République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

LAOUAR Sihem

Titre :

Information de Fisher

Membres du Comité d'Examen :

Dr.	**SAYAH Abdallah**	UMKB	Président
Pr.	**NECIR Abdelhakim**	UMKB	Encadreur
Dr.	**TOUBA Sounia**	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie ce modeste travail à mes chers parents.

REMERCIEMENTS

Avant tout je remercie ***Dieu*** *le Tout-Puissant de m'avoir accordé la volonté*

ainsi que la patience pour accomplir ce modeste travail.

Je tiens aussi à remercier très chaleureusement mon encadreur

mémoire Monsieur **Pr. Necir Abdelhakim**

pour ses précieux conseils et son orientation tout au long de mes recherches.

Un grand remerciement également à tous les membres du jury

d'avoir participé à la commission des examinateurs en vue d'une juste et prompte.

Je remercie ***mes chers parents***, ***mes frères*** *et* ***soeurs***

et sans oublier ***mes précieux amis*** *pour tout le soutien, les conseils*

ainsi que la confiance qu'ils ont pu me donner tout au long de mon travail.

Je remercie également tous le corps enseignant

pour la formation et l'enseignement de qualité qu'ils nous ont donné.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'objectif d'une procédure d'estimation est de révéler de l'information sur une caractéristique de la population à partir d'un échantillon. L'échantillon $X_1, ..., X_n$ contient une certaine information vis-à-vis d'un paramètre inconnu θ que l'on cherche à estimer. Lorsque la taille n de l'échantillon $(X_1, ..., X_n)$ est très grande et pour inférer sur les caractéristiques d'une propriété de X, le statisticien utilise des fonctions mesurables de l'échantillon T_n sont appelées statistiques, qui résument les observations tout en conservant l'intégralité de l'information sur le paramètre inconnu θ : c'est les statistiques exhaustives. De plus, pour mesurer l'information fournie par une statistique sur un modèle paramétrique dominé au sujet d'un paramètre, le statisticien définit une quantité mathématique mesurant l'information sur le paramètre apportée par l'échantillon est exprimée par l'information de Fisher.

Dans ce travail, on s'intéresse à deux notions de base de la statistique mathématique, à savoir l'estimation paramétrique ponctuelle et l'exhaustivité.

Le contenu de ce mémoire est réparti en deux chapitres comme suit :

Chapitre 1 : rassemble des rappels sur des notions de la base de statistique particulièrement de l'estimation paramétrique ponctuelle qui seront essentiels dans le deuxième chapitre. Avant de définir la quantité d'information de Fisher, nous avons étudié les propriétés des estimateurs ponctuels et une méthode d'estimation comme-ci celle dite du "maximum de vraisemblance" dans le cas des données complètes.

Chapitre 2 : nous définissons le concept principal en inférence statistique : la quantité

d'information au sens de Fisher dans le cas des données complètes puis nous abordons la notion d'hessienne. On présente les définitions dans le cas où le paramètre inconnu θ est un scalaire, puis nous avons étendu ces définitions au cas vectoriel. En outre, on expose l'intérêt et le lien entre l'information de Fisher et la statistique et enfin on définit l'information de Fisher au cas des données incomplètes (censure et troncature).

Chapitre 1

Estimastion paramétrique ponctuelle

1.1 Généralités

La procédure qui utilise des informations obtenues à partir de l'échantillon qui permet de déduire des résultats concernant l'ensemble de la population est appelée estimation. L'estimation consiste à donner des valeurs approchées aux paramètres d'une population (espérance, variance, etc.) à l'aide d'un échantillon de n observations issues de cette population. On peut se tromper sur la valeur exacte, mais on donne la "meilleure valeur" possible que l'on peut supposer. Pour plusieurs informations détaillées sur l'exhaustivité on réfère la lecture à [8], [12] et [14], ...

1.1.1 Estimateur ponctuel

Supposant qu'on voudrait estimer un paramètre θ d'une population (cela peut être sa moyenne μ, son écart-type σ, une proportion p). Un estimateur de θ est une statistique T (donc une fonction mesurable de $X_1, ..., X_n$) dont la réalisation est considérée comme une "bonne valeur" du paramètre θ. On parle d'estimation de θ associée à cet estimateur la valeur observée lors de l'expérience.

Exemple 1.1.1 *Pour estimer l'espérance $E(X)$ de la loi de X, un estimateur naturel est*

la moyenne empirique $\bar{X}$ qui produit une estimation $\bar{x}$, moyenne descriptive de la série des valeurs observées.

1.1.2 Propriétés des estimateurs ponctuels

Le contexte classique est celui d'un échantillon de n variables aléatoires indépendantes $(X_1, X_2, ..., X_n)$ suivant la même loi, à savoir celle d'une variable aléatoire X (variable dite parente), de densité de probabilité $f(x, \theta)$ (resp de loi de probabilité $P(X = x)$ dans le cas discret), θ étant un paramètre unidimensionnel ou multidimensionnel dont on se propose d'évaluer l'estimation $\hat{\theta}$ à travers une statistique $\theta = T(X_1, X_2, ..., X_n)$ (dite estimateur), [3].

Qualités d'un bon estimateur

Afin de choisir entre plusieurs estimateurs possibles d'un même paramètre il faut définir les qualités demandées d'un estimateur.
Soit θ le paramètre à estimer et T un estimateur, c'est-à-dire une fonction des X_i à valeurs dans un domaine acceptable pour θ.

Définition 1.1.1 *Un estimateur T est dit **convergent** si $E(T)$ tend vers θ lorsque n tend vers l'infini.*
*Il sera dit **consistant** si T converge en probabilité vers θ lorsque n tend vers l'infini, [5].*

Théorème 1.1.1 *Si T est convergent et de variance tendant vers 0 lorsque n tend vers l'infini alors T est **consistant**, [5].*

Preuve. On a, pour tous réels θ et $\alpha > 0$,

$$|T - \theta| > \alpha \Rightarrow |T - E(T)| > \alpha - |\theta - E(T)| .$$

Si la limite $E(T) = \theta$, alors à partir d'un certain rang N, on a : $|\theta - E(T)| < \frac{\alpha}{2}$. Ainsi

$$\begin{aligned} P(|T-\theta| > \alpha) &\leq p(|T - E(T)| > \alpha - |\theta - E(T)|) \\ &\leq p(|T - E(T)| > \frac{\alpha}{2}) \\ &\leq \frac{4}{\alpha^2} Var(T) \text{ (par Bienaymé-Chebishev),} \end{aligned}$$

borne supérieure qui tend vers 0 lorsque n tend vers l'infini. ■

Définition 1.1.2 *On appelle* ***biais*** *de T pour θ la valeur*

$$b_\theta(T) = E(T) - \theta.$$

Un estimateur T est dit ***sans biais*** *si $E(T) = \theta$, voir la figure (1.1)*

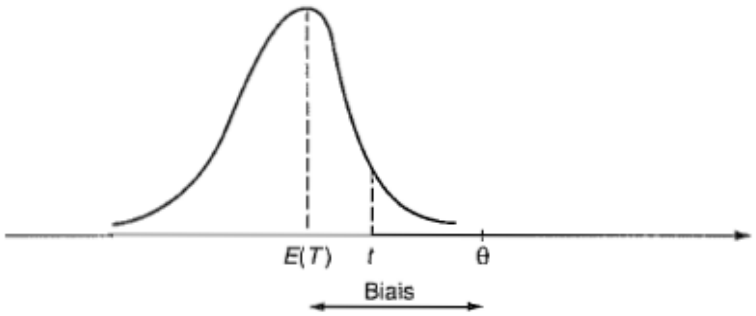

FIG. 1.1 – Biais

Comparaison des estimateurs

Le risque quadratique est un bon critère de comparaison d'estimateurs.

