)$.

Théorème 15 *Soient $X_1, \cdots, X_n$ v.a.i.i.d. suivant $(P_\theta)_{\theta \in \Theta}$ où $(P_\theta)_{\theta \in \Theta}$ est une famille à rapport de vraisemblance monotone. Pour tester*

$$H_0 : \theta \leq \theta_0 \text{ contre } H_1 : \theta > \theta_1$$

il existe un test UPP de niveau α de la forme

$$\Psi(X_1, \cdots, X_n) = \begin{cases} 1 \text{ si } U(X_1, \cdots, X_n) > c \\ \gamma \text{ si } U(X_1, \cdots, X_n) = c \\ 0 \text{ si } U(X_1, \cdots, X_n) < c \end{cases}$$

où U est la statistique de la définition 33 et γ et c sont définies par $E_{\theta_0}(\Psi(X_1, \cdots, X_n)) = \alpha$.

Preuve - a. On vérifie d'abord que le test est de niveau α.
Soient les hypothèses $H_0 : \theta = \theta'$ contre $H_1 : \theta = \theta''$ avec $\theta' < \theta''$. On considère le test du théorème. Ce test est exactement le test de Neyman-Pearson pour tester $H_0 : \theta = \theta'$ contre $H_1 : \theta = \theta''$. Son niveau vaut $E_{\theta_0}(\Psi(X_1, \cdots, X_n))$ que l'on note α'. On sait d'après le lemme de Neyman-Pearson que Ψ est UPP parmi tous les tests de niveau α'.
Soit minatenant le test $\psi(X_1, \cdots, X_n) = \alpha'$ pour tout $(X_1, \cdots, X_n)$. Pour tester $H_0 : \theta = \theta'$ contre $H_1 : \theta = \theta''$, $E_\theta(\psi(X_1, \cdots, X_n)) = \alpha'$ donc ψ est de niveau α' et Ψ est plus puissant que ψ. Donc

$$E_{\theta''}(\Psi(X_1, \cdots, X_n)) \geq E_{\theta''}(\Psi(X_1, \cdots, X_n)) = \alpha'$$

Ainsi pour tout $\theta' < \theta''$, $\alpha' = E_{\theta'}(\Psi(X_1, \cdots, X_n)) \leq E_{\theta''}(\psi(X_1, \cdots, X_n))$.
En particulier pour $\theta'' = \theta_0$, pour tout $\theta' < \theta_0$, on a

$$E_{\theta'}(\Psi(X_1, \cdots, X_n)) \leq E_{\theta_0}(\Psi(X_1, \cdots, X_n))$$

donc $E_{\theta'}(\Psi(X_1, \cdots, X_n)) = \alpha$ et le niveau de Ψ est bien α.

b. On montre que Ψ est UPP.
Le test de Neyman-Pearson pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ où $\theta_1 > theta_0$ est exactement Psi, sa forme ne dépend pas de θ_1. Ce test est le plus puissant pour tester $\theta = \theta_0$ contre $\theta = \theta_1$ d'après le lemme de Neyman-Pearson. Si Ψ n'était pas UPP pour tester $\theta \leq \theta_0$ contre $\theta > \theta_0$ alors il existerait un test $\Psi^{(\alpha)}$ plus puissant que Ψ au moins

pour un $\theta_1 > \theta_0$ c'est à dire $E_{\theta_1}(\Psi^{(\alpha)}(X_1, \cdots, X_n)) \geq E_{\theta_1}(\Psi(X_1, \cdots, X_n))$. Ce qui est impossible puisque Ψ est UPP pour tester $\theta = \theta_0$ contre $\theta = \theta_1$. $\diamond$

Exemple - Loi de Gauss de moyenne inconnue.
Soient $X_1, \cdots, X_n$ v.a.i.i.d. suivant une loi de Gauss de moyenne θ et de variance 1. On veut tester

$$H_0 : \theta \leq \theta_0 \text{ contre } H_1 : \theta > \theta_1$$

La vraisemblance est

$$\mathcal{L}(\theta; X_1, \cdots, X_n) = \frac{1}{(\sqrt{1\pi})^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right)$$

Il faut montrer que le modèle est à rapport de vraisemblance monotone.
On peut soit montrer que la loi appartient à la famille exponentielle, soit le montrer directement. Considérons $\theta' < \theta''$,

$$\begin{aligned}\frac{\mathcal{L}(\theta''; x_1, \cdots, x_n)}{\mathcal{L}(\theta'; x_1, \cdots, x_n)} &= \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(_i - \theta'')^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta')^2\right) \\ &= \exp\left((\theta'' - \theta')\sum_{i=1}^{n} x_i + \frac{n}{2}((\theta')^2 - (\theta'')^2)\right)\end{aligned}$$

C'est une fonction monotone de $U(x_1, \cdots, x_n) = \sum_{i=1}^{n} x_i$ car $\theta'' > \theta'$. On peut appliquer le théorème précédent pour conclure que le test UPP au niveau α pour tester $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_1$ est

$$\Psi(X_1, \cdots, X_n) = \begin{cases} 1 \text{ si } \sum_{i=1}^{n} X_i > c \\ \gamma \text{ si } \sum_{i=1}^{n} X_i = c \\ 0 \text{ si } \sum_{i=1}^{n} X_i < c \end{cases}$$

où γ et c sont définies par $E_{\theta_0}(\Psi(X_1, \cdots, X_n)) = \alpha$. Comme la loi est absolument continue par rapport à la mesure de Lebesgue, on peut choisir $\gamma = 0$ et le test se ramène à un test pour la région critique

$$\text{RC} = \left\{\sum_{i=1}^{n} x_i > c\right\}$$

où c est déterminée par

$$E_{\theta_0}(\Psi(X_1, \cdots, X_n)) = P_{\theta_0}\left(\sum_{i=1}^{n} X_i > c\right) = \alpha$$

Pour $\theta = \theta_0$, $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\theta_0, n)$ donc $c = n\theta_0 + \sqrt{n}\Phi^{-1}(1 - \alpha)$.

