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

BENAISSA MOHAMMED

Titre :

Théorie d'estimation

Membres du Comité d'Examen :

Dr.	**CHERFAOUI MOULOUD**	UMKB	Président
Dr.	**TOUBA SONIA**	UMKB	Encadreur
Dr.	**DHIABI SAMRA**	UMKB	Examinateur

Septembre 2020

Dédicace

Je dédie ce modeste travail :

A Ma tendre Mère Fatiha .

A Mon très cher Père Sayeh ,Qua Dieu bénisse son âme .

A Ma future femme

A mon cher frères

A mes très chère amis :Noureddine,Ibrahim,Karim,Houssem Abderaouf.

A tous mes enseignants depuis mes premières années d'études.

A tous ceux qui me sens chers et que j'ai omis de citer

Enfin, je ne pourrais manquer de souligner l'apport et le soutien de mes collègues étudiants à UNIVERSITÉ MOHAMED KHIDER BISKRA de qui ont rendu cette pério qui de ma vie si intéressante et si enrichissante.

BENAISSA MOHAMMED

REMERCIEMENTS

Ce travail de mémoire de master est la première expérience dans l'activité de recherche.

Il n'aurait pas été aussi fructueux sans l'aide de plusieurs personnes.

Je vais donc m'essayer à trouver les mots justes pour exprimer spécifiquement ma reconnaissance à tous ceux qui ont contribue de près ou de loin à ce travail.

Avant tout, , je souhaite rendre grâce à **ALLAH**, le Clément et Miséricordieux de m'avoir donné la force, le courage et la patience de mener à bien ce modeste travail

Je tiens à remercier sincèrement mon encadreur :

TOUBA SOUNIA, de m'avoir suivi et dirigé tout au long de la réalisation de ce mémoire.

Je souhaite également remercier les membres du jury :

Mouloud Cherfaoui.

Dhiabi Samra

Maitres de conférences à UNIVERSITÉ MOHAMED KHIDER, BISKRA, pour l'honneur qu'ils m'ont fait en portant leur attention sur ce travail

Merci à toutes et à tous.

BENAISSA MOHAMMED

Table des matières

Table des figures

Liste des tableaux

Introduction

Dans une étude statistique, un dénombrement complet de la population est très souvent pratiquement impossible, soit parce que la population totale est inconnue, soit parce qu'elle comprend beaucoup trop d'individus pour qu'une telle étude soit complètement réalisable. Toutefois, le but d'une étude statistique est d'obtenir des connaissances sur I'ensemble de la population. Or, si une étude sur I'ensemble de la population est difficilement envisageable, il nous faut malgré tout trouver d'autres moyens pratiques d'y parvenir. Un moyen efficace est de procéder a un échantillonnage, qui consiste a choisir parmi les éléments de la population un certain nombre d'unités pour lesquelles nous obtiendrons des observations. Si l'échantillon étudié est bien choisi, les observations permettront d'acquérir les connaissances voulues sur la population à étudier avec un degré spécifié de précision.

Le but de ce mémoire est de présenter les différentes méthodes d'échantillonnage et d'estimation.

Ce mémoire est divisé en trois chapitres :

Dans le premier chapitre, nous nous limitons à un bref rappel sur des notions et définitions de base sur les paramètres statistique et sur quelques éléments fondamentaux(fonction de vraisemblance, statistique exhaustive) que nous jugeons utiles pour la suite de notre travail.

Dans le deuxième chapitre nous nous intéressons essentiellement aux estimation ponctuelle , est partagé en deux section de base , première section étudie la qualité d'un estimateur et deuxième section étudie les méthodes d'estimation ponctuelle .

Dans le troisième chapitre,

à présenter l'estimation par l'intervalle de confiance pour évaluer la confiance que l'on peut avoir en une estimation et exemple appliqué sous logiciel R .

Chapitre 1

Généralités sur les paramètres

L'objectif de l'estimation statistique est le suivant : évaluer certaines grandeurs as- sociées à une population à partir d'observations faites sur un échantillon. Bien souvent, grandeurs sont des moyennes ou des variances. On prendra soin de distinguer ces grandeurs théoriques (inconnues et à estimer) de celles observées sur un échantillon.

1.1 Echantillon

Soit $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$une population de taille N . Soit X le caractère que líonvoudrait étudier sur cette population. Avec l'échantillon aléatoire simple : soit X_k le résultata aléatoire du k^{ieme} tirage, císt une variable aléatoire qui suit la même loi que X. On note x_k le résultat du k^{ieme} tirage et on note $(X_1, X_2, ..., X_n)$ le résultat aléatoire de ces n tirages. $(X_1, X_2, ..., X_n)$ est un n -uplet de variables aléatoires indépendantes de meme loi (celle de X). Il est appelé n-échantillon aléatoire simple.

1.2 statitique

Une statistique Y sur un échantillon $(X_1, X_2, ..., X_n)$ est une variable aléatoire fonction mesurable des X_i : $Y = f(X_1, X_2, ..., X_n)$ prend la valeur $f(x_1, x_2, ..., x_n)$.

Les statistiques sont utilisées pour estimer les caractéristiques de la population totale. Les statistiques les plus utilisées sont la moyenne empirique, la variance empirique, la fréquence empirique.

1.3 Estimateur

Soient $(X_1, X_2, ..., X_n)$ un n-échantillon aléatoire simple issu d'une variable aléatoire X (discrète ou continue) et θ un paramètre associé a la loi probabilité $\mathbb{P}_\theta$ de θ. Un estimateur de θ est une variable aléatoire T fonction des X_i

$$T = f(X_1, X_2, ..., X_n).$$

Si on considère n observations $x_1, x_2, ..., x_n$, l'estimateur T fournira une estimation de notée :

$$\hat{\theta} = f(x_1, x_2, ..., x_n).$$

1.4 Quelques statistiques classiques

Rappels :

- $E(aX + b) = aE(X) + b$.

- $E(X + Y) = E(X) + E(Y)$.

- $V(aX + b) = a^2V(X)$.

- $V(X) = E(X^2) - [E(X)]^2 = E([X - E(X)]^2)$.

si X, Y indépendantes,

- $V(X + Y) = V(X) + V(Y)$.

1.5 La moyenne empirique et la variance empirique

1.5.1 La moyenne empirique

Définition 1.5.1 *On appelle moyenne empirique de l'échantillon* $(X_1, X_2, ..., X_n)$ *de* X*, la statistique*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Sa réalisation est $\overline{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$ *(qui est la moyenne de l'échantillon) aussi appelée moyenne observée. (on verra plus tard que* $\overline{X}$ *estimera l'espérance* $E(X)$*)*

propriétés

1. Soit $(X_1, X_2, ..., X_n)$ un n-échantillon issu d'une variable aléatoire X de moyenne μ et de variance σ^2. Alors $E(X) = \mu$ (on dit que X est une statistique sans biais).
   En effet : $E(\overline{X}) = \frac{1}{n}\sum\limits_{i=1}^{n} E(X_i) = \mu$.
2. $Var(\overline{X}) = \frac{\sigma^2}{n}$.
   En effet : $Var(\overline{X}) = \dfrac{1}{n^2}\sum\limits_{i=1}^{n} Var(X_i) = \dfrac{1}{n^2} n\sigma^2 = \dfrac{1}{n}\sigma^2$.
3. $\overline{X} \xrightarrow{ps} \mu$ (Loi forte des grands nombres).