Définition 1.1.3 *[5] La qualité d'un estimateur se mesure également par l'****erreur quadratique moyenne*** *(ou risque quadratique) définie par :*

$$R(T;\theta) = \|T - \theta\|^2 = E((T-\theta)^2).$$

On peut alors comparer deux estimateurs.

Définition 1.1.4 *On dit que T_1 est meilleur estimateur que T_2 :*

si pour tout $\theta \in \Theta$, on a

$$R(T_1;\theta) \leq R(T_2;\theta),$$

et s'il existe un $\theta' \in \Theta$ tel que

$$R(T_1;\theta') < R(T_2;\theta'),$$

L'erreur quadratique moyenne de T se décompose en deux termes :

Théorème 1.1.2 *[5]Soit T un estimateur du paramètre θ à étudier. On a*

$$E((T-\theta)^2) = Var(T) + [E(T)-\theta]^2.$$

Preuve. Il est clair que

$$E([T-\theta]^2) = E([T-E(T)+E(T)-\theta]^2),$$

et par la linéarité de l'espérance mathématique

$$E([T-\theta]^2) = E([T-E(T)]^2) + E([E(T)-\theta]^2) + 2E([T-E(T)][E(T)-\theta]).$$

D'une part : $E([E(T)-\theta]^2) = [E(T)-\theta]^2$, et d'autre part

$$E([T-E(T)][E(T)-\theta]) = [E(T)-\theta]\, E([T-E(T)] = 0.$$

(car $E(T)-\theta$ est une constante et que $E[T-E(T)] = 0$), il vient

$$E([T-\theta]^2) = Var(T) + (E(T)-\theta)^2.$$

■

Cette décomposition permet de se ramener à une discussion sur la variance pour les estimateurs sans biais de θ.

D'après la décomposition biais-variance, la comparaison d'estimateurs sans biais revient à la comparaison de leurs variances ; on parle alors d'efficacité.

Définition 1.1.5 *L'estimateur T_1 est dit plus* ***efficace*** *que T_2 s'il est meilleur au sens de variance :*

$$Var_\theta(T_1) \leq Var_\theta(T_2), \text{ pour tout } \theta \in \Theta,$$
$$Var_{\theta'}(T_1) < Var_{\theta'}(T_2), \text{ s'il existe } \theta' \in \Theta.$$

On dit que l'estimateur sans biais T_1 est de variance minimal si $Var_\theta(T_1) \leq Var_\theta(T_2)$ pout tout estimateurs sans biais T_2 et pout tout $\theta \in \Theta$.

1.1.3 Quelques estimateurs

1. $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais de la moyenne μ. Son estimation $\bar{x}$ est la moyenne observée dans une réalisation de l'échantillon.
2. $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ est un estimateur consistant de σ^2 (mais biaisé).
3. $S^{*2} = \frac{n}{n-1}S^2$ est un estimateur sans biais et consistant de σ^2 où σ est l'écart-type observé dans une réalisation de l'échantillon.
4. Si p est la fréquence d'un caractère, F constitue un estimateur sans biais et consistant de p. Son estimation est notée f.

Remarque 1.1.1 *Si la moyenne μ de X est connue, $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ est un meilleur estimateur que S^{*2}, [5]*

En effet

$$Var(T) = \frac{1}{n^2} Var(\sum_{i=1}^{n} (X_i - \mu)^2) = \frac{1}{n} Var\left[(X_i - \mu)^2\right]$$
$$= \frac{1}{n}\left[E(X-\mu) - \left[E(X-\mu)^2\right]^2\right] = \frac{1}{n}\left(\mu_4 - \sigma^4\right),$$

et

$$Var(S^{*2}) = \left(\frac{n}{n-1}\right)^2 Var(S^2) = \left(\frac{n}{n-1}\right)^2 \frac{n-1}{n^3}\left[(n-1)\,\mu_4 - (n-3)\,\sigma^4\right]$$
$$= \frac{1}{n}\left[\mu_4 - \frac{n-3}{n-1}\sigma^4\right],$$

Donc

$$Var(T) < Var(S^{*2}).$$

1.1.4 Estimation par la méthode du maximum de vraisemblance (Données complètes)

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.

Soit X une variable aléatoire réelle de loi paramétrique (discrète ou continue), dont on veut estimer le paramètre $\theta \in \Theta$. Alors on définit une fonction f telle que :

$$f(x;\theta) = \begin{cases} f(x) \text{ si } X \text{ est une v.a continue de densité } f. \\ P(X = x) \text{ si } X \text{ est une v.a discrète de probabilité ponctuelle } P. \end{cases}$$

Définition 1.1.6 *On appelle fonction de vraisemblance de θ pour une réalisation $(x_1, ..., x_n)$*

d'un échantillon, la fonction de θ :

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Définition 1.1.7 *[5]La méthode consistant à estimer par la valeur qui maximise L (vraisemblance) s'appelle méthode du maximum de vraisemblance.*

$$\hat{\theta} = \left\{ \theta / L(\hat{\theta}) = \sup_{\theta} L(\theta) \right\},$$

Ceci est un problème d'optimisation. On utilise généralement le fait que si L es dérivable et si L admet un maximum global en une valeur, alors la dérivée première s'annule en et que la dérivée seconde est négative.

Réciproquement, si la dérivée première s'annule en $\theta = \hat{\theta}$ et que la dérivée seconde est négative en $\theta = \hat{\theta}$ alors $\hat{\theta}$ est un maximum local (non global) de $L(x, ..., x_i, ..., x_n; \theta)$. Il est alors nécessaire de vérifier qu'il s'agit bien d'un maximum global. La vraisemblance étant positive et le logarithme népérien une fonction croissante, il est équivalent et souvent plus simple de maximiser le logarithme népérien de la vraisemblance (le produit se transforme en somme, ce qui est plus simple à dériver).

Ainsi en pratique :

1. La condition nécessaire :

$$\frac{\partial L(x_1, ..., x_n; \theta)}{\partial \theta} = 0 \text{ ou } \frac{\partial \ln L(x_1, ..., x_n; \theta)}{\partial \theta} = 0,$$

   permet de trouver la valeur $\hat{\theta}$.

2. $\theta = \hat{\theta}$ est un maximum local si la condition suffisante est remplie au point critique :

$$\frac{\partial^2 L(x_1, ..., x_n; \theta)}{\partial \theta^2}(\hat{\theta}) \leq 0 \text{ ou } \frac{\partial^2 \ln L(x_1, ..., x_n; \theta)}{\partial \theta^2}(\hat{\theta}) \leq 0.$$

Exemple 1.1.2 *Avec une loi discrète*

On souhaite estimer le paramètre λ d'une loi de Poisson à partir d'un n-échantillon. On a

$$f(x;\lambda) = P_\lambda(X = x) = \exp(-\lambda)\frac{\lambda^x}{x!}.$$

La fonction de vraisemblance s'écrit :

$$L(x_1, ..., x_n; \lambda) = \prod_{i=1}^{n} \exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = \exp(-\lambda n)\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!},$$

Il est plus simple d'utiliser le logarithme, la vraisemblance étant positive :

$$\ln L(x_1, ..., x_n; \lambda) = \ln\exp(-\lambda n) + \ln\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!} = -\lambda n + \ln\lambda\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\ln(x_i!),$$

la dérivée première

$$\frac{\partial \ln L(x_1, ..., x_n; \theta)}{\partial\theta} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i,$$

s'annule pour $\hat{\lambda} = n^{-1}\sum_{i=1}^{n} x_i$. La dérivée seconde

$$\frac{\partial^2 \ln L(x_1, ..., x_n; \theta)}{\partial\theta^2} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i,$$

est toujours négative ou nulle. Ainsi l'estimation donnée par $\Lambda = n^{-1}\sum_{i=1}^{n} X_i = \overline{X}$ conduit à un estimateur du maximum de vraisemblance égal $\hat{\lambda} = \overline{x}$. Il est normal de retrouver la moyenne empirique qui est le meilleur estimateur possible pour le paramètre (qui représente aussi l'espérance d'une loi de Poisson), [12]

Exemple 1.1.3 *Avec une loi continue*

On souhaite estimer les paramètres μ et σ d'une loi normale à partir d'un n-échantillon.