5.4.2 Test bilatéral $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$

Il n'existe pas en général de test UPP pour le test bilatéral $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. En effet, pour être UPP, un test doit être le plus puissant pour tester $H_0 : \theta = \theta_0$ contre

$H_1 : \theta = \theta_1$ pour tout $\theta_1 \neq \theta_0$. Cependant, selon le lemme de Neyman-Pearson, la forme des tests les plus puissants diffère selon que $\theta_1 > \theta_0$ ou $\theta_1 < \theta_0$.

Soient $X_1, \cdots, X_n$ des v.a.i.i.d. de loi binomiale de paramètres n et θ et supposons que l'on veuille tester

$$H_0 : \theta = \theta_0 \text{ contre } H_1 : \theta \neq \theta_0$$

à un certain niveau α. Considérons tout d'abord, les hypothèses

$$H_0' : \theta = \theta_0 \text{ contre } H_1' : \theta = \theta_1$$

avec $\theta_1 \neq \theta_0$. Le lemme de Neyman-Pearson indique que le test UPP de H_0' contre H_1' est basé sur la statistique de test :

$$T = \frac{\mathcal{L}(\theta_1; X_1, \cdots, X_n)}{\mathcal{L}(\theta_0; X_1, \cdots, X_n)} = \left(\frac{1-\theta_0}{1-\theta_1}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum_{i=1}^n X_i}$$

Si $\theta_1 > \theta_0$, il est facile de vérifier que T est une fonction croissante de $\sum_{i=1}^n X_i$: donc un test plus puissant de H_0' contre H_1' rejetera H_0' pour de grandes valeurs de $\sum_{i=1}^n X_i$. Mais, si $\theta_1 < \theta_0$, T est une fonction décroissante de $\sum_{i=1}^n X_i$ et un test plus puissant va rejeter H_0' pour de petites valeurs de $\sum_{i=1}^n X_i$. On comprend donc qu'on ne pourra pas trouver de test UPP.

Une solution consiste à se restreindre à la classe des tests sans biais.

Définition 34 - *Un test est dit sans biais si sa puissance est toujours supérieure à son niveau, autrement dit si pour tester $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$, pour tout $\theta_1 \in \Theta_1$,*

$$E_{\theta_1}(\Psi(X_1, \cdots, X_n)) \geq \sup_{\theta \in \Theta_0} E_\theta(\Psi(X_1, \cdots, X_n))$$

c'est à dire qu'on a souvent raison quand on conlcut H_1.

Proposition 10 *Un test UPP est forcément sans biais.*

Preuve - Soit Ψ un test UPP de niveau α pour tester $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$. Soit $\psi(x_1, \cdots, x_n) = \alpha$ pour tout $(x_1, \cdots, x_n)$. ψ est de niveau α car

$$\sup_{\theta \in \Theta_0} E_\theta(\psi(X_1, \cdots, X_n)) = \sup_{\theta \in \Theta_0} \alpha = \alpha$$

Comme Ψ est UPP parmi les tests de niveau α, pour tout $\theta_1 \in \Theta_1$,

$$E_{\theta_1}(\Psi(X_1, \cdots, X_n)) = E_{\theta_1}(\psi(X_1, \cdots, X_n)) = \alpha$$

Donc la puissance de Ψ est toujours supérieure à son niveau et donc Ψ est sans biais. $\diamond$

Il est parfois possible de construire des tests uniformément plus puissants parmi les tests dans biais. Supposons qu'on veuille tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ au niveau α, on

peut construire un test UPP parmi les tests sans biais [3] en combinant des tests UPP pour tester $H_0' : \theta \leq \theta_0$ contre $H_0'' : \theta \geq \theta_0$. Plus précisement, supposons que $\Phi_1(X_1, \cdots, X_n)$ est un test UPP de niveau α_1 et $\Phi_2(X_1, \cdots, X_n)$ est un test UPP de niveau α_2 tels que $\alpha_1 + \alpha_2 = \alpha$. Alors $\Phi = \Phi_1 + \Phi_2$ sera un test de niveau α si

$$\Phi_1(X_1, \cdots, X_n) + \Phi_2(X_1, \cdots, X_n) \leq 1$$

Ainsi, en choisissant bien α_1 et α_2 il est possible d'avoir Φ un test UMPU. Le choix naturel est $\alpha_1 = \alpha_2 = \alpha/2$ mais en général, ça ne conduit pas à un test sans biais.