1.5.2 La variance empirique

Définition 1.5.2 *On appelle variance empirique de l'échantillon* $(X_1, X_2, ..., X_n)$*de* X *, la statistique*

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

Sa réalisation est $s^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$ *(qui est la variance de l'échantillon), aussi appelée variance observée.*

propriétés

1. $E(S^2) = \frac{\sigma^2(n-1)}{n}$,
2. $S^2 \xrightarrow{ps} \sigma^2$,
3. $Var(S^2) = \frac{(n-1)}{n^3}\left[(n-1)\mu_4 - (n-3)\sigma^4\right]$, ou $\mu_4 = E(X - \overline{X})^4$.

1.6 Fonction de vraisemblance

Comme un échantillon aléatoire simple $(X_1, X_2, ..., X_n)$ est une variable aléatoire à n dimention, on peut parler de la disribution de probabilité jointe $(X_1, X_2, ..., X_n)$. Cette distribution est appelée vraisemblance de l'échantillon et elle est notée par $L(x_1, x_2, ..., x_n; \theta)$

$\forall\theta \in \Theta$

Soit

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

s'il s'agit des variables continues

Ou $f(x_i; \theta)$ est la densité de probabilité de X_i.

s'il s'agit des variables discrètes alors la vraisemblance devient :

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} P_\theta(X = x_i).$$

1.7 Statistique exhaustive

Dans un probléme statistique ou figure un paramètre θ inconnu, un échantillon nous apporte certaine information sur ce paramètre. Lorsque l'on résume cet échantillon par une statistique, il s'agit de ne pas perdre cette information. Une statistique qui conserve l'information sera dite suffisante (exhaustive).

1.7.1 Définition d'une statistique exhaustive

Soit $(X_1, X_2, ..., X_n)$ un n-échantillon et $L(x,\theta)$ sa vraisemblance. Soit T une statistique fonction de $X_1, X_2, ..., X_n$ de loi $g(t,\theta)$ dans le cas continue $P(T=t)$ dans le cas discret. T est dite exhaustive si $L(x,\theta) = g(t,\theta)h(x)$ (principe de factorisation, de Neyman-Fisher), en d'autre termes si la loi conditionnelle de l'échantillon est indépendante de θ.

$P(X = x|T(x) = t) = k(x,t)$ indépendante de θ.

Exemple 1.7.1 *Soit $X_1, X_2, ..., X_n$ issu de $X \rightsquigarrow P(\theta)$, θ inconnu. La statistique $T = \frac{1}{n}\sum_{i=1}^{n} X_i$ est-elle exhaustive pour θ ?*

On a :

$$f(x,\theta) = e^{-\theta}\frac{\theta^x}{x!}.$$

la loi de Poisson est stable donc $T = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow P(n\theta)$

$g(x,\theta) = e^{-n\theta}\frac{n\theta^t}{t!}$, $\forall x_1, x_2, ..., x_n \in \mathbb{N}^n, \forall t \in \mathbb{N}$.

Notons P_θ la quantité : $\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n | T = t)$.

$$\begin{aligned}
\mathbb{P}_\theta &= \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n, T = t)}{P(T=t)} \\
&= \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n, X_n = t - \sum_{i=1}^{n-1} x_i)}{\mathbb{P}(T=t)} \\
&= \frac{\mathbb{P}_\theta(X_1 = x_1), ..., \mathbb{P}_\theta(X_n = x_n)}{\mathbb{P}(T=t)} \\
&= \frac{\prod_{i=1}^{n} f(x_i,\theta)}{f(t,\theta)} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.\frac{t!}{e^{-n\theta}(n\theta)^t} \\
&= \frac{t!}{\prod_{i=1}^{n} x_i! n^t} = \frac{\left(\sum_{i=1}^{n} x_i\right)!}{\prod_{i=1}^{n} x_i!}\left(\frac{1}{n}\right)^{\sum_{i=1}^{n} x_i} \quad \textit{indépendante de } \theta.
\end{aligned}$$

Donc T est exhaustive pour θ.

Ou bien :

$$
\begin{aligned}
L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta) = \prod_{i=1}^{n} \left(\frac{e^{-\theta}\theta^{x_i}}{x_i!} \right) \quad &= \frac{e^{-n\theta}\theta^{t}}{\prod\limits_{i=1}^{n} x_i!} \\
&= \frac{e^{-n\theta}\left(n\theta\right)^{t}}{t!} . \frac{t!}{n^{t}\left(\prod\limits_{i=1}^{n} x_i!\right)} \\
&= g(t,\theta)h(x).
\end{aligned}
$$

D'aprés le principe de factorisation, T est exhaustive pour θ.

Chapitre 2

Estimation ponctuelle

On souhaite 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 de$(X_1, ..., X_n)$ dont la réalisation est envisagée comme une "bonne valeur" du paramètre θ . On parle d'estimation de θ associée à cet estimateur la valeur observée lors de l'expérience, c'est-à-dire (i.e) la valeur prise par la fonction (fct) au point

observé$(X_1, ..., X_n)$.

2.1 Qualité d'un estimateur

Dans cette partie on va introduire quelques qualités d'un estimateur .

2.1.1 Estimateur sans biais

Définition 2.1.1 *On appelle **biais** de T pour θ la valeur*

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

*Un estimateur Test dit **sans biais** si $E(T) = \theta$.*

2.1.2 Estimateur convergent

Définition 2.1.2 *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.*

Théorème 2.1.1 *Si T est convergent et de variance tendant vers 0 lorsque n tend vers l'infini alors T est consistant.*

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

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

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

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

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

De façon générale, on peut écrire

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

ainsi

- la grandeur $T - E(T)$ représente les fluctuations de T autour de sa moyenne.
- $E(T) - \theta$ représente l'erreur systématique (biais).

2.1.3 Erreur quadratique moyenne

Définition 2.1.3 *La qualité d'un estimateur se mesure également par* ***l'erreur quadratique moyenne*** *(ou risque quadratique) définie par $E((T-\theta)^2)$.*

Théorème 2.1.2 *Soit T un estimateur du paramètre θ à étudier. On a :*

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

Preuve. (évident). ■

2.1.4 Quelques estimateurs clssiques

1. $\overline{X}$ est un estimateur sans biais de la moyenne μ. Son estimation $\overline{x}$ est la moyenne observée dans une réalisation de l'échantillon.
2. $\tilde{S}^2$est un estimateur consistant de σ^2 (mais biaisé).
3. $S^2 = \dfrac{n}{n-1}\tilde{S}^2$ est un estimateur sans biais et consistant de σ^2.Son estimation $S^2 = \dfrac{n}{n-1}\ \sigma_\varepsilon^2$ est où σ_ε 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.

2.2 Estimation ponctuelle des paramétres usuels

2.2.1 Estimation de la moyenne

Soit X une v.a dont on veut estimer la moyenne (ou espérance) $\mu = E(X)$ à partir d'un n -échantillon $(X_1, ..., X_n)$ de X On ne suppose rien sur la loi de X.

Théorème 2.2.1 *la moyenne empirique* $\overline{X} = \dfrac{1}{n}(X_1 + ... + X_n)$. *, est un estimateur efficace de* μ.*car sans biais* $E(\overline{X}) = \mu$ *et de plus* $V(\overline{X}) = \dfrac{V(X)}{n} \to 0$ *pour* $n \to \infty$, *et* $\forall\ T$, *un autre estimateur de* μ ,$V(T) > V(\overline{X})$.