La loi normale $\mathcal{N}(\mu;\sigma)$ a pour fonction densité :

$$f(x;\mu,\sigma) = f_{(\mu,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{1.1}$$

Ecrivons la fonction de vraisemblance pour une réalisation d'un échantillon de n variables indépendantes :

$$f(x_1,...,x_n;\mu,\sigma) = \prod_{i=1}^{n} f(x_i;\mu,\sigma) = \left(\frac{1}{2\pi\sigma}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right), \tag{1.2}$$

or (théorème de König)

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2,$$

où $\bar{x}$ représente la moyenne de l'échantillon. Ainsi la fonction de vraisemblance peut être écrite sous la forme :

$$L(x_1,...,x_n;\mu,\sigma) = \left(\frac{1}{2\pi\sigma}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right),$$

et sa dérivée par rapport à μ est

$$\frac{\partial \ln L}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\ln(\frac{1}{2\pi\sigma})^{\frac{n}{2}} - \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)\right) = 0 - \frac{-2n(\bar{x}-\mu)}{2\sigma^2}.$$

On obtient donc l'estimateur par le maximum de vraisemblance de l'espérance :

$$\hat{\mu} = \overline{x} = n^{-1}\sum_{i=1}^{n} x_i = \overline{x}$$

Pour le second paramètre, on calcule

$$\frac{\partial \ln L}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left(\ln(\frac{1}{2\pi\sigma^2}\right) - \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{\sigma^3},$$

Donc

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

On vérifie que c'est bien des maxima locaux :

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} \leq 0$$

$$\frac{\partial^2 \ln L}{\partial \sigma^2} = -\frac{n}{\sigma^2} - \frac{3}{\sigma^4}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right),$$

Au point $\hat{\sigma}$:

$$\frac{\partial^2 \ln L}{\partial \sigma^2}(\hat{\sigma}) = -\frac{n}{\sigma^2} - \frac{3}{\sigma^4}\left(n\hat{\sigma}^2 + n(\bar{x}-\mu)^2\right) \leq 0.$$

La méthode fournit un estimateur non biaisé de la moyenne $(E(\hat{\mu}) = \mu)$ *mais par contre, l'estimateur de la variance est biaisé* $\left(E(\hat{\sigma}^2) = \frac{n}{n-1}\sigma^2\right)$, *[12].*

Chapitre 2

Exhaustivité

Dans un problème statistique où figure un (ou plusieurs) paramètre θ inconnu, un échantillon apporte une certaine information sur ce paramètre. Lorsque l'on résume cet échantillon par une statistique T, il s'agit de ne pas perdre cette information ; une statistique qui conserve l'information sera qualifiée d'exhaustive. Pour plus information détaillée sur l'exhaustivité on réfère la lecture à [2], [11] et [14], ...

2.1 Statistique exhaustive

soit T une statistique fonction de $X_1, X_2..., X_n$ de la loi $g(t;\theta)$ (densité dans le cas continue, $P(T = t)$ dans le cas discret), $L(x_1, ..., x_n;\theta)$ soit la densité de $(X_1, X_2..., X_n)$ si X est absolument continue et $P(X_1 = x_1 \cap X_2 = x_2... \cap X_n = x_n)$ la probabilité conjointe si X est discrète.

Définition 2.1.1 *T sera dite exhaustive si l'on a $L(x,\theta) = g(t,\theta)h(x)$ (principe de factorisation) en d'autres termes si la densité conditionnelle de l'échantillon est indépendante du θ.*

Exemple 2.1.1 *La loi normal tel que μ connu, σ inconnu.*

1.1, posons $T = \sum_{i=1}^{n}(X_i - \mu)^2$, *et comme* $T/\sigma^2 \sim \chi^2_n$. *Alors la densité de* T *est*

$$\begin{aligned} g(t,\sigma) &= \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\left(\frac{t}{\sigma^2}\right)^{\frac{n}{2}-1}\exp\left(-\frac{1}{2}\sigma^2\right)\frac{1}{\sigma^2} \\ &= \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\frac{t^{\frac{n}{2}-1}}{\sigma^n}\exp\left(-\frac{1}{2}\sigma^2\right). \end{aligned}$$

d'où

$$L(x,\sigma) = g(t,\theta)\frac{\Gamma\left(\frac{n}{2}\right)}{\pi^{\frac{n}{2}}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]^{\frac{n}{2}-1}} = g(t,\theta)h(x).$$

Donc T *est exhaustive pour* σ^2.

En fait, on peut caractériser les lois de probabilité pour lesquelles les modèles d'échantillon admettent une statistique exhaustive : celles qui appartiennent à la famille exponentielle.

2.2 Famille exponentielle

Définition 2.2.1 *Soit* X *une v.a réelle dont la loi de probabilité dépend d'un paramètre* $\theta \in \mathbb{R}^k$. *On dit que la loi de* X *appartient à la famille exponentielle si et seulement si* $P(X = x;\theta)$ *(cas discret) ou* $f(x;\theta)$ *(cas continue) est de la forme :*

$$\exp\left[\sum_{j=1}^{n} a_j(x)\alpha_j(\theta) + b(x) + \beta(\theta)\right].$$

La plupart des lois usuelles appartiennent à la famille exponentielle :

Exemple 2.2.1 *Cas d'une loi discrète : loi de Bernoulli* $\mathcal{B}(p)$

$$P(X=x;\theta) = \begin{cases} p \ si\ x=1 \\ 1-p \ si\ x=0 \end{cases} = p^x(1-p)^{1-x}$$
$$= \exp\left[x\ln p + (1-x)\ln(1-p)\right]$$
$$= \exp x\left[\ln p - \ln(1-p)\right] + \ln(1-p)$$
$$= \exp\left[x\ln\frac{p}{1-p} + \ln(1-p)\right].$$

avec

$$d=1, a(x)=x, \alpha(p)=\ln\frac{p}{1-p}, b(x)=0 \ et\ \beta(p)=\ln(1-p).$$

Cas d'une loi continue : loi normal $\mathcal{N}(\mu,\sigma^2)$

$$2.2 = \exp\left[\left(-\frac{x^2}{2\sigma^2}\right) + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \ln\sigma\sqrt{2\pi}\right].$$

avec

$$d=2, a_1(x)=x^2, \alpha_1(\mu,\sigma^2) = -\frac{1}{2\sigma^2}, a_2(x)=x, \alpha_2(\mu,\sigma^2)=\frac{\mu}{\sigma^2},$$
$$b(x)=0 \ et\ \beta(\mu,\sigma^2) = -\frac{\mu^2}{2\sigma^2} - \ln\sigma\sqrt{2\pi}.$$

Le lien entre la famille exponentielle et l'exhaustivité est donné par le théorème de Darmois :

Théorème 2.2.1 (Darmois) *Soit* X *une v.a dont le domaine de définition ne dépend pas de* θ*. Une condition nécessaire et suffisante pour que l'échantillon* $(X_1,...,X_n)$ *admette une statistique exhaustive est que la forme de la densité soit :*

$$f(x;\theta) = \exp\left[a(x)\alpha(\theta) + b(x) + \beta(\theta)\right], \ \textit{(famille exponentielle).}$$

Si la densité est de cette forme et si de plus l'application $x_1 \longrightarrow \sum_{i=1}^{n} a(X_i)$ *est bijective et continument différentiable pour tout* i, *alors* $T = \sum_{i=1}^{n} a(X_i)$ *est une statistique exhaustive particulière.*