Soit X une variable aléatoire continue de densité

$$f(x; \theta) = \theta x^{\theta - 1} \text{ pour } 0 \leq x \leq 1$$

et supposons qu'on veuille tester

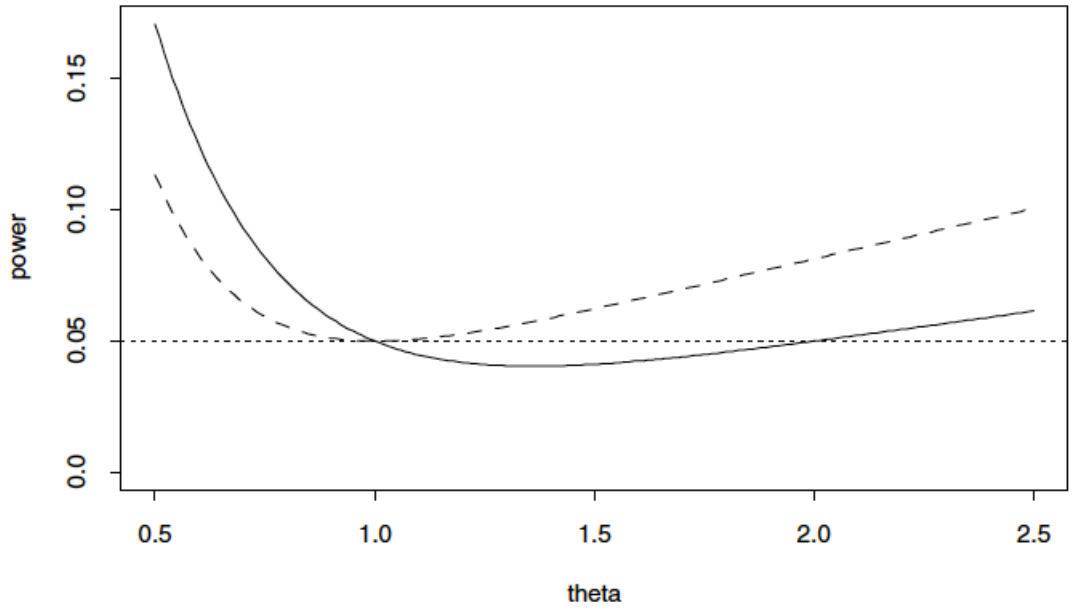
$$H_0 : \theta = 1 \text{ contre } H_1 : \theta_1 \neq 1$$

au niveau 5%. On va rejeter H_0 if $x \leq 0.025$ ou si $x \geq 0.975$; ce test est clairement de niveau $\alpha = 5\%$ puisque $P_{\theta=1}(X \leq 0.025) = P_{\theta=1}(X \geq 0.975) = 0.025$.
La fonction puissance est alors

$$\Pi(\theta) = \int_0^{0.025} \theta x^{\theta-1} dx + \int_{0.975}^1 \theta x^{\theta-1} dx = 1 + 0.025^\theta + 0.975^\theta$$

Si on évalue $\Pi(\theta)$ pour θ proche de 1 il est facile de voir que le test n'est pas sans biais. En effet $\Pi(\theta) < 0.05$ pour $1 < \theta < 2$.
Cependant, il est possible de trouver un test sans biais pour tester H_0 contre H_1. Ce test rejette H_0 si $x \leq 0.0085$ ou si $x \geq 0.9585$. Les deux fonctions puissance sont représentées ci-dessous avec la fonction puissance du test sans biais en tirets :

3. UMPU : unbiased most powerfull test



Dans cet exemple, le test sans biais a une puissance plus grande pour $\theta > 1$ mais plus petite pour $\theta < 1$. Ceci illustre le fait qu'en choisissant un test sans biais, on sacrifie de la puissance dans certaines régions.

5.5 Généralisation

Jusqu'à présent, nous avons proposé des méthodes constructives permettant d'obtenir des tests UPP (ou localement UPP) dans le cas d'un paramètre unique. Ce type de test optimal n'existe pas en général pour les modèles à plus d'un paramètre. Il existe une méthode qui permet de développer des tests dans des cas plus généraux.

Considérons le test

$$H_0 : \theta \in \Theta_0 \text{ contre } \theta \in \Theta_1$$

avec $\Theta_0 \cap \Theta_1 = \emptyset$. On va utiliser le même type d'idée que dans le lemme de Neyman-Pearson.

Définition 35 *La statistique* Λ *du rapport de vraisemblance est*

$$\Lambda = \frac{\sup_{\theta \in \Theta} \mathcal{L}(X_1, \cdots, X_n; \theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(X_1, \cdots, X_n; \theta)}.$$

Un test du rapport de vraisemblance pour $H_0 : \theta \in \Theta_0$ *contre* $\theta \in \Theta_1$ *va rejetter* H_0 *pour de grandes valeurs de* Λ.

Pour utiliser ces tests du rapport de vraisemblance, on a besoin de connaître (exactement ou approximativement) la distribution de la statistique Λ sous H_0. Dans certains cas, Λ est fonction d'une autre statistitque T dont on connait la loi et on peut alors utiliser cette statistique. Sinon on utilise un résultat limite.

Considérons le test bilatéral

$$H_0 : \theta = \theta_0 \text{ contre } \theta \neq \theta_1$$

et $X_1, \cdots, X_n$ v.a.i.i.d. de densité (ou fonction de fréquence) f dans (P_θ). La statistique Λ s'écrit

$$\Lambda_n = \prod_{i=1}^{n} \frac{f(X_i; \hat{\theta}_n)}{f(X_i; \theta_0)}$$

avec $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance.