– $\overline{x}$ *est la réalisation de* $\overline{X}$ *et donc une estimation efficace de* μ.

2.2.2 Estimation de la variance d'une population Gaussienne

Soit X une v.a qui suit une loi normale $N\ (\mu, \sigma)$. On veut estimer la variance σ de X.

a) μ connue

1. – $T^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i - \mu)^2$. est un estimateur efficace de σ^2.

En effet :

$$\begin{aligned}
E\left(T^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\frac{1}{n}\mu X_i + \mu^2\right) \\
&= \frac{1}{n}E\left(\sum_{i=1}^{n}X_i^2\right) - 2\mu\frac{1}{n}\sum_{i=1}^{n}E\left(X_i\right) + \mu^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}E\left(X_i^2\right) - \mu^2 = \frac{1}{n}\sum_{i=1}^{n}\left[V\left(X_i\right) + \left(E\left(X_i\right)\right)^2\right] - \mu^2 \\
&= \sigma^2 + \mu^2 - \mu^2 = \sigma^2.
\end{aligned}$$

donc sans biais

$$\begin{aligned}
V\left(T^2\right) &= V\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n}V\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n^2}\sum_{i=1}^{n}V\left((X_i-\mu)^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E((X_i-\mu)^4) - (E((X_i-\mu)^2))^2\right] = \ldots \to 0
\end{aligned}$$

b) μ inconnue

- $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2.$,i.e la variance empirique, est un estimateur biaisé de σ^2, mais asymptotiquement sans biais.

  En effet :

$$\begin{aligned}
E(S^2) &= \frac{n-1}{n}\sigma^2 \\
B(S^2) &= E(S^2) - \sigma^2 = \left(1-\frac{1}{n}\right)\sigma^2 = -\frac{1}{n}\sigma^2 \\
V(S^2) &\to 0 \text{ pour } n \to \infty.
\end{aligned}$$

- $\left(S'\right)^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2.$ est un estimateur sans biais de σ^2.

  En effet :

$$E\left(\left(S'\right)^2\right) = \frac{n}{n-1}E(S^2) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

  donc sans biais

- n grand, $E(S^2) \approx E((S')^2)$ et on préfère S^2.
- n petit, on préfère $(S')^2$.

2.3 L'information de Fisher

Définition 2.3.1 *On appelle quantité d'information de Fisher $I_n(\theta)$ apportée par un échantillon sur le paramètre θ, la quantité suivante positive ou nulle (si elle existe) :*

$$I_n(\theta) = E\left[\left(\frac{\partial \ln L(X,\theta)}{\partial \theta}\right)^2\right].$$

avec $L(X,\theta) = \prod_{i=1}^{n} f(X_i,\theta)$.

Théorème 2.3.1 *Si le domaine de définition de X ne dépend pas de θ alors :*

$$I_n(\theta) = E\left[\left(\frac{\partial^2 \ln L(X,\theta)}{\partial \theta^2}\right)\right].$$

2.3.1 Propriété d'additivité de l'information de Fisher :

Soit Y v.a indépendante de X dont la loi dépend du même paramètre θ.Soit $f(x,\theta), g(y,\theta), h(x,y,\theta)$ les densités de X, Y et du couple (X,Y) (resp.), d'information de Fisher $I_x(\theta)$; $I_y(\theta)$, $I(\theta)$ (resp.) Si les domaines de définition de X, Y ne dépendent pas de θ alors :

$$I(\theta) = I_x(\theta) + I_y(\theta).$$

Preuve. Puisque X, Y sont indépendantes :

$$\begin{aligned} h(x,y,\theta) &= f(x,\theta).g(y,\theta) \\ \ln h(x,y,\theta) &= \ln f(x,\theta). \ln g(y,\theta) \\ \frac{\partial^2 \ln h(x,y,\theta)}{\partial \theta^2} &= \frac{\partial^2 \ln f(x,\theta)}{\partial \theta^2}.\frac{\partial^2 \ln g(y,\theta)}{\partial \theta^2} \\ E\left(\frac{\partial^2 \ln h(X,Y,\theta)}{\partial \theta^2}\right) &= E\left(\frac{\partial^2 \ln f(X,\theta)}{\partial \theta^2}\right) E\left(\frac{\partial^2 \ln g(Y,\theta)}{\partial \theta^2}\right) \\ -I(\theta) &= -I_x(\theta) - I_y(\theta) \Longrightarrow I(\theta) = I_x(\theta) + I_y(\theta). \end{aligned}$$

■

Conclusion 2.3.1 *:Si le domaine de définition ne dépend pas de θ alors :* $I_n(\theta) = nI_1(\theta)$.

avec

$$I_1(\theta) = -E\left(\frac{\partial^2 \ln f(X,\theta)}{\partial \theta^2}\right).$$

i.e : que chaque observation a la même importance. Ce qui n'est pas le cas pour la loi uniforme sur $[0, \theta]$ où la plus grande observation est la plus intéressante.

Exemple 2.3.1 *Soit $X \rightsquigarrow N(\mu, \sigma^2)$. Calculer l'information de Fisher apportée par un n-échantillon issu de X sur le paramètre μ :*

$$\begin{aligned}
I_n(\mu) &= nI_1(\mu) \\
I_1(\mu) &= -E\left(\frac{\partial^2 \ln f(X,\mu)}{\partial \mu^2}\right) \\
f(x,\mu) &= \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \\
\ln f(x,\mu) &= -\ln\left(\sigma\sqrt{2\pi}\right)\frac{-1}{2\sigma^2}(x-\mu)^2 \\
\frac{\partial \ln L}{\partial \mu} &= \frac{1}{\sigma^2}(x-\mu) \\
\frac{\partial^2 \ln L}{\partial \mu^2} &= \frac{-1}{\sigma^2} \Longrightarrow I_n(0) = \frac{n}{\sigma^2}.
\end{aligned}$$

2.4 Information de Fisher et suffisance

On montre que l'information portée par une statistique est inférieure ou égale à celle apportée par un échantillon. En effet : soit T la statistique de densité $g(t, \theta)$ que l'on substitue à l'échantillon.

On a $L(x, \theta) = g(t, \theta).h(x, \theta|t)$ où $h(x, \theta|t)$ est la densité conditionnelle de l'échantillon.

$$\begin{aligned}
\ln L(x,\theta) &= \ln g(t,\theta) + \ln h(x,\theta|t) \\
\frac{\partial^2 \ln L(x,\theta)}{\partial \theta^2} &= \frac{\partial^2 \ln g(t,\theta)}{\partial \theta^2} + \frac{\partial^2 \ln h(x,\theta|t)}{\partial \theta^2}.
\end{aligned}$$

En prennant l'espérance mathématique :

$$I_n(\theta) = I_T(\theta) - E\left[\frac{\partial^2 \ln h}{\partial \theta^2}\right] = I_T(\theta) + I_{n/T}(\theta)$$
$$I_{n/T}(\theta) = -E\left[\frac{\partial^2 \ln h}{\partial \theta^2}\right] \geq 0.$$

D'où :

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

Si T est suffisante $I_n(\theta) = I_T(\theta)$.

La réciproque est vraie si le domaine de X est indépendant de θ.