Preuve. Voir [14], page 293. ■

Le tableau (2.1) suivant contient la plupart des lois usuelles appartement à la famille exponentielle, [11] :

Loi	**Paramètre**	$\mathbf{a}_1(x)$	$\mathbf{a}_2(x)$
Bernoulli $\mathcal{B}(p)$	p	x	-
Binomial $\mathcal{B}(n,p)(n$ inconnu$)$	p	x	-
Binomial négative $\mathcal{BN}(r,p)(r$ connu$)$	p	x	-
Poisson $\mathcal{P}(\lambda)$	λ	x	-
exponentielle $\xi(\lambda)$	λ	x	-
Gamma $\gamma(r,\lambda)(r$ connu$)$	λ	x	-
normal $\mathcal{N}(\mu,\sigma^2)$	(μ,σ^2)	x^2	x
Pareto$(a$ connu$)$	θ	$\ln(x)$	-
Beta(α,β)	(α,β)	$\ln(x)$	$\ln(1-x)$

TAB. 2.1 – Principales lois usuelles appartement à la famille exponentielle

2.3 Information de Fisher (Données complètes)

L'information de Fisher est une notion statistique développée dans les années 1920 par le statisticien britannique Ronald Aylmer Fisher $(1890-1962)$ et anticipant d'une vingtaine d'années la fonction " entropie " de Shannon, base de la théorie de l'information, l'information de Fisher dont l'intérêt est souligné par l'inégalité de Cramer Rao présentée ci-après, a pour objet la quantification de l'information pertinente contenue dans les données.
On considère le modèle statistique $(\mathcal{X}, \mathcal{A}, P_\theta)$; $\theta \in \Theta$ où Θ désigne un ouvert de $\mathbb{R}^k$, $k \geq 1$. Soit X une v.a.r de loi P_θ admet une densité $f(x;\theta)$ (cas continue). On suppose dans cette section que les hypothèses suivantes sont satisfaites :

H1 pour tout $\theta \in \Theta$, $x \in \mathcal{X}$, les lois P_θ ont toutes même support $\{x : f(x;\theta) > 0\}$, qui.ne dépend pas de θ.

H2 pour tout $\theta \in \Theta$, $x \in \mathcal{X}$, $\frac{\partial}{\partial\theta}\ f(x;\theta)$ et $\frac{\partial^2}{\partial\theta^2} f(x;\theta)$ existent et continuent, en d'autres termes : $f(x;\theta) \in C^2(\Theta)$.

H3 pour tout $\theta \in \Theta$, $A \in \mathcal{A}$: on peut dérivé deux fois $\int\limits_{\mathcal{A}} f(x;\theta)dx$ par rapport à θ sous le signe d'intégration, comme suit :

$$\frac{\partial}{\partial\theta_j}\int\limits_{\mathcal{A}} f(x;\theta)dx = \int\limits_{\mathcal{A}} \frac{\partial}{\partial\theta_j} f(x;\theta)dx,\ j = 1, ..., k.$$
$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\int\limits_{\mathcal{A}} f(x;\theta)dx = \int\limits_{\mathcal{A}} \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(x;\theta)dx,\ i, j = 1, ..., k.$$

(Dans le cas discret, les hypothèses portent sur les sommations en lieu et place des intégrations).

Remarque 2.3.1 **H1**, **H2** *et* **H3** *sont vérifiées si P_θ appartient à la famille exponentielle.*

Sous ces hypothèses, on peut définir les quantités suivantes :

2.3.1 Information de Fisher sur un paramètre réel

Définition 2.3.1 *On appel quantité d'information de Fisher $I(\theta)$ apportée sur un paramètre réel θ $(k = 1)$, par une seule observation :*

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right] = E\left[\left(\frac{f'_\theta(X;\theta)}{f(X;\theta)}\right)^2\right]. \tag{2.1}$$

Exemple 2.3.1 •*Dans le cas discret*

Soit X une v.a de la loi de Bernoulli $\mathcal{B}(p)$.

$$f(x;p) = P(X = x;\theta) = p^x(1-p)^{1-x},$$

Par calculs, on trouve

$$\ln\left(f(x;p)\right) = x\ln(p) + (1-x)\ln(1-p),$$

et

$$\frac{\partial}{\partial p}\ln f(x;p)=\frac{x}{p}-\frac{(1-x)}{1-p},$$
$$\frac{\partial^2}{\partial p^2}\ln f(x;p)=-\frac{x}{p^2}-\frac{(1-x)}{(1-p)^2}.$$

Comme $E[X]=p$, *on déduit*

$$I(p)=E\left[\frac{\partial^2}{\partial p^2}\ln f(x;p)\right]=\frac{1}{p(1-p)}.$$

•*Dans le cas continue*

Soit $X\sim\mathcal{N}(\mu,\sigma^2)$, ($\sigma^2$ *connu*) :

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

alors

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

et

$$\frac{\partial\ln f(x;\mu)}{\partial\mu}=\frac{x-\mu}{\sigma^2},$$

et donc

$$I(\mu)=E\left[\left(\frac{x-\mu}{\sigma^2}\right)^2\right]=\frac{1}{\sigma^2}E\left[\left(\frac{x-\mu}{\sigma}\right)^2\right]=\frac{1}{\sigma^2}.$$

Car : $E(x)=\mu$ *et* $\frac{x-\mu}{\sigma}\sim\mathcal{N}(0,1)$.

Remarque 2.3.2 *L'information apportée par un échantillon de taille* $n\geq 1$ *est*

$$I_n(\theta):=E\left[\left(\frac{\partial\ln L(X_1,...,X_n;\theta)}{\partial\theta}\right)^2\right]=E\left[\left(\frac{L'_\theta}{L}\right)^2\right].$$

Où

$$L(X_1, ..., X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

Voir le tableau (2.2) on résume les expressions de l'information de Fisher pour certaines lois de probabilité usuelles, [11] :

Loi	**Paramètre**	**information**
Bernoulli $\mathcal{B}(p)$	p	$1/p(1-p)$
Binomial $\mathcal{B}(n,p)$ (n inconnu)	p	$n/p(1-p)$
Binomial négative $\mathcal{BN}(r,p)$ (r connu)	p	$r/p^2(1-p)$
Poisson $\mathcal{P}(\lambda)$	λ	$1/\lambda$
exponentielle $\xi(\lambda)$	λ	$1/\lambda^2$
Gamma $\gamma(r,\lambda)$ (r connu)	λ	r/λ^2
normal $\mathcal{N}(\mu,\sigma^2)$ (μ connu)	σ^2	$1/2\sigma^4$
normal $\mathcal{N}(\mu,\sigma^2)$ (σ^2 connu)	μ	$1/\sigma^2$
pareto(a,θ) (a connu)	θ	$1/\theta^2$
uniforme $\mathcal{U}[0,\theta]$	θ	$1/\theta^2$

TAB. 2.2 – Résumé de l'information sur les paramètre des lois usuelles

Théorème 2.3.1 *Sous les hypothèses* **H1**, **H2** *et* **H3**, *on obtient pour* $I(\theta)$, *une seconde expression fort utile pour les calculs, à savoir :*

$$I(\theta) = -E\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right),$$

si cette quantité existe.

Preuve. En effet, supposant par exemple X à valeurs sur $\mathbb{R}$ et y satisfaisant les conditions $f(X;\theta)$ strictement positive, deux fois dérivable et de support indépendant de θ ; on a par définition $\int\limits_{\mathbb{R}} f(X;\theta)dX = 1$, pour tout θ.
En dérivant les deux membres par rapport à θ et en remarquant que

$$\frac{\partial}{\partial\theta}\left(\int\limits_{\mathbb{R}} f(X;\theta)dX\right) = 0 = \int\limits_{\mathbb{R}}\left(\frac{\partial}{\partial\theta} f(X;\theta)dX\right),$$

ce qui implique

$$\int_{\mathbb{R}} \left(\frac{\partial}{\partial \theta} \ln f(X;\theta) \right) . f(X;\theta) dX = 0.$$

Soit l'espérance mathématiquement : $E\left(\frac{\partial}{\partial \theta} \ln f(X;\theta)\right)$; ce qui prouve que la v.a $\frac{\partial \ln f(X;\theta)}{\partial \theta}$ est centrée et que $I(\theta) = Var\left(\frac{\partial \ln f(X;\theta)}{\partial \theta}\right)$.