Théorème 16 *Soient $X_1, \cdots, X_n$ v.a.i.i.d. qui admettent une densité (ou une fonction de fréquence) vérifiant les hypothèses (A1) à (A5) (du chapitre précédent) avec $I(\theta) = J(\theta)$. Si l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ satisfait un théorème de limite centrale*

$$\sqrt{n}(\hat{\theta}_n - \theta) \to Z \sim \mathcal{N}(0, 1/I(\theta))$$

alors la statistique Λ satisfait

$$2\ln(\Lambda) \to V \sim \chi^2(1)$$

quand $H_0 : \theta = \theta_0$ est vraie.

Preuve - Notons $\ell(x;\theta) = \ln(f(x;\theta))$ et $\ell'(x;\theta)$ et $\ell''(x;\theta)$ ses dérivées par rapport à θ. Sous les hypothèses du théorème, sous H_0,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to Z \sim \mathcal{N}(0, 1/I(\theta_0))$$

En prenant le $\log(\Lambda_n)$ et en faisant un développement de Taylor, on a

$$\begin{aligned}
\ln(\Lambda_n) &= \sum_{i=1}^{n} \Big(\ell(X_i; \hat{\theta}_n) - \ell(X_i; \theta_0) \Big) \\
&= (\theta_0 - \hat{\theta}_n) \sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) - \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \ell''(X_i; \theta_n^*) \\
&= -\frac{1}{2} n (\hat{\theta}_n - \theta_0)^2 \frac{1}{n} \ell''(X_i; \theta_n^*)
\end{aligned}$$

où θ_n^* est entre θ_0 et $\hat{\theta}_n$. Sous les hypothèses (A4) et (A5), du chapitre précédent, on a donc sous H_0

$$\frac{1}{n}\ell''(X_i; \theta_n^*) \to_P -E_{\theta_0}[\ell''(X_i; \theta_0)] = I(\theta_0)$$

On a aussi

$$n(\hat{\theta}_n - \theta_0)^2 \to_d \frac{V}{I(\theta_0)}$$

et on conclut avec le lemme de Slutsky. $\diamond$

Exemple - Soient $X_1, \cdots, X_n$ v.a.i.i.d. de loi de Gauss de moyenne μ et de variance σ^2 (toutes deux inconnues) et supposons qu'on veut tester

$$H_0 : \mu = \mu_0 \text{ contre } H_1 : \mu \neq \mu_0$$

Les estimateurs du maximum de vraisemblance de μ et σ^2 sont

$$\hat{\mu} = \bar{X} \text{ et } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Sous H_0, l'estimateur du maximum de vraisemblance de σ s'écrit

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$$

Et on a donc

$$\begin{aligned}
\Lambda_n &= \frac{(2\pi\sigma_0^2)^{n/2}}{(2\pi\hat{\sigma}^2)^{n/2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \frac{1}{\hat{\sigma}_0^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right) \\
&= \left(\frac{\sigma_0^2}{(\hat{\sigma}^2)}\right)^{n/2} \exp\left(\frac{1}{2\hat{\sigma}^2}n\hat{\sigma}^2 + \frac{1}{2\hat{\sigma}_0^2}n\hat{\sigma}_0^2\right) \\
&= \left(\frac{\sigma_0^2}{(\hat{\sigma}^2)}\right)^{n/2}
\end{aligned}$$

et la région critique est donc de la forme $\left\{\left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right)^{n/2} \geq c\right\}$ avec c déterminé par

$$P_{\theta_0}\left(\left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right)^{n/2} \geq c\right) = \alpha$$

car le test est pur (loi continue). La distribution de Λ n'est pas triviale ; cependant, on remarque que Λ est une fontion monotone de $\hat{\sigma}_0^2/\hat{\sigma}^2$ et

$$\begin{aligned}
\frac{\sigma_0^2}{\hat{\sigma}^2} &= \frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= 1 + \frac{n(\bar{X} - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= 1 + \frac{1}{n-1}\left(\frac{n(\bar{X} - \mu_0)^2}{S^2}\right) \\
&= 1 + \frac{T^2}{n-1}
\end{aligned}$$

où

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

et

$$T = \sqrt{n}(\bar{X} - \mu_0)/S.$$

Or on sait que sous H_0, T suit une distribution de Student à $(n-1)$ degrés de liberté et donc que T^2 suit une distribution de Fisher à 1 et $(n-1)$ degrés de liberté.

On peut généraliser le théorème précédent aux de paramètres dans $\mathbb{R}^p$ pour $p > 1$.

Théorème 17 *Soient $X_1, \cdots, X_n$ v.a.i.i.d. qui admettent une densité (ou une fonction de fréquence) vérifiant les hypothèses $A1$ à $A5$ dans le cas d'un paramètre de dimension p et avec $I(\theta) = J(\theta)$ où $\theta = (\theta_1, \cdots, \theta_p)$. Si l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ satisfait un théorème de limite centrale*

$$\sqrt{n}(\hat{\theta}_n - \theta) \to Z \sim \mathcal{N}(0, 1/I(\theta))$$

alors la statistique Λ pour tester $H_0 : \theta_1 = \theta_{10}, \cdots, \theta_r = \theta_{r0}$ satisfait

$$2\ln(\Lambda) \to V \sim \chi^2(r)$$

quand H_0 est vraie.

La preuve du théorème repose sur le fait que la log-vraisemblance peut être approchée par une fonction quadratique près de la vraie valeur du paramètre.