2.5 Borne de Freshet-Damois-Cramer-Rao (FDCR)

Le résultat suivant nous indique que la variance d'un estimateur ne peut être inférieur à une certaine borne, qui dépend de la quantité d'information de Fisher apportée par l'échantillon sur le paramètre θ.

Théorème 2.5.1 *(Inégalité de FDCR)*

Si le domaine de définition de X ne dépend pas de θ alors pour tout estimateur sans biais :

$$Var(T) \geq \frac{1}{I_n(\theta)}.$$

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

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

Preuve.

$$\begin{aligned}
cov\left(T,\frac{\partial\ln L}{\partial\theta}\right) &= E\left(T\frac{\partial\ln L}{\partial\theta}\right) \text{ puisque } \frac{\partial\ln L}{\partial\theta} \text{ est centrée} \\
&= \int_{\mathbb{R}^n} t\frac{\partial\ln L}{\partial\theta}L(x;\theta)dx \\
&= \int_{\mathbb{R}^n} t\frac{\partial L(x;\theta)}{\partial\theta}dx \\
&= \frac{d}{d\theta}\int_{\mathbb{R}^n} tL(x;\theta)x \\
&= \frac{d}{d\theta}E'(T) \\
&= h'(\theta).
\end{aligned}$$

D'autre part l'inégalité de Schwartz donne :

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

i.e :

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

■

2.6 Méthodes de calcul d'un estimateur

2.6.1 Méthode des moments

Soit $(X_1, X_2, ..., X_n)$ un n-échantillon issu d'une v.a X de densité $f(x, \theta_1, \theta_2, ..., \theta_k)$ où : $\theta_1, \theta_2, ..., \theta_k$ sont des paramètres inconnus. La méthode des moments consiste à estimer les paramètres $\theta_1, \theta_2, ..., \theta_k$, en égalisant les moments empiriques calculés à partir de l'échantillon avec les moments théoriques de même ordre.

Soit $\mu'_r = E(X^r)$, $r = 1, 2, ..., k$ moments d'ordre r de la population (théorique) et on note $M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$. moment empirique d'ordre r de l'échantillon.

La solution du système $M_r = \mu'_r$, $r = \overline{1,k}$ nous donne les estimateurs de $\theta_1, \theta_2, ..., \theta_k$

$$\begin{cases} M_1 = \mu'_1 \\ M_2 = \mu'_2 \\ \quad . \qquad\qquad k \text{ équations à } k \text{ inconnus}; \\ \quad . \\ \quad . \\ M_r = \mu'_r. \end{cases}$$

Remarque 2.6.1 *Dans la plupart des cas, les estimateurs obtenus par la méthode des moments sont consistants, convergents, asymptotiquement normaux mais en général ne sont pas efficaces.*

Exemple 2.6.1 *Soit* $(X_1, X_2, ..., X_n)$ *un* n*-échantillon issu de* $X \rightsquigarrow N(\mu, \sigma^2)$*,où* μ, σ^2 *sont inconnus. Estimer* μ *et* σ^2 *par la méthode des moments.*

$$\begin{aligned} \mu'_1 &= E(X) = \mu \\ \mu'_2 &= E\left(X^2\right) = Var(X) + (E(X))^2 = \sigma^2 + \mu^2 \\ M_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ M_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

$$\begin{cases} \mu = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{cases} \quad \textit{L'estimateur de } \mu \textit{ est } \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \widehat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}\right)^2 \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2. \end{aligned}$$

2.6.2 Méthode du maximum de vraisemblance

Il s'agit de trouver un estimateur de θ, qui maximise la fonction de vraisemblance de l'échantillon. La valeur $\widehat{\theta}$ qui maximise $L(x,\theta)$ serait un bon estimateur car elle donne la plus grande probabilité pour cet échantillon.

Cette méthode consiste à résoudre : $\begin{cases} \frac{\partial L(x,\theta)}{\partial\theta} = 0, & \text{pour trouver } \widehat{\theta}; \\ \frac{\partial^2 L(x,\theta)}{\partial\theta^2} \mid_{\theta=\widehat{\theta}} < 0, & \text{pour assurer l'existence du } \max_{\theta\in\Theta} L(x,\theta). \end{cases}$

Remarque 2.6.2 *Maximiser* $L(x,\theta)$ *revient à maximiser* $\ln L(x,\theta)$ *. Il est plus commode de maximiser* $\ln L(x,\theta)$

Exemple 2.6.2 *Soit* $(X_1, X_2, ..., X_n)$ *un n-échantillon issu de* $X \rightsquigarrow N(\mu,\sigma^2)$ *,trouver l'estimateur EMV de* μ *et* σ^2.

La densité deX est :

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

$$\begin{aligned}
\frac{\partial \ln L(x,\mu,\sigma^2)}{\partial\mu} &= \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma^2}\right) = 0 \\
&\Longrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \\
&\Longrightarrow n\mu = \sum_{i=1}^{n} x_i. \\
\widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i. \\
\frac{\partial^2 \ln L(x,\mu,\sigma^2)}{\partial\mu^2} &= \frac{-1}{\sigma^2} < 0.
\end{aligned}$$

d'où : $\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}$ *est l'EMV de* μ.

$$
\begin{aligned}
L(x,\mu,\sigma^2) &= \frac{1}{(\sigma^2 2\pi)^{\frac{n}{2}}}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\\
\ln L(x,\mu,\sigma^2) &= \frac{-n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\\
\frac{\partial \ln L}{\partial \sigma^2} &= \frac{-n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{2\sum\limits_{i=1}^{n}(x_i-\mu)^2}{4\left(\sigma^2\right)^2}\\
&= \frac{-n}{2\sigma^2} + \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\left(\sigma^2\right)^2} = 0\\
&\Longrightarrow \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{2\left(\sigma^2\right)^2} = \frac{n}{2\sigma^2}\\
&\Longrightarrow \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} = n\\
&\Longrightarrow \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 .
\end{aligned}
$$

On a :

$$
\begin{aligned}
\frac{\partial^2 \ln L(x,\mu,\sigma^2)}{\partial\left(\sigma^2\right)^2} &= \frac{2n}{4\left(\sigma^2\right)^2}\frac{-4\sigma^2\sum\limits_{i=1}^{n}(x_i-\mu)^2}{4\left(\sigma^2\right)^4}\\
&= \frac{n}{2\left(\sigma^2\right)^2} - \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\left(\sigma^2\right)^3}\Big|_{\sigma^2=\widehat{\sigma}^2}\\
&= \frac{1}{\left(\widehat{\sigma}^2\right)^2}\left[\frac{n}{2}\frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\right]_{\sigma^2=\widehat{\sigma}^2}\\
&= \frac{1}{\left(\widehat{\sigma}^2\right)^2}\left[\frac{n}{2}-n\right] = \frac{-n}{2\widehat{\sigma}^2} < 0.
\end{aligned}
$$

D'où : $\begin{cases} S_1^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\mu)^2\,, \textit{si } \mu \textit{ est connue .}\\ S_2^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2\,, \textit{si } \mu \textit{ est inconnu .}\end{cases}$.

Chapitre 3

Estimation par intervalle de confiance

L'estimation ponctuelle d'une paramètre θ donne une valeur numérique unique à ce paramètre, mais n'apporte aucune information sur la précision des résultats.

Pour évaluer la confiance que l'on peut avoir en une estimation, il est nécessaire de lui associer un intervalle qui contient, avec une certaine probabilité, la vraie valeur du paramètre, c'est l'estimation par intervalle de confiance.