Dérivant une deuxième fois $E\left(\frac{\partial}{\partial \theta} \ln f(X;\theta)\right)$, il vient

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}} \left(\frac{\partial}{\partial \theta} \ln f(X;\theta) \right) . f(X;\theta) dX = 0,$$

$$\int_{\mathbb{R}} \frac{\partial^2 \left(\ln f(X;\theta)\right)}{\partial \theta^2} f(X;\theta) dX + \int_{\mathbb{R}} \left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 f(X;\theta) dX = 0,$$

ce qui implique

$$\int_{\mathbb{R}} \left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 f(X;\theta) dX = - \int_{\mathbb{R}} \frac{\partial^2 \left(\ln f(X;\theta)\right)}{\partial \theta^2} f(X;\theta) dX. \tag{2.2}$$

Ainsi,

$$I(\theta) = E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 \right] = \int_{\mathbb{R}} \left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 f(X;\theta) dX.$$

D'après 2.2 on a aussi

$$I(\theta) = - \int_{\mathbb{R}} \frac{\partial^2 \left(\ln f(X;\theta)\right)}{\partial \theta^2} f(X;\theta) dX = -E\left[\frac{\partial^2 \left(\ln f(X;\theta)\right)}{\partial \theta^2} \right],$$

ce qui démontre le théorème. ■

Propriété de $I(\theta)$

***La positivité** la quantité d'information de Fisher est positive ; d'après la formule 2.1 on déduit que

la v.a $f'_\theta(X;\theta)/f(X;\theta)$ étant centrée alors on a

$$I(\theta) = Var(\frac{f'_\theta(X;\theta)}{f(X;\theta)}) \geq 0.$$

***L'additivité** soient X et Y deux variables aléatoires indépendantes de lois respectives $f(X,\theta)$ et $g(Y,\theta)$, la quantité d'information $I(\theta)$ liée au couple (X,Y), soit

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta),$$

des quantités d'informations liées respectivement à X et Y.

Démonstration

Ecrivons

$$I_{(X,Y)}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln\left(f(X;\theta).g(Y;\theta)\right)\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\left(\ln f(X;\theta) + \ln g(Y;\theta)\right)\right],$$

et par la linéarité de l'espérance

$$I_{(X,Y)}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\left(\ln f(X;\theta)\right)\right] - E\left[\frac{\partial^2}{\partial\theta^2}\left(\ln g(Y;\theta)\right)\right].$$

Ceci établit le résultat annoncé.

Corollaire 2.3.1 *Soit un échantillon $(X_1,...,X_n)$ de n variables aléatoires indépendantes X_i, sous les hypothèses* **H1**, **H2** *et* **H3**, *on a*

$$I_n(\theta) = I_{X_1}(\theta) + I_{X_2}(\theta) + ... + I_{X_n}(\theta) = nI_1(\theta).$$

Ceci veut dire que chaque observation apporte la même information, ce qui n'est pas le cas pour la loi uniforme sur $[0,\theta]$ où la plus grande observation est la plus intéressante.

Preuve. Le résultat découe du fait que $X_1, ..., X_n$ sont identiquements distribuées. Alors

$$I_{X_1}(\theta) = I_{X_2}(\theta) = ... = I_{X_n}(\theta) = I(\theta),$$

d'où le résultat.

Au contraire, dans le cas pour la loi uniforme sur $[0, \theta]$, $\theta > 0$. Il est clair que le support est de X dépend de θ, d'une part, on a

$$I(\theta) = E\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2 = E\left[\frac{\partial(-\ln\theta)}{\partial\theta}\right]^2 = \frac{1}{\theta^2},$$

et d'autre part

$$I_n(\theta) = E\left[\frac{\partial \ln L(X_1, ..., X_n;\theta)}{\partial\theta}\right]^2 = E\left[\frac{\partial(-n\ln\theta)}{\partial\theta}\right]^2 = \frac{n^2}{\theta^2},$$

ce qui implique : $I_n(\theta) \neq nI(\theta)$. ■

Exemple 2.3.2 •*Loi normal :*

Pour un échantillon $(X_1, ..., X_n)$ *d'une population* $X \sim \mathcal{N}(\mu, \sigma^2)$, *on a*

$$L(x_1, ..., x_n;\mu,\sigma) = \prod_{i=1}^{n} f(x_i;\mu,\sigma) = \left(\frac{1}{2\pi\sigma}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right),$$

par calcule on trouve

$$\frac{\partial \ln L}{\partial\mu} = \frac{\partial}{\partial\mu}\left(\ln\left(\frac{1}{2\pi\sigma}\right)^{\frac{n}{2}}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)\right) = 0 - \frac{-2n(\bar{x}-\mu)}{2\sigma^2}.$$

D'où

$$\frac{\partial^2 \ln L}{\partial\mu^2} = -\frac{n}{\sigma^2},$$

par conséquent, on obtient

$$I_n(\mu) = \frac{n}{\sigma^2} = nI(\mu).$$

2.3.2 Information de Fisher sur un paramètre vectoriel

Soit X une v.a de densité f dépendant d'un paramètre vectoriel $\theta = (\theta_1, ..., \theta_k)^T \in \mathbb{R}^k$.

Supposant que les hypothèses sont vérifées :

Définition 2.3.2 *Le score est le gradient de la log-vraisemblance*

$$S(X;\theta) = \nabla \ln L(X;\theta) = (S_1(X;\theta), ..., S_k(X;\theta))^T.$$

où pour tout $j \in \{1, ..., k\}$, $S_j(X;\theta) = \frac{\partial}{\partial \theta_j} \ln L(X;\theta)$, *alors le score est un vecteur aléatoire de dimension* k.

Remarque 2.3.3 *Si* $\theta \in \mathbb{R}$ *la v.a.*$S(X;\theta) = \frac{\partial}{\partial \theta} \ln L(X;\theta)$.

Propriété du score :

1. $E(S(X;\theta)) = 0$ c-à-d le score est un vecteur aléatoire centré.

   Démonstration 2.3.1 *Pour tout* $j \in \{1, ..., k\}$ *on a*

$$\begin{aligned} E(S_j(X;\theta)) &= E(\frac{\partial}{\partial \theta_j} \ln L(X;\theta)) \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial \theta_j} \ln L(x;\theta) L(x;\theta) d\mu(x) \\ &= \int_{\mathcal{X}} \frac{\frac{\partial}{\partial \theta_j} L(x;\theta)}{L(x;\theta)} L(x;\theta) d\mu(x) \text{ d'après les hypothèses } \mathbf{H1}, \mathbf{H2} \text{ et } \mathbf{H3} \\ &= \frac{\partial}{\partial \theta_j} P(X \in \mathcal{X}) = \frac{\partial}{\partial \theta_j} 1 = 0. \end{aligned}$$

2. Soient X et Y deux variables aléatoires indépendantes, le score dans ce cas est définit comme suit

$$S((X,Y);\theta) = S(X;\theta) + S(Y;\theta), \text{ pour tout } \theta,$$

et donc le vecteur score est additif.