Exemple - Soient $X_1, \cdots X_m$ des variables aléatoires i.i.d. de loi exponentielle de paramètre λ et $Y_1, \cdots Y_n$ des variables aléatoires i.i.d. de loi exponentielle de paramètre θ. On suppose que les X_i sont indépendants des Y_i. On veut tester

$$H_0 : \lambda = \theta \text{ contre } H_1 : \lambda \neq \theta$$

Les estimateurs du maximum de vraisemblance de λ et θ sont, dans le cas général,

$$\hat{\lambda} = 1/\bar{X} \text{ et } \hat{\theta} = 1/\bar{Y}$$

et sous H_0,

$$\hat{\lambda}_0 = \hat{\theta}_0 = \left(\frac{m\bar{X} + n\bar{Y}}{m+n}\right)^{-1}$$

On a donc

$$\Lambda = \left(\frac{m}{n+m} + \frac{n}{n+m}\frac{\bar{Y}}{\bar{X}}\right)^m \left(\frac{n}{n+m} + \frac{m}{n+m}\frac{\bar{X}}{\bar{Y}}\right)^n$$

On remarque que Λ ne dépend que de $T = \bar{X}/\bar{Y}$. On peut déduire un test de T ou construire un test asymptotique avec Λ.

Exemple - Soient $(X_1, Y_1), \cdots, (X_n, Y_n)$ couples i.i.d. de variables aléatoires continues de densité jointe

$$f(x, y; \theta, \lambda, \alpha) = \frac{2\theta\lambda\alpha}{(\theta x + \lambda y + \alpha)^3} \text{ pour } x, y > 0$$

avec $\theta, \lambda, \alpha > 0$. Les densités marginales de X_i et Y_i sont

$$f_X(x; \theta, \alpha) = \frac{\theta\alpha}{(\theta x + \alpha)^2} \text{ pour } x > 0$$

$$f_Y(y; \lambda, \alpha) = \frac{\lambda\alpha}{(\lambda y + \alpha)^2} \text{ pour } y > 0$$

On veut tester $H_0 : \theta = \lambda$.
On peut reparamétriser ce problème de différentes façons. Par exemple, on peut définir $\eta_1 = \theta - \lambda$, $\eta_2 = \theta$ et $\eta_3 = \alpha$ ou $\eta_1 = \theta/\lambda$. On exprime alors H_0 en fonction de η_1 et on s'attend à ce que la statistique de test du rapport de vraisemblance suive approximativement une loi du χ^2 à 1 degré de liberté pour n grand.

5.6 Autres tests basés sur le maximum de vraisemblance

5.6.1 Test de Wald

$$H_0 : \theta = \theta_0 \text{ contre } \theta \neq \theta_0$$

Dans le test de Wald, l'estimateur du maximum de vraisemblance $\hat{\theta}$ du paramètre θ est comparé à la valeur θ_0, sous l'hypothèse que la différence est distribuée approximativement selon une loi de Gauss. En pratique le carré de la différence est comparé à un seuil de la loi du chi 2. Dans le cas univarié, la statistique de Wald est

$$\frac{(\widehat{\theta} - \theta_0)^2}{\text{var}(\hat{\theta})}$$

Si on compare la différence à un quantile de la loi de Gauss, la statistique de test est

$$\frac{\widehat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$$

où $\text{se}(\widehat{\theta})$ est l'écart-type de l'estimateur du maximum de vraisemblance. Un estimateur raisonnable de cet écart-type est donné par $\frac{1}{\sqrt{I_n(MLE)}}$, où I_n est l'information de Fisher du paramètre.
Dans le cas univarié, un test sur plusieurs paramètres simultanément est réalisé en utilisant une matrice de variance. Par exemple, on utilise ce test pour une variable catégorielle recodée en plusieurs variables dichotomiques.

5.6.2 Score test (ou test des multiplicateurs de Lagrange)

Le test des multiplicateurs de Lagrange utilise le fait que si l'hypothèse nulle est fausse alors, le gradient de la log vraisemblance ne doit pas être proche de 0. Plus précisément, en notant $S_i(\theta)$ le gradient de $\log(f(X_i;\theta)$ en fonction de θ, alors sous H_0, on a

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\hat{\theta})$$

tend en loi vers une variable de loi de Gauss centrée où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ sous H_0.
Comme pour le test de Wald, on rejette H_0 pour les grandes valeurs de $S_n = \frac{1}{n} \left(\sum_{i=1}^{n} S_i(\hat{\theta}) \right)^2 / I_n(\theta)$ et, sous H_0 on a S_n tend en loi vers une variable qui suit un chi 2 à r degrés de liberté ($r = dim(\theta)$).

5.7 Tests classiques

5.7.1 Tests paramétriques pour des moyennes, des variances ou des corrélations

– Test de Student : comparaison de moyennes dans le cadre de la loi de Gauss ou pour des grands échantillons.

– Test de Fisher : comparaison de moyennes dans le cadre de la loi de Gauss. Ce test permet notamment de tester l'égalité des variances de 2 échantillons.
– Analyse de la variance (variance inter-classe/ variance intra-classe)
– Test de Pearson : test de corrélation. $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$.