3.1 Généralité

Il est plus réaliste et plus intéressant de fournir une estimation du type

$$t_1 < \theta < t_2$$

plutôt que d'écrire séchement $\theta = t$, car on sait que la valeur estimée t diffère toujours de la valeur exacte du paramètre recherché, θ. Il est donc souhaitable de donner la précision de l'estimation en acceptant de faire une erreur α sur celle-ci.

Définition 3.1.1 *Soit X une v.a. dont la loi dépend d'un paramètre inconnu θ, on appelle INTERVALLE DE CONFIANCE pour θ de niveau $1-\alpha$ (ou de seuil α), un intervalle qui a la probabilité $1-\alpha$ de contenir la vraie valeur de θ.*

$[t_1, t_2]$ *est un intervalle de confiance de niveau* $1-\alpha$ *pour* θ *signifie*

$$\mathbb{P}(t_1 < \theta < t_2) = 1 - \alpha.$$

(plus le niveau de confiance est élevé, plus la certitude est grande que la méthode d'estimation produira une estimation contenant la vraie valeur de θ*).*

• les niveaux de confiance les plus fréquemment utilisés sont 90%, 95%, 99%.

• α est appelé le seuil (le risque) ; on choisira dans la plupart des cas un intervalle à risques symétriques, c-à-d t.q.

$$\mathbb{P}(\theta < t_1) = \frac{\alpha}{2}, \mathbb{P}(\theta > t_2) = \frac{\alpha}{2}.$$

Remarque 3.1.1 *: Si on augmente le niveau de confiance* $1-\alpha$*, on augmente la longueur de l'intervalle.*

3.2 Principe de construction

Le principe de la méthode d'estimation par intervalle de confiance est le suivant :

Soit T un estimateur de θ (meilleur estimateur possible) dont on connait la loi en fonction

de θ.

On détermine par la suite un intervalle de probabilité de niveau $1-\alpha$ pour T i.e :

$$\mathbb{P}(t_1(\theta) < \theta < t_2(\theta)) = 1 - \alpha.$$

Il faut ensuite inverser cet intervalle pour T, dont les bornes dépendent de θ pour obtenir un intervalle pour,

$$\begin{aligned}&\text{i.e } \mathbb{P}_\theta(a(T) \leq \theta \leq b(T)) = 1 - \alpha,\\ &\text{ou bien } \mathbb{P}(t_1^{-1}(t) < \theta < t_2^{-1}(t)) = 1 - \alpha.\end{aligned}$$

Graphiquement :

Verticalement, on lit l'intervalle de probabilité et horizontalement l'intervalle de confiance

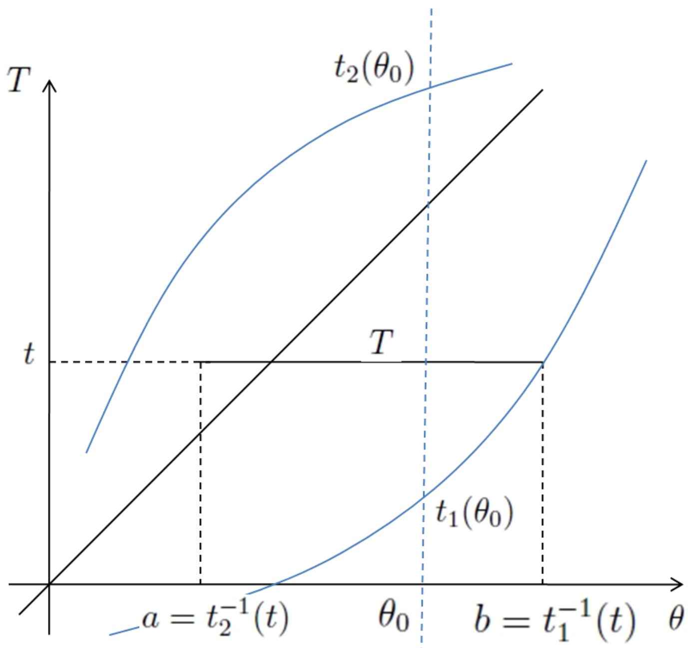

FIG. 3.1 – Intervalle de probabilité et intervalle de confiance

Remarque 3.2.1 *Si $1 - \alpha$ augmente, on augmente la longueur de l'intervalle de probabilité donc les courbes sécartent.*

Remarque 3.2.2 *Si n augmente, comme T est supposé convergent, alors $Var(T)$ diminue donc $[t_1, t_2]$ diminue est les courbes se rapproche de la première bissectrice, donc l'intervalle de confiance diminue.*

3.3 Intervalle de confiance pour les paramètres d'une loi normale

3.3.1 Théorème centrale limite

L'étude de sommes de variable indépendance et de meme loi joue un role capital en statistique.

Le théorème suivant connu sous le nom de théorème centrale limite (il vaudrait mieuxdire théorème de la limite centrée) établit la convergence vers la loi de Gauss sous de hypothèses peu contraignantes.

Théorème 3.3.1 *Soit (X_n) une suit de vas indépendantes de meme loi d'espérance μ et d'écart-type σ alors :*

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + X_2 + ... + X_n - \mu}{\sigma}\right) \xrightarrow{\mathcal{L}} LG(0;1).$$

Théorème 3.3.2 *Si $(X_1, ..., X_n)$ est un échantillon de la loi $N(\mu, \sigma^2)$, alors :*

1. $\dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ *suit la loi normale $N(0,1)$ (d'aprés le théorème centrale limite(TCL)).*
2. $\dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n-1}}}$ *suit une loi de student $\check{T}_{n-1}$ de degré de liberté $(n-1)$.*
3. $\dfrac{nS^2}{\sigma^2}$ *suit la loi du khi-deux χ^2_{n-1} degré de liberté $(n-1)$.*

3.4 Intervalle de confiance pour une moyenne

a <u>Cas où n, la taille de l'échantillon est petite $n < 30$:</u>

On suppose que $X \sim N(\mu, \sigma)$.

On distingue deux cas σ connu et σ inconnu.

a-1) <u>σ connu :</u>

- $\overline{X} \sim N(\mu, \dfrac{\sigma}{\sqrt{n}})$ d'aprés un résultat du théorème (ou bien $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$)
- On se fixe le risque α et on cherche dans la table de la loi normale la valeur $u_{1-\frac{\alpha}{2}}$ telle que

$$P(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < u_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

$$\Updownarrow$$

$$P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < u_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}.$$

$u_{1-\frac{\alpha}{2}}$ est le fractile d'order $1-\dfrac{\alpha}{2}$ de la loi normale centrée réduite.

$$P(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < u_{1-\frac{\alpha}{2}}) = 1-\alpha.$$
$$\Updownarrow$$
$$P(\overline{X} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

Conclusion 3.4.1 *Si $\overline{x}$ est une réalisaion de $\overline{X}$, l'intervalle de confiance de μ de seuil α est*

$$I = \left[\overline{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right].$$