Donc facilement on définit l'information de Fisher à partir du vecteur score :

La matrice d'information de Fisher $I(\theta)$ est la matrice de covariance du score, de terme générale :

$$\begin{aligned} I(\theta) &= cov(S(X;\theta); S(X;\theta)) \\ &= E(S(X;\theta)S(X;\theta)^t) \text{ car } E(S(X;\theta)) = 0 \\ &= \begin{pmatrix} E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial \theta_1}\right)^2\right] & . & . & . & E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta_1}\frac{\partial \ln f(X;\theta)}{\partial \theta_k}\right] \\ . & & . & & . \\ . & & . & & . \\ . & & . & & . \\ E\left[\frac{\partial \ln f(X;\theta)}{\partial \theta_1}\frac{\partial \ln f(X;\theta)}{\partial \theta_k}\right] & . & . & . & E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial \theta_k}\right)^2\right] \end{pmatrix}, \end{aligned}$$

c'est une matrice symétrique définie positive.

Corollaire 2.3.2 *Pour l'échantillon $(X_1, ..., X_n)$, le vecteur score $S((X_1, ..., X_n); \theta)$ sera noté $S_n(\theta)$ et l'information de Fisher associée sera notée $I_n(\theta)$.*

Par indépendance des X_j, on a

$$S_n(\theta) = \nabla(\sum_{i=1}^{n} \ln f(X_i;\theta)),$$

et par la linéarité du gradient donne

$$S_n(\theta) = \sum_{j=1}^{n} (S(X_j;\theta)).$$

Or les vecteurs scores $S(X_1;\theta), ..., S(X_n;\theta)$ sont iid (de la même loi que $S(X;\theta)$). On a donc

$$I(\theta) = Var(S_n(X;\theta)) = \sum_{j=1}^{n} Var\,(S(X_j;\theta)) = nI(\theta).$$

Remarque 2.3.4 *Pour avoir la loi asymptotique du score il suffit d'appliqué le théorème limite centrale (TLC) aux $S(X_i;\theta)$ donne immédiatement. Pour tout $\theta \in \Theta$ on a*

$$\frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta)).$$

Proposition 2.3.1 *Sous les hypothèses* **H1**, **H2** *et* **H3**, *on a*

$$\begin{aligned} I(\theta) &= -E\left[\nabla(S(X;\theta)^T)\right] \\ &= \begin{pmatrix} -E\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta_1^2}\right] & . & . & . & -E\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta_1\partial\theta_k}\right] \\ . & & . & & . \\ . & & . & & . \\ . & & . & & . \\ -E\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta_1\partial\theta_k}\right] & . & . & . & -E\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta_k^2}\right] \end{pmatrix} \end{aligned}$$

Donc

$$I(\theta) = -E\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta_i \partial\theta_j}\right) \ ; \ \textit{pour tout } i, \ j = 1, ..., k$$

Exemple 2.3.3 *Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ tel que $\theta = (\mu, \sigma^2)^T \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$ et la densité 1.1 avec $x \in \mathbb{R}$.*

On a : le support $S_\theta = \mathbb{R}$ ne dépend pas des paramètres et les hypothèses **H1**, **H2** *et* **H3** *sont satisfaites, alors*

$$I(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix},$$

la matrice d'information est diagonale (les paramètres μ et σ^2 sont orthogonaux).

En effet,

$$\ln f(x;\mu,\sigma^2) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2)\frac{1}{2\sigma^2}\ln(x-\mu)^2,$$

et par calculs

$$\frac{\partial \ln(f(x;\mu,\sigma^2))}{\partial \mu} = \frac{x-\mu}{\sigma^2}, \frac{\partial \ln(f(x;\mu,\sigma^2))}{\partial(\sigma^2)} = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}$$

$$\begin{aligned}
&\frac{\partial^2 \ln(f(x;\mu,\sigma^2))}{\partial \mu^2} = -\frac{1}{\sigma^2} \Rightarrow -E\left(\frac{\partial^2 \ln(f(x;\mu,\sigma^2))}{\partial \mu^2}\right) = \frac{1}{\sigma^2},\\
&\frac{\partial^2 \ln(f(x;\mu,\sigma^2))}{\partial(\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6} \Rightarrow -E\left(\frac{\partial^2 \ln(f(x;\mu,\sigma^2))}{\partial(\sigma^2)^2}\right) = \frac{1}{2\sigma^4},\\
&\frac{\partial^2 \ln(f(x;\mu,\sigma^2))}{\partial \mu \partial(\sigma^2)} = -\frac{x-\mu}{\sigma^4} = \frac{\mu-x}{\sigma^4} \Rightarrow E\left(\frac{\partial^2 \ln(f(x;\mu,\sigma^2))}{\partial \mu \partial(\sigma^2)}\right) = 0.
\end{aligned}$$

Propriétés de $I(\theta)$

***** L'information de Fisher est une matrice symétrique définie positive. En effet, étant donné que le score est centré :

$$I(\theta) = Var(S(X;\theta)) \geq 0.$$

***** Soient X et Y deux v.a r.indépendantes à valeurs dans $\mathcal{X}$ et $\mathcal{Y}$ de la loi respective P_θ et Q_θ, alors

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta), \text{pour tout } \theta \in \Theta,$$

c-à-d que l'information de Fisher est additive, car c'est la variance d'une somme de scores indépendants.

2.3.3 Inégalité de Fréchet-Darmois- Cramer- Rao(FDCR)

L'intérêt principal de la quantité d'information de Fisher est qu'elle fournit une borne inférieure pour la variance de n'importe quel estimateur sans biais de θ, grâce à l'inégalité de FDCR pour n'importe quelle statistique T.

Théorème 2.3.2 *Sous les hypothèses et $E(T)$ est dérivable sous le signe somme, on a*

pour tout estimateur T sans biais de θ

$$Var(T) \geq \frac{1}{I_n(\theta)}, \text{ où } 0 < I_n(\theta) < +\infty,$$

la quantité $I_n^{-1}(\theta)$ est appelée la borne de Cramer-Rao.

Si T est un estimateur sans biais de $h(\theta)$, alors

$$Var(T) \geq \frac{[h'(\theta)]^2}{I_n(\theta)}.$$

Démonstration 2.3.2 *D'une part, on considère que*

$$cov(T, \frac{\partial \ln L}{\partial \theta}) = E(T\frac{\partial \ln L}{\partial \theta}) - E(T)E(\frac{\partial \ln L}{\partial \theta}),$$

est comme $\frac{\partial \ln L}{\partial \theta}$ est centrée c-à-d $E(\frac{\partial \ln L}{\partial \theta}) = 0$. Donc

$$\begin{aligned} cov(T, \frac{\partial \ln L}{\partial \theta}) &= \int t\frac{\partial \ln L}{\partial \theta} L dx = \int t\frac{\partial L}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta}\int tL dx = \frac{\partial}{\partial \theta}E(T) \\ &= h'(\theta). \end{aligned}$$

D'autre part, l'inégalité de Schwarz donne

$$\left[cov(T, \frac{\partial \ln L}{\partial \theta})\right]^2 \leq Var\,(T)\, Var\left(\frac{\partial \ln L}{\partial \theta}\right),$$

et donc par substitution, on trouve

$$[h'(\theta)]^2 \leq Var(T) I_n(\theta).$$

Définition 2.3.3 *L'estimateur T est efficace si*

$$Var(T) = \frac{[h'(\theta)]^2}{I_n(\theta)}, \text{ pour tout } \theta \in \Theta$$

T est donc un estimateur sans biais de variance minimale de $h(\theta)$.

Remarque 2.3.5 •*Un estimateur efficace est de variance minimale, c-à-d optimal.*

•*Un estimateur peut être sans biais, de variance minimale, mais ne pas atteindre la borne de Cramer-Rao, donc ne pas être efficace.*

•*L'efficacité est une notion qui fait le lien entre la théorie de l'information et l'estimation : plus l'information de Fisher est grande et plus la borne de Cramer Rao est petite, i.e. plus on a une chance de trouver un estimateur sans biais de faible variance.*

Exemple 2.3.4 ·*Cas d'une loi discrète*

Considérons un n-échantillon $(X_1, ..., X_n)$ de la loi de Bernoulli de paramètre p. L'estimateur $\overline{X}_n$ est un estimateur sans biais et efficace de p.