5.7.2 Tests non paramétriques pour des moyennes , des variances ou des corrélations

On les utilise quand on a de petits échantillons dont on ne connait pas la distribution.
– Test du signe, test des signes et rangs de Wilcoxon ou mann-Whitney Wilcoxon
– Test de Kruskal-Wallis
– test de Spearman, Test du τ de Kendall

Test des signes et rangs de Wilcoxon

Hypothèses : Nous supposons que la distribution de la variable dans la population est *symétrique* et *continue*.
Etant donné un échantillon de n mesures indépendantes, nous pouvons au lieu de noter seulement les signes des écarts à la médiane spécifiée dans H_0, relever aussi la grandeur de chaque écart. Si H_0 est vraie, les écarts d'un grandeur donnée ont autant de chance, pour une distribution symétrique, d'être positifs que négatifs ; et une valeur dépassant θ de 4 ou 5 unités a la même probabilité d'être observée qu'une valeur inférieure à θ de 4 à 5 unités. C'est sur cette idée que se base le test des **signes et rangs** de Wilcoxon [4]
Reprenons l'exemple de la taille des poissons. En notant θ la longueur médiane, les hypothèses de test sont

$$H_0 : \theta = 220 \text{ contre } H_1 : \theta \neq 220$$

Nous rappelons que nous avions observé l'échantillon suivant :

126	142	156	228	245	246	370	419	433	454	478	503

TABLE 5.2 – Longueur de 12 poissons

Formulation et postulat - Nous rangeons par ordre croissant les écarts à 220 (écarts en valeurs absolue), puis nous associons à chaque écart son signe (c'est à dire un signe + si l'observation correspondante est supérieure à la médiane spécifiée sous H_0 et un signe - sinon). On calculons la somme S_p des rangs des écarts positifs et la somme S_n des rangs des écarts négatifs. Si H_0 est vraie, on s'attend à ce que ces deux sommes soit presque égales. La statistique de test est la plus petite des deux sommes. Pour construire la région de rejet, nous utilisons la table des signes et rangs de Wilcoxon.

Le problème des ex aequo - Nous avons supposé que la distribution de la variable d'intérêt est continue dans la population. Or pour une distribution continue, la probabilité d'obtenir des observations égales est nulle de même que celle d'obtenir des observations égales à la médiane de la population. Cependant, en pratique, les observations ne sont pas strictement continues (arrondis ou précision limité des appareils de mesure). Si une ou plusieurs valeurs coïncident avec la médiane spécifiée sous H_0, nous leur attribuons le rang 0.

4. En anglais, on dit *signed ranks test* ce qui est aussi traduit **test des rangs signés**.

Si un grand échantillon comporte des valeurs égales à la médiane sous H_0 ou des ex aequo, on modifie Z de la façon suivante

$$Z = \frac{S - \frac{n(n+1)}{4} - d_0(d_0+1)}{\sqrt{n(n+1)(2n+1)/24 - d_0(d_0+1)(2d_0+1)/24 - \sum_{i=1}^{n_{ge}}(d_i^3 - d_i)/48}}$$

où d_0 est le nombre de valeurs égales à la médiane spécifiée sous H_0, n_{ge} est le nombre de groupes d'ex eaquo et d_i le nombre d'ex aequo dans le $i$ème groupe.

5.7.3 Test d'adéquation ou de comparaison de distribution

Test d'adéquation [5].
- Tets du χ^2 : loi discrête
- Test de Kolmogorov, Cramer-von Mises, : loi continue quelconque
- Test de Shapiro-Wilk : loi normale

Test du chi2

Pour une distribution discrète on utilise le test d'adéquation du chi 2.
Exemple : On suppose que le nombre de pièces défectueuses produites en un jour par une machine suit une loi de Poisson, de paramètre inconnu. Rappelons les caractéristiques de cette loi : si une variable aléatoire X suit une loi de Poisson de paramètre λ , alors $E(X) = \lambda$, $Var(X) = \lambda$, et pour tout $k \in \mathbb{N}$,

$$P(X = k) = \frac{\lambda^k e^\lambda}{k!}$$

On pose les hypothèses de test

$$H_0 : X \text{suit une loi de Poisson de paramètre 1.2 } \textit{contre } H_1 : X \text{ suit une autre loi}$$

On observe 100 jours de production de cette machine, voici les résultats, regroupés en 5 classes.

Nombre de pièces défectueuses	0	1	2	3	4 et plus
Nombre d'observations	27	41	21	7	4
Fréquence empirique	0,27	0,41	0,21	0,07	0,04

‘

On utilise une statistique de test du chi2 donnée par

$$T = \sum_{k=1}^{K} \frac{(\hat{n} f_k - n p_k)^2}{n p_k}$$

avec f_k et p_k les fréquences empirique et théorique de la classe k et K le nombre de classes. T suit une loi du chi 2 à $K-1$ degrés de liberté.
Pour l'exemple considéré, les fréquences théoriques sont données ci-dessous.

Nombre de pièces défectueuses	0	1	2	3	4 et plus
Fréquence théorique	0,30	0,36	0,22	0,09	0,03

La statistique de test $T = 0,0112$; or d'après la table du chi 2 on rejette H_0 au risque 5% si $T > 5.99$.

5. En anglais : *Goodness-of-fit tests*

Test de Kolmogorov-Smirnov

Soient $x_1, \cdots, x_n$, n réalisations d'une variable aléatoire X. On se demande s'il est raisonnable de supposer que X suit la loi caractérisée par la fonction de répartition F et on pose les hypothèses de tests :

$$H_0 : X \text{suit la loi } F \text{ contre } H_1 : X \text{suit une autre loi}$$

On propose de construire une statistique de test basée sur la distance entre les fonction F et une estimation de la fonction de répartition de X obtenue à partir des observations.