Exemple 3.4.1 $n = 15$, $\alpha = 3.75$, $\alpha = 5\%$, $\sum_{i=1}^{15} x_i = 2400$ *alors* $x = 2400/15 = 160$, $u_{1-\frac{\alpha}{2}} = 1.96$ *car* $P(U < -1.96) = 0.025$

on suppose X gaussienne et on obtient l'intervalle de confiance :

$$\left[160 - 1.96\frac{3.57}{\sqrt{15}}, 160 + 1.96\frac{3.57}{\sqrt{15}}\right].$$

a-2) <u>σ inconnu :</u>

- $\dfrac{\overline{X}-\mu}{S\sqrt{n-1}} \sim t_{n-1}$ d'aprés le théorème.
- On chareche dans la tabe de la loi de student α étant fixé, la valeur $t_{n-1(1-\frac{\alpha}{2})}$ telle que

$$P(-t_{n-1(1-\frac{\alpha}{2})} < \frac{\overline{X}-\mu}{S/\sqrt{n-1}} < t_{n-1(1-\frac{\alpha}{2})}) = 1-\alpha$$
$$\Updownarrow$$
$$P(\frac{\overline{X}-\mu}{S/\sqrt{n-1}} < t_{n-1(1-\frac{\alpha}{2})}) = 1-\alpha/2$$

On a

$$P(-t_{n-1(1-\frac{\alpha}{2})} < \frac{\overline{X}-\mu}{S/\sqrt{n-1}} < t_{n-1(1-\frac{\alpha}{2})}) = 1-\alpha$$
$$\Updownarrow$$
$$P(\overline{X} - t_{n-1(1-\frac{\alpha}{2})}\frac{S}{\sqrt{n-1}} < \mu < \overline{X} + t_{n-1(1-\frac{\alpha}{2})}\frac{S}{\sqrt{n-1}}) = 1-\alpha$$

Conclusion 3.4.2 *Si $\overline{x}$ est une réalisationde $\overline{X}$ et s une réalisation de S, l'intervalle de confiance de μ de seuil α est*

$$I = \left[\overline{x} - t_{n-1(1-\frac{\alpha}{2})}\frac{S}{\sqrt{n-1}}, \overline{x} - t_{n-1(1-\frac{\alpha}{2})}\frac{S}{\sqrt{n-1}}\right].$$

Exemple 3.4.2 $n = 30$, $\sum_{i=1}^{30} x_i = 1673$, $\sum_{i=1}^{30} x_i^2 = 98285$, $\alpha = 10\%$ *alors* $\overline{x} = 55.77$, $s^2 = 165,87$, $s = 12.88$, $t_{20(10\%)} = 1.699$

$$I = \left[55.77 - 1.699\frac{12.88}{\sqrt{29}}, 55.77 + 1.699\frac{12.88}{\sqrt{29}}\right]$$

b Cas où n la taille de l'échantillon est grande n > 30 :

Il n'est plus nécessaire de supposer que$X$ est Gaussienne.

b-1) σ connu :

• D'aprés le théorème $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ pour $n \to \infty$

La démarache est la meme que a-1)

• Conclusion : Si $\overline{x}$ est une réalisationde $\overline{X}$ et s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$I = \left[\overline{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

b-2) σ inconnu :

On peut prendre comme intervalle de confiance celui de la section a-2). On peut également utiliser l'approximation suivante :

• $\frac{\overline{X} - \mu}{S\sqrt{n}} \sim N(0,1)$

• On se fixe l'erreur α et on charche dans la table de la loi norale la valeur $u_{1-\frac{\alpha}{2}}$ telle que

$$P(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < u_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

$$\Updownarrow$$

$$P(\frac{\overline{X} - \mu}{S/\sqrt{n}} < u_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}.$$

On a

$$P(-u_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < u_{1-\frac{\alpha}{2}}) = 1-\alpha.$$

$$\Updownarrow$$

$$P(\overline{X} - u_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}) = 1-\alpha.$$

Conclusion 3.4.3 *Si $\overline{x}$ est une réalisaion de $\overline{X}$, et s est une réalisation de S, l'intervalle de confiance de μ de seuil α est :*

$$I = \left[\overline{x} - u_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{x} - u_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right].$$

Remarque 3.4.1 *Plus n est grand, plus I es petit(car $1/\sqrt{n}$ ou bien $1/\sqrt{n-1}$ est petit) et donc meilleure est la précision de l'estimation.*

3.5 Intervalle de confiance pour la variance d'une variable gaussienne

On suppose que $X \sim N(\mu,\sigma)$.

a μ connue (peu fréquent) :

- $T^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)$ est un estimateur efficace de σ^2 (voir estimation ponctuelle) ; sa réalisation est $t^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$. Comme $\frac{X_i-\mu}{\sigma} \sim N(0,1)$.

$\frac{nT^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ est une somme de n v.a indépendantes qui suivant la loi normale $N(0,1)$ et danc

$$\frac{nT^2}{\sigma^2} \sim \chi^2_n.$$

- L'erreur α étant fixée, on charche dans la table χ^2_n les valeurs $k_{n(1-\frac{\alpha}{2})}$ et $k_{n(1-\alpha/2)}$ telle que

$$P(k_{n(\frac{\alpha}{2})} < \frac{nT^2}{\sigma^2} < k_{n(1-\frac{\alpha}{2})}) = 1 - \alpha \tag{3.1}$$

$$\Uparrow$$

$$\begin{cases} P(\frac{nT^2}{\sigma^2} k_{n(1-\frac{\alpha}{2})}) = 1 - \alpha/2 \\ P(\frac{nT^2}{\sigma^2} k_{n(\frac{\alpha}{2})}) = \alpha/2 \end{cases}$$

$$(3.1) \Longleftrightarrow P(\frac{nT^2}{k_{n(1-\frac{\alpha}{2})}} < \sigma^2 < \frac{nT^2}{k_{n(\frac{\alpha}{2})}}) = 1 - \alpha$$

Conclusion 3.5.1 *Si t^2 est une réalisation de T^2, l'intervalle de confiance de σ^2 de seuil α est*

$$I = \left[\frac{nt^2}{k_{n(1-\frac{\alpha}{2})}}, \frac{nt^2}{k_{n(\frac{\alpha}{2})}} \right]$$

L'intervalle de confiance pour σ au seuil α est

$$I = \left[t\sqrt{\frac{n}{k_{n(1-\frac{\alpha}{2})}}}, t\sqrt{\frac{n}{k_{n(\frac{\alpha}{2})}}} \right]$$

$n = 10$, $\mu = 6$, $\sum_{i=1}^{10} x_i^2 = 402$, $\alpha = 5\%$

alors

$$t^2 = 40.2 - 36 = 4.2, k_{10(0.025)} = 20.5, k_{10(0.975)} = 3.25$$

$$I = \left[\frac{10 \times 4.2}{20.5}, \frac{10 \times 4.2}{3.25} \right] = [2.05, 12.92]$$

b <u>μ inconnu :</u>

- On a

$$\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

- On cherche dans la table χ^2_{n-1} les valeurs $k_{n-1(1-\frac{\alpha}{2})}$ et $k_{n-1(\alpha/2)}$ telle que

$$P(k_{n-1(\frac{\alpha}{2})} < \frac{nS^2}{\sigma^2} < k_{n-1(1-\frac{\alpha}{2})}) = 1 - \alpha \tag{3.2}$$

$$\Uparrow$$

$$\begin{cases} P(\dfrac{nS^2}{\sigma^2} k_{n-1(\frac{\alpha}{2})}) = \alpha/2 \\ P(\dfrac{nS^2}{\sigma^2} k_{n-1(1-\frac{\alpha}{2})}) = 1-\alpha/2 \end{cases}$$

$$(3.2) \Longleftrightarrow P(\frac{nS^2}{k_{n-1(1-\frac{\alpha}{2})}} < \sigma^2 < \frac{nS^2}{k_{n-1(\frac{\alpha}{2})}}) = 1-\alpha$$