En effet, des calculs donnent

$$E(\overline{X}_n) = p, \ Var(\overline{X}_n) = \frac{p(1-p)}{n}.$$

La fonction de vraisemblance est donnée par

$$L(x_1, ..., x_n; p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \mathbf{1}_{x_i = \{0,1\}}.$$

En dérivant deux fois par rapport à p la log-vraisemblance, on obtient pour $x_i \in \{0, 1\}$, $i = 1, ..., n$:

$$\frac{\partial^2}{\partial p^2} \ln \left[(L(x_1, ..., x_n; p)) \right] = -\frac{\sum_{i=1}^{n} x_i}{p^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-p)^2},$$

l'information de Fisher est donnée par

$$I_n(p) = -E\left(-\frac{\sum_{i=1}^{n} x_i}{p^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-p)^2}\right) = \frac{n}{p(1-p)},$$

d'où le résultat

$$Var(\overline{X}_n) = \frac{1}{I_n(p)}.$$

Exemple 2.3.5 ·*Cas d'une loi continue*

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ *où* μ *connue et* $(X_1, ..., X_n)$ *iid, L'estimateur* S_n^{*2} *est un estimateur sans biais, non efficace de* σ^2 *tel que*

$$\begin{aligned} E(S_n^{*2}) &= \frac{n}{n-1} E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right) \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right)^2 - E(\overline{X}_n)^2\right) \\ &= \frac{n}{n-1}\left((\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2)\right) \\ E(S_n^{*2}) &= \sigma^2 \end{aligned}$$

$$Var(S_n^{*2}) = \frac{2\sigma^4}{n-1}$$

$$\ln\left[\left(L(x_1, ..., x_n; \sigma^2)\right)\right] = -\ln(\sigma^n(2\pi)^{\frac{n}{2}}) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}, \text{ pour tout } x_i \in \mathbb{R},\ i = 1, ..., n.$$

On obtient

$$\frac{\partial^2}{\partial(\sigma^2)^2} \ln\left[\left(L(x_1, ..., x_n; \sigma^2)\right)\right] = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2.$$

L'information de Fisher est donnée par

$$I_n(\sigma^2) = -E\left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2\right) = \frac{n}{2\sigma^4},$$

et donc

$$\frac{2\sigma^4}{n-1} \neq \frac{2\sigma^4}{n},$$

ce qui démontre que S_n^2 *est non efficace.*

Par contre, si σ^2 *connue, l'estimateur* $\overline{X}_n$ *est un estimateur sans biais et efficace de* μ.

Par calculs, on a

$$E(\overline{X}_n) = \mu,\ Var(\overline{X}_n) = \frac{\sigma^2}{n}$$

et

$$I_n(\mu) = \frac{n}{\sigma^2},$$

d'où le résultat.

2.3.4 Lien entre l'information au sens de Fisher et la statistique

Le résultat suivant établit le lien étroit qui existe entre les notions de statistique et d'information au sens de Fisher. Soit $I_n(\theta) = nI(\theta)$ l'information de Fisher de l'échantillon $(X_1, ..., X_n)$:

Théorème 2.3.3 (Dégradation de l'information) *Sous les hypothèses* **H1**, **H2** *et* **H3**, *on a*

$$0 \leq I_T(\theta) \leq I_n(\theta).$$

c-à-d l'information apportée par une statistique T *est inférieur ou égale à celle apportée par l'échantillon, où l'information de Fisher apportée par la statistique* T *sur le paramètre inconnu* θ

$$I_T(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(T;\theta)\right)^2\right]$$

et

$$I_T(\theta) = I_n(\theta) \Leftrightarrow T \text{ est exhaustive.}$$

Preuve. Soit T de densité $g(t;\theta)$ la statistique que l'on substitue à l'échantillon, on a

$$L(x;\theta) = g(t;\theta)h(x_1,...,x_n;\theta/t),$$

où : $h(x_1,...,x_n;\theta/t)$ est la densité conditionnelle de l'échantillon .On a donc prenant l'espérance des dérivées secondes

$$I_n(\theta) = I_T(\theta) - E\left(\frac{\partial^2 \ln h}{\partial\theta^2}\right),$$

le dernier terme est la quantité d'information conditionnelle $I_{n/T}(\theta)$ (ou information supplémentaire) ; elle est positive ou nulle, donc

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

■

Définition 2.3.4 (statistique libre) *Une statistique T d'un modèle paramétrique est dite libre si sa loi ne dépend pas du paramètre θ. Une statistique libre n'apporte donc aucune information sur le paramètre θ.*

Remarque 2.3.6 *L'information de Fisher apportée par une statistique libre est nulle. En autres termes on a*

$$I_T(\theta) = 0 \Leftrightarrow T \text{ est libre.}$$

2.4 Information de Fisher (Données incomplètes)

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information : au lieu d'observer des réalisations i.i.d de durées X, on observe la réalisation de la variable X soumise à diverses perturbations, indépendantes ou non du phénomène

étudié. Pour plus d'informations sur les données incomplètes, on recommande de voir [7], [13] et [16],...

2.4.1 Censure

La durée de survie n'est pas toujours complètement observée, parce que, pour certains individus, l'évènement de début et/ou de fin n'est pas observé. On parle alors de données censurées.

Pour l'individu i, considérons :

•son temps de survie X_i,

•son temps de censure C_i,

•la durée réellement observée T_i.

Censure à droite

Le phénomène le plus souvent à l'origine de ces données incomplètes est la censure à droite. La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées ; pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue.

Il existe trois différents types de censure à droite :

1. **La censure de type I :** Soit C une valeur fixée, au lieu d'observer les variables $X_1, X_2..., X_n$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i \leq C$; sinon on sait uniquement que $X_i > C$. On utilise la notation suivante

$$T_i = X_i \wedge C = min(X_i, C).$$

2. **La censure de type II :** Elle est présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude

à ce moment-là. Soient $X_{(i)}$ et $T_{(i)}$ les statistiques d'ordre des variables X_i et T_i : La date de censure est donc $X_{(k)}$ et on observe les variables suivantes

$$T_{(1)} = X_{(1)}$$
$$\vdots$$
$$T_{(k)} = X_{(k)}$$
$$T_{(k+1)} = X_{(k+1)}$$
$$\vdots$$
$$T_{(n)} = X_{(k)}$$

3. **La censure de type III (ou censure aléatoire de type I) :** Soient $C_1, C_2..., C_n$ des variables aléatoires i.i.d. On observe les variables $T_i = X_i \wedge C_i$:

   L'information disponible peut être résumée par :

   •La durée réellement observée T_i,

   •Un indicateur : $\delta_i = \mathbf{1}_{\{X_i \leq C_i\}}$

   $= \begin{cases} 1 \text{ si l'événement est observé (d'où } T_i = X_i\text{).On observe les vraies durées (complètes).} \\ 0 \text{ si l'individu est censuré (d'où } T_i = C_i\text{).On observe des durées incomplètes (censurées).} \end{cases}$

2.4.2 Maximum de vraisemblance (Données censurées)

La méthode du maximum de vraisemblance pour estimer, au vu de données censurées, les paramètres réels d'un modèle d'analyse des durées de vie.

Nous considérons le cas d'une censure aléatoire à droite, les observations sont les couples $(T_1, \delta_1), ..., (T_n, \delta_n)$, où $T_i = min(X_i; C_i)$ est la durée observée, et l'indicateur de censure $\delta_i = \mathbf{1}_{\{X_i \leq C_i\}}$.