5.7.4 Estimer la fonction de répartition

On construit naturellement un estimateur de la fonction de répartition de X d'après l'équation (**??**).

$$F_n(x) = \frac{\text{Card}(\{i | x_i \leq x\}}{n})$$

Exercice - On considère les données d'un essai visant à déterminer la solidité d'une corde d'escalade. Un morceau de 1 m corde est mis sous tension jusqu'à cassure. On se demande si la corde pour casser à n'importe endroit. On obtient les résultats suivants :

0.1	0.4	0.4	0.6	0.7	0.7	0.8	0.9	0.9	0.9

1. Tracer l'estimation de la fonction de répartition.
2. Ajouter sur le graphique la fonction de répartition théorique pour ce problème.

Kolmogorov propose d'utiliser la statistique de test suivante :

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

La loi de cette statistique de test est données dans la table de Kolmogorov.
Si on considère de nouveau les données de l'exercice, on obtient

observations	0.1	0.4	0.4	0.6	0.7	0.7	0.8	0.9	0.9	0.9
F_n	1/10	3/10	3/10	4/10	6/10	6/10	7/10	1	1	1
F	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$F - F_n$	0	0.1	0	0	0.1	0	0.2	0.1	0	

d'où $D_n = 0.2$ or on rejette H_0 au risque $\alpha = 5\%$ si $D_n > 0.369$. Ainsi ici on peut supposer que les observations sont issues d'une loi uniforme sur $[0, 1]$.

Cas de la loi normale

La version du test de Kolmogorov adaptée pour la loi de Gauss s'appelle le test de Lilliefors. Ce test est peu puissant et on lui préfère le test de Shapiro-Wilk. Ce dernier est basé sur une comparaison de deux estimateurs de la variance qui ne peuvent conduire à la même estimation que si les observations sont issues d'une loi de Gauss.
Pour vérifier qu'une série d'observation suit une loi normale, on peut en première approche utiliser une méthode graphique : la droite de Henry (*quantile-quantile plot* ou *qqplot*).

Soit $\{x_1, \cdots, x_n\}$ une suite d'observations. Si cette suite constitue une suite de réalisation d'une variable gaussienne, alors les points de coordonnées $(x_i, \Phi^{-1}((i-1/2)/n))$ sont alignés sur la droite d'équation

$$y = \frac{x - \overline{x}}{\hat{\sigma}}$$

Cette droite est appelée *droite de Henry*.

Comparaison de deux distributions

On peut généraliser le test de Kolmogorov au cas de deux échantillons afin de comparer leurs distributions. Le test s'appelle alors test de Kolmogorov-Smirnov. L'hypothèse nulle est que les deux échantillons proviennent de la même distribution ; l'alternative est qu'ils proviennent de distributions ayant des répartitions différentes. On ne spécifie aucune forme particulière pour leur différence. Et la statistique de test est basée sur un écart en valeur absolue entre la fonctions de répartition empiriques des deux suites d'observations.
Il existe d'autres tests permettant de comparer des distributions. Par exemple, le test de Cramér-von Mises repose sur la somme des carrés des écarts en valeurs absolue entre les deux fonctions de répartition. En notant, S_d^2 cette somme, la statistique de test est

$$T = \frac{nmS_d^2}{n+m}$$

avec m et n les nombres d'observation des deux groupes.
Pour un test bilatéral, on rejette H_0 au niveau de signification 5% (resp. 1%) si T est supérieur à 0.461 (resp. 0.743).
Le test de Cramér-von Mises est souvent plus puissant que le test de Kolmogorov-Smirnov et il est plus facile à utiliser grâce à la bonne approximation qui évite le recours à des tables.

Chapitre 6

Estimation par intervalles

On a vu dans le chapitre sur l'estimation qu'une statistique n'est en géréral pas exactement égale à la valeur du paramètre qu'elle est censée estimer (si la loi de la statistique a une densité, cet événement est même de probabilité nulle). Donc, il est important que toute procédure d'estimation soit accompagnée d'une indication sur la précision de l'estimation.

6.1 Exemple

On considère un n-échantillon $X_1, \cdots, X_n$ de la loi gaussienne de paramètres (μ, σ^2). On suppose que σ^2 est connu. On estime μ. On a déjà vu toutes les qualités de l'estimateur $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. On sait que la variable $\sqrt{n}\sigma^{-1}(\bar{X}_n - \mu)$ est gaussienne standard. Sa loi ne dépend pas du paramètre μ, ce qu'on a déjà utilisé pour construire un test de niveau α pour tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$ ou contre $H_1 : \mu > \mu_0$. On sait par exemple que pour tout μ

$$P_\mu\left(\mu - \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \tag{6.1}$$

Or ceci peut aussi se lire

$$P_\mu\left(\bar{X}_n - \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \tag{6.2}$$

ce qui change tout car dans (6.2), l'intervalle est aléatoire (la probabilité qu'il contienne le paramètre est $1-\alpha$), tandis que dans (6.1), l'intervalle est fixé et la variable aléatoire $\bar{X}_n$ a une probabilité $1-\alpha$ de se trouver dedans.

On dit que

$$I(X_1, \cdots, X_n)) = [\bar{X}_n - \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}}, \bar{X}_n + \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}}]$$

est un intervalle de confiance de niveau $1-\alpha$ pour le paramètre μ.