Conclusion 3.5.2 *Si s^2 est une réalisation de S^2, l'intervalle de confiance de σ^2 de seuil α est*

$$I = \left[\frac{ns^2}{k_{n-1(1-\frac{\alpha}{2})}}, \frac{ns^2}{k_{n-1(\frac{\alpha}{2})}} \right]$$

L'intervalle de confiance pour σ au seuil α est

$$I = \left[s\sqrt{\frac{n}{k_{n-1(1-\frac{\alpha}{2})}}}, s\sqrt{\frac{n}{k_{n-1(\frac{\alpha}{2})}}} \right]$$

Remarque 3.5.1 *Si dans les tables du χ^2_n ou de t_n vous ne trouvez pas les valeurs correspondantes à $\frac{\alpha}{2}$ et $1-\alpha/2$, on prendra un risque asymétrique.*

$n = 30, \sum_{i=1}^{30} x_i = 1683, \sum_{i=1}^{30} x_i^2 = 98295, \alpha = 10\%$

alors

$$\bar{x} = 55.77, \; s^2 = 165.87, \; k_{29(0.05)} = 42.6, \; k_{29(0.95)} = 17.7$$

$$I = \left[\frac{30 \times 165.87}{42.6}, \frac{30 \times 165.87}{17.7} \right] = [116.81, 281.14]$$

3.6 Intervale de confiance pour une proportion

Soient $X_1, .., X_n$ i.i.d selon $B(p)$ et $X = \sum_{i=1}^{n} X_i \sim B(n,p)$ Notons $F_n = \dfrac{X}{n}$ estimateur sans biais de p

Dans le cas de grands échantillon :

En approchant une loi binomiale vers une loi normale on a : $\sqrt{n}\frac{F_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{loi} N(0,1)$.

Ce qui permet décrire $P\left(-u \leq \sqrt{n}\frac{F_n-p}{\sqrt{p(1-p)}} \leq u\right) = 1-\alpha$ où u est le fractile d'order $1-\dfrac{\alpha}{2}$ de la loi $N(0,1)$.

Et donc l'intervalle de confiance bilatéral symétrique pour une proportion p au niveau $1-\alpha$ s'obtient en résoivant l'inéquation $\sqrt{n}\left|\frac{F_n-p}{\sqrt{p(1-p)}}\right| \leq u$

Ce qui danne en notant f_n la réalisation de F_n sur l'échantillon :

$$IC(p) = \left[\frac{f_n + \frac{u^2}{2n} - \frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n} + f_n(1-f_n)}}{1+\frac{u^2}{n}}, \frac{f_n + \frac{u^2}{2n} + \frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n} + f_n(1-f_n)}}{1+\frac{u^2}{n}}\right]$$

Pour une taille d'échantillon importante, on considère l'approximation suivante :

$$IC(p) = \left[f_n - u\sqrt{\frac{f_n(1-f_n)}{n}}, f_n + u\sqrt{\frac{f_n(1-f_n)}{n}}\right]$$

Cette approximation est parfaitement justifiée sur le plan théorique

En effet, d'aprés le théorème de Slutsky. on a : $\sqrt{F_n(1-F_n)} \xrightarrow{p} \sqrt{p(1-p)}$

On déduit donc que $\sqrt{n}\frac{F_n-p}{\sqrt{F_n(1-F_n)}} \underset{n\to\infty}{\xrightarrow{loi}} N(0,1)$.

d'ou : $P\left(-u \leq \sqrt{n}\frac{F_n-p}{\sqrt{F_n(1-F_n)}} \leq u\right) = 1-\alpha$ où u est le fractile d'order $1-\dfrac{\alpha}{2}$ de la loi $N(0,1)$.

Quand n est grand liintervalle de confiance bilatéral symétrique pour une proportion s'écrit donc au niveau $1-\alpha$ sous forme indiquée :

$IC(p) = \left[f_n - u\sqrt{\frac{f_n(1-f_n)}{n}}, f_n + u\sqrt{\frac{f_n(1-f_n)}{n}}\right]$ f_n est la réalisation de F_n sur l'échantillon.

Sinon construction d'intervalles de confiance exacte :

On construit ces intervalles en considérant la fonction de répartition de la loi binomiale . Si la probabilité de recouvrement de l'intervalle ne vaut pas exactement $1-\alpha$, on prend d'itervalle ayant la plus petite probabilité de recouvrement parmi ceux ayant une probabilité de recouvrement supérieure à $1-\alpha$.

Exemple 3.6.1 *Dans un échantillon pris au hasard de* 100 *automobiliste, on constate que* 25 *d'entre eux possèdent une voiture de cylindrée supérieure à* 1600 *cc. Quel est l'intervalle de confiance pour le proportion d'automobilistes possèdent une voiture cylindrée supérieure à*1600 *cc (intervalle bilatéral, à risque symétrique, seuil de confiance* 95%*) ?*

- Estimation ponctuelle de p *: lenombre d'automobilistes possédant une voiture de cylindréesupérieure à* 1600 *cc dans un échantillon de taillen* $= 100$ *suit la loi binomiale* $B(100, p)$.

Un estimateur non biaisé pour la proportion p *est donné par la fréqunce* $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$.

D'ou l'estimation ponctuelle de p : $\hat{p} = \frac{25}{100}$

- Intervalle de confiance pour p*, on utilise l'approximation normale :*

$$P\left(\hat{p}_n - 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq p \leq \hat{p}_n + 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) = 0.95$$

D'ou les trois méthodes :

- remplacer p *par son estimation (*0.25*). Intervalle de confiance :*

$$P(0.25 - 0.085 < p < 0.25 + 0.085) = P(0.165 < p < 0.335) = 0.95$$

- remplacer p *par* $\frac{1}{2}$*. Intervalle de confiance:*

$$P(0.25 - 0.098 < p < 0.25 + 0.098) = P(0.152 < p < 0.348) = 0.95$$

- résoudre l'inéquation

$$p^2\left(1 + \frac{u^2}{n}\right) - 2p\left(\hat{p}_n + \frac{u^2}{2n}\right) + \hat{p}_n^2 \leq 0 \iff 1.038p^2 - 0.538p + 0.0625 \leq 0$$

On obtient pour intervalle de confiance :

$$I = [p_1; p_2] = [0.176; 0.342]$$

3.7 Intervalle de confiance pour la moyenne d'une loi quelconque

Quelle que soit la taille de l'échantillon, on prendra pour estimateurs sans biais de la moyenne, la statistique $\overline{X}$, et de la variance, la statistique $S^{'2}$. Si la taille de l'échantillon est grande, en pratique $n > 30$, grâce au théorème central limite, on utilise les résultats précédents pour calculer l'intervalle de confiance pour la moyenne.

Si la série prélevée est faible, il faut utiliser les lois suivies par les paramètres étudiés.