Hypothèse fondamentale :

On suppose que C_i de l'individu i est une v.a indépendante de la durée de vie X_i.

Proposition

Sous l'hypothèse fondamentale d'indépendance $X_i \coprod C_i$ pour $i = 1, ..., n$. La vraisemblance s'écrit

$$L((t_1, \delta_1), ..., (t_n, \delta_n); \theta) = \prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} S_\theta(t_i)^{1-\delta_i},$$

où f_θ est la densité commune des T_i et S_θ la fonction de survie associe.

Démonstration 2.4.1 *Soit la suite des délais de censure $C_1, C_2..., C_n$ i.i.d de densité commune g et G la survie associée, i.e. : $G(c) = P(C_1 > c)$.*
le couple de v.a (T_1, Δ_1) admet pour densité

$$g(t_i)S_\theta(t_i) \text{ si } \delta_i = 0 \text{ (observations censurées) ;}$$
$$f_\theta(t_i)G(t_i) \text{ si } \delta_i = 1 \text{ (dures } t_i = x_i \text{ observes),}$$

que l'on peut aussi écrire de façons équivalente

$$[f_\theta(t_i)G(t_i)]^{\delta_i} [g(t_i)S_\theta(t_i)]^{1-\delta_i},$$

de plus les couples $(T_1, \Delta_1), ..., (T_n, \Delta_n)$ sont indépendants donc la vraisemblance des observations s'écrit :

$$\prod_{i=1}^{n} [f_\theta(t_i)G(t_i)]^{\delta_i} [g(t_i)S_\theta(t_i)]^{1-\delta_i},$$

comme la loi des C_i ne fait pas intervenir le paramètre θ, la partie utile de la vraisemblance se réduit à :

$$L(\theta) = \prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} S_\theta(t_i)^{1-\delta_i}.$$

2.4.3 Troncature

Les données censurées ne sont pas le type unique de données incomplètes. L'autre cas classique de données incomplètes est celui des données dites tronquées. Les troncatures Les troncatures différentes des censures au sens où elles concernent l'échantillonnage lui-

même. Une observation est dite tronquée si elle est conditionnelle à un autre évènement. On dit que la variable T de durée de vie est tronquée si T n'est observable que sous une certaine condition dépendante de la valeur de T.

Troncature à gauche et droite

Soit C une variable aléatoire indépendante de X, on dit qu'il y a troncature à gauche lorsque X (la durée de survie) n'est observable que si $X > C$: On observe le couple $(X; C)$; avec $X > C$. De même, il y a troncature à droite lorsque X n'est observable que si $X < C$.

2.4.4 Maximum de vraisemblance (Données tronquées)

La vraisemblance représente la probabilité d'observer l'échantillon d'après le modèle et est le produit des n contributions individuelles

$$L = \prod_{i=1}^{n} L_i,$$

Soit t_i le temps de participation du sujet i. Considérons le cas de données tronquées à gauche de manière aléatoire, la contribution individuelle d'un sujet i dont l'observation est tronquée à gauche en a_i est

$$L_i = \begin{cases} \frac{f_\theta(t_i)}{S_\theta(a_i)} \text{ si } \delta_i = 1 \\ \frac{f_\theta(t_i)}{S_\theta(a_i)} \text{si } \delta_i = 0 \end{cases}$$

Conclusion

Dans ce mémoire nous avons conclu que la théorie de l'information de Fisher fournit un cadre mathématique pour quantifier l'information contenue pour l'échantillon $(X_1, ..., X_n)$. Cette dernière a une grande importance sur l'étude de la qualité d'un estimateur en terme d'efficacité, basée sur FDCR. En effet, un estimateur sans biais de faible variance ayant une information de Fisher très grande. En outre, nous avons défini l'information de Fisher au cas des données aléatoirement censurées et tronquées à droite, dans le but de l'utiliser aux statistiques des valeurs extrêmes, voir les travaux récents de : Jan Beirlant, Brahim Brahimi, Djamel Meraghni, Abdelhakim Necir.

Bibliographie

[1] Barra, J. R., & Linnik, U. V. (1971). Notions fondamentales de statistique mathématique : maîtrise de mathématiques et applications fondamentales.

[2] Borovkov, A. (1987). Statistique Mathématique. Edition Mir Moscou.

[3] Boulay, J-P. (2010). Statistique Mathématique. Ellipses Édition Marketing.S.A.

[4] Dauxois, J-Y. (2011 – 2012). CTU, licence de mathématiques statistique inférentielle. Université de Franche-Comté.
https ://lmb.univ-fcomte.fr/IMG/pdf/cours_stat_inf.pdf.

[5] Dusart, P. (2015). Cours de Statistiques inférentielles .
http ://www.unilim.fr/pages_perso/pierre.dusart/Probas/cours_stat_S4.pdf.

[6] Gassama, M. (2016). Estimation du risque attribuable et de la fraction préventive dans les études de cohorte (Doctoral dissertation, Paris Saclay).

[7] Hughes, E. J. (1962). Maximum likelihood estimation of distribution parameters from incomplete data.

[8] Hurlin, C., & Mignon, V. (2015). Statistique et probabilités en économie-gestion. Dunod.

[9] Lecoutre, J. P. (2009). Statistique et probabilités-4e édition. Dunod.

[10] Lehmann,E.I.,Casella,G.,(2003).Theory of point estimation.Springer.

[11] Lejeune,M.,Statistique,la théorie et ses applications.Springer,Paris.

[12] Necir, A. (2016). Cours de troisième année licence. Université Mohamed Khider de Biskra.

[13] Saint-Pierre, P. (2015). Introduction à l'analyse des durées de survie. Cours Université Pierre et Marie Curie.

[14] Saporta, G., 2006. Probabilité, Analyse de Données et Statistique. Technip.

[15] Spall, J. C. (2005). Introduction to stochastic search and optimization : estimation, simulation, and control (Vol. 65). John Wiley & Sons.

[16] Touraine, C. (2013). Modèles illness-death pour données censurées par intervalle : application à l'étude de la démence (Doctoral dissertation, Bordeaux 2).

Annexe A : Rappel

Inégalité de Beinaymé-Chebychev

Soit X une v.a d'espérance μ et de varaince finie σ^2 (l'hypothèse de variance finie garantit l'existence de l'espérance). L'inégalité de Bienaymé-Chebychev s'énonce de la façon suivante : pour tout réel ε strictement positif :

$$\text{pour tout } \varepsilon > 0,\ P\left(|X - \mu| > \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2}$$

Théorème centrale limite

Soit $(X_1, ..., X_n)$ un échantillon iid, avec $E\left[X_i\right] = \mu$ et $Var\left(X_i\right) = \sigma^2 < \infty$. On a alors

$$\overline{X} \xrightarrow{\mathcal{L}} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Annexe B : Abréviations et Notations

Les différentes abrévariations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

Notations	Signification
$(X_1, ..., X_n)$	échantillon de taille n de v.a's.
v.a	variable.aléatoire.
$L(x_1, ..., x_n; \theta)$	fonction de vraisemblance de θ pour une réalisation $(x_1, ..., x_n)$ d'un échantillon.
$E[X]$ ou μ	l'éspérance mathématique ou moyenne du v.a X.
$Var(X)$	variance du v.a X.
cov	covariance.
R	risque.
θ	paramètre inconnu.
Θ	ensemble des valeurs possibles du paramètre θ.
$I(\theta)$	quantité d'information de Fisher apportée sur θ.
FDCR	Fréchet Darmois Cramer Rao
$b_\theta(T)$	le biais de l'estimateur T pour θ.
c-à-d / i.e	c'est-à-dire.
iid	indépendantes identiquement distribuées.
exp	fonction exponentielle.
$\xrightarrow{\mathcal{L}}$	convergence en lois
lim	limite.
$\mathbb{R}$	les nombres réelles.