Remarque - On a choisi un intervalle symétrique : est ce nécessaire ?
Dans une certaine mesure la réponse est oui. En effet, soit U une variable gaussienne

standard. Il est facile de vérifier que, $P(U \in [x, x + 2\Phi^{-1}(1-\alpha/2)])$ est maximale (et vaut $1-\alpha$) quand $x = -\Phi^{-1}(1-\alpha/2)$. On en déduit que parmi les intervalles tels que $P(U \in [x,y]) = 1-\alpha$, le plus court est $[-\Phi^{-1}(1-\alpha/2), \Phi^{-1}(1-\alpha/2)]$. Autrement dit l'intervalle proposé en (6.2) est le plus précis des intervalles de confiance de niveau $1-\alpha$ pour le paramètre μ. Ceci dit, on peut avoir d'autres critères que la précision pour choisir l'intervalle de confiance. On peut vouloir par exemple une demi-droite de confiance si le seul souci est de garantir que le paramètre est suffisamment grand (ou suffisamment petit).

6.2 Méthode générale pour construire des intervalles de confiance

On cherche à estimer $\theta \in \Theta$ à partir de $(X_1, \cdots, X_n)$, variables indépendantes et de même loi P_θ. Ici, $\Theta \subset \mathbb{R}^d$.

Définition 36 *On appelle fonction pivotale, une fonction $h(\theta; X_1, \cdots, X_n)$ à valeurs dans $\mathbb{R}^k$ possédant les propriétés suivantes :*
- *Propriété 1 : quelque soit $\theta \in \Theta$, la fonction $h(\theta; X_1, \cdots, X_n)$ est mesurable.*
- *Propriété 2 : la loi de $h(\theta; X_1, \cdots, X_n)$ ne dépend pas de θ.*

Il est facile de s'appuyer sur une fonction pivotale pour construire des intervalles de confiance de la façon suivante. Supposons qu'il existe B, un borélien de $\mathbb{R}^k$, tel que

$$P_\theta(h(\theta; X_1, \cdots, X_n) \in B) = 1-\alpha, \forall \theta \in \Theta \tag{6.3}$$

on défini alors la région de confiance

$$I(X_1, \cdots, X_n) = \{\theta \in \Theta | h(\theta; X_1, \cdots, X_n) \in B\}$$

Comme dans l'exemple introductif, l'ensemble $l(X_1, \cdots, X_n)$ est aléatoire et a, sous P_θ, une probabilité $1-\alpha$ de contenir θ. Dans l'exemple on avait

$$h(\mu; X_1, \cdots, X_n) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

et le borélien B était l'intervalle $[-\Phi^{-1}(1-\alpha/2), \Phi^{-1}(1-\alpha/2)]$

Remarques -
- L'égalité (6.3) s'écrit aussi $\pi(B) = 1-\alpha$ où π désigne la loi (ne dépendant pas de θ, d'après la Propriété 2) transportée de P_θ^n par $h(\theta, x)$. Bien sûr, un tel B peut ne pas exister. On recherchera alors un borélien tel que $\pi(B) > 1-\alpha$.
- On peut aussi bien se servir de la fonction pivotale pour construire un test de l'hypothèse $\theta = \theta_0$. Pour cela, il suffit de prendre comme région de rejet l'ensemble des $(X_1, \cdots, X_n)$ tels que $h(\theta_0; X_1, \cdots, X_n) \notin B$.

6.3 Lien avec les tests

On peut construire des intervalles de confiance à l'aide des tests. La région d'acceptation du test définit un intervalle de confiance.

Théorème 18 *Soit $\mathbf{X} = X_1, \cdots, X_n$ un n-échantillon de loi $(P_\theta)_{\theta\in\Theta}$. Pour chaque $\theta_0 \in \Theta$ on considère le problème de tester $H_0 : \theta = \theta_0$ au niveau α et soit $A(\theta_0) = RC(\theta_0)^c$ sa région d'acceptation.*
Pour chaque $(x_1, \cdots, x_n) \in \mathbb{R}^n$ on définit $C(x)$ par

$$C(x_1, \cdots, x_n) = \{\theta_0 \in \Theta : \mathbf{x} \in A(\theta_0)\}.$$

Alors $C(X_1, \cdots, X_n)$ est une région de confiance pour θ de niveau $1 - \alpha$.
Réciproquement, soit $C(X)$ une région de confiance pour θ de niveau $1 - \alpha$. Pour tout $\theta_0 \in \Theta$ on définit

$$A(\theta_0) = \{(x_1, \cdots, x_n) : \theta_0 \in C(\mathbf{x})\}$$

Alors $A(\theta_0)$ est la région d'acceptation d'un tests de niveau α pour $H_0 : \theta = \theta_0$.

Preuve - Comme $A(\theta_0)$ est la région d'acceptation d'une test de niveau α, on a

$$P_{\theta_0}(\mathbf{X} \notin A(\theta_0)) \leq \alpha$$

ou encore

$$P_{\theta_0}(\mathbf{X} \in A(\theta_0)) \geq 1 - \alpha$$

Comme θ_0 est quelconque, on peut écrire θ à la place de θ_0. Alors, d'après l'inégalité précédente,

$$P_\theta(\theta \in C(\mathbf{X})) = P_{\theta_0}(\mathbf{X} \in A(\theta_0)) \geq 1 - \alpha$$

donc C est une région de confiance pour θ de niveau $1 - \alpha$.
Par ailleur, on voit que l'erreur de 1ère espèce pour H_0 avec pour région d'acceptation $A(\theta_0)$ est

$$P_{\theta_0}(\mathbf{X} \notin A(\theta_0)) = P_{\theta_0}(\theta_0 \notin C(\mathbf{X})) \leq \alpha$$

donc le test est de niveau α sous H_0. $\diamond$