Exemple 3.7.1 *On considère un échantillon de* 40 *paquets de biscuits provenant d'une production de* 2000 *unités. Le poids moyen obtenu pour cet échantillon est égal à* 336*g et l'écart-type empirique, c'est-à-dire la quantité s, est égal à* $0,86g$. *Quelle est l'estimation, par intervalle de confiance, du poids moyen de ces paquets de biscuits, pour l'ensemble de la fabrication, avec les seuils de confiance* $0,90$, $0,98$ *et* $0,99$ *?*

La distribution du poids de ces paquets est inconnue ainsi que la variance de la population. Cependant, pour un grand échantillon ($n = 40 > 30$)*, l'intervalle de confiance est de la forme :*

$$\overline{x} - h\frac{s}{\sqrt{n-1}} < \mu < \overline{x} + h\frac{s}{\sqrt{n-1}} \qquad \text{avec } \overline{x} = 336,\ s = 0.86,\ n = 40.$$

Le fractile h *est lu sur la table de Student, degré de liberté* $n - 1 = 39$, *il dépend du seuil choisi.*

Les résultats sont regroupés dans le tableau suivant.

Niveaux de confiance	h	Limites inférieures	Limites supérieures
$0,90$	$1,685$	$336 - 0,2325335,768$	$336 + 0,2325336,232$
$0,98$	$2,429$	$336 - 0,3345335,666$	$336 + 0,3345336,334$
$0,99$	$2,708$	$336 - 0,3735335,627$	$336 + 0,3735336,373$

TAB. 3.1 – Limites de l'intervalle de confiance.

Ainsi, l'intervalle de confiance, qui a 98 *chances sur* 100 *de contenir la vraie valeur du poids moyen de cette fabrication de biscuits, est :* $[335,666,\ 336,334]$.

3.8 Intervalle de confiance de la différence des moyennes de deux lois normales indépendantes

Soient X_1 et X_2 deux variables aléatoires indépendantes, suivant les lois normales $N(\mu_1; s_1)$ et $N(\mu_2; s_2)$.

Le paramètre à estimer est la différence des moyennes $D = \mu_1 - \mu_2$.

On considère deux échantillons indépendants de taille n_1 pour la variable X_1 et de taille n_2 pour la variable X_2.

La variable aléatoire $\overline{D} = \overline{X}_1 - \overline{X}_2$ suit une loi normale d'espérance $\mu = \mu_1 - \mu_2$ et d'écart-type :

$$\sigma_{\overline{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Les écarts-types σ_1 et σ_2 étant inconnus, il faut chercher une estimation qui prend en compte toutes les données du problème. En utilisant la relation :

$$\frac{nS^2}{\sigma^2} = \chi^2(n-1)$$

et les propriétés de la loi du chi-deux, on montre que la statistique :

$$\frac{1}{n_1 + n_2 - 2}\left[\sum_1 (X_i - \overline{X})^2 + \sum_2 (X_i - \overline{X})^2\right]$$

est un estimateur sans biais de la variance commune des deux distributions prenant en compte toutes les observations (l'indice 1 indique que la sommation est faite sur le premier échantillon et l'indice 2 sur le deuxième).

Pour obtenir ce résultat, on a supposé que les écarts-types inconnus σ_1 et σ_2 étaient égaux. Un test de Fisher permet de vérifier ce résultat.

Toutes les propriétés démontrées entraînent que la variable aléatoire :

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\sum_1 (X_i - \overline{X})^2 + \sum_2 (X_i - \overline{X})^2}} \times \frac{n_1 + n_2 - 2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une loi de Student à $(n_1 + n_2 - 2)$ degrés de liberté.

Pour déterminer un intervalle de confiance pour la différence des moyennes de deux lois normales ou pour tester l'égalité de ces deux moyennes, il suffit d'utiliser la table statistique de Student.

Remarque 3.8.1 *En « mélangeant » les échantillons, on a perdu un degré de liberté.*

Exemple 3.8.1 *Des tubes sont construits par deux procédés de fabrications A et B. On dispose de deux échantillons de tubes dont on mesure les diamètres (en mm) :*

- Procédé A : échantillon 1 *de taille* $n_1 = 5$. *Résultats :* $63,12; 63,57; 62,81; 64,32; 63,76.$

- Procédé B : échantillon 2 *de taille* $n_2 = 4$. *Résultats :* $62,51; 63,24; 62,31; 62,21.$

On suppose que les diamètres des tubes de chaque fabrication sont distribués suivant des lois normales de variances inconnues mais égales.

- Intervalle de confiance bilatéral, à risques symétriques, pour la différence $\mu_1 - \mu_2$ *des moyennes des deux lois (seuil de confiance* $0,95$*) :*

- Échantillon 1 :

$$\overline{x}_1 = 63,516 \quad \sum (x_i - \overline{x})^2 = 1,3638.$$

Échantillon 2 :

$$\overline{x}_2 = 62,5675 \quad \sum (x_i - \overline{x})^2 = 0,6498.$$

- La variable $\dfrac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sum_1 (x_i - \overline{x})^2 + \sum_2 (x_i - \overline{x})^2}} \times \dfrac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ *suit une loi de Student à* $n_1 + n_2 - 2 = 7$ *degrés de liberté.*

- Avec les valeurs numériques données, on obtient :

$$\frac{0,9485-(\mu_1-\mu_2)}{\sqrt{1,3638+0,6498}}\times\frac{\sqrt{5+4-2}}{\sqrt{\frac{1}{5}+\frac{1}{4}}}=\frac{0,9485-(\mu_1-\mu_2)}{0,3598}$$

$$P((-2,365<t(7)<2,365)=0,95$$

D'où l'intervalle de confiance :

$$P(0,9485-2,365\times 0,3598<(\mu_1-\mu_2)<0,9485+2,365\times 0,3598)=0.95$$

L'intervalle $[0,10,1,80]$ *a une probabilité égale à* $0,95$ *de contenir la vraie valeur de la différence des moyennes des deux lois normales.*

3.9 Intervalle de confiance pour le rapport des variances de deux lois normales

Soient X_1 et X_2 deux variables aléatoires indépendantes, suivant les lois normales $N(\mu_1;s_1)$ et $N(\mu_2;s_2)$.

On a démontré le résultat suivant :

$$\frac{n_1S_1^2}{(n_1-1)\sigma_1^2}\times\frac{(n_2-1)\sigma_2^2}{n_2S_2^2}=F(n_1-1;n_2-1)$$

ou, en introduisant la statistique S'^2 :

$$\frac{S_1'^2}{\sigma_1^2}\times\frac{\sigma_2^2}{S_2'^2}=F(n_1-1;n_2-1)$$

Ce résultat permet de déterminer un intervalle de confiance pour le rapport de deux variances ou pour tester l'égalité de deux variances. On utilise les méthodes exposées dans les paragraphes précédents. On obtient par exemple :

$$P\left(\frac{s_1'^2}{s_2'^2} \times \frac{1}{F_{\alpha/2}(n_1-1;n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1'^2}{s_2'^2} \times \frac{1}{F_{1-\alpha/2}(n_1-1;n_2-1)}\right) = \alpha - 1$$

où s' désigne la valeur de la statistique $S^{'}$ donnée par l'échantillon considéré ; les valeurs de $F_{\alpha/2}(n_1-1;n_2-1)$ et de $F_{1-\alpha/2}(n_1-1;n_2-1)$sont lues sur les tables $7.1, 7.2, 7.3$ ou 7.4 selon la valeur du seuil a. On en déduit l'égalité ou non des variances.

Exemple 3.9.1 *Une firme d'expertises en contrôle de matériaux a demandé à un laboratoire d'effectuer une vérification sur la résistance à la compression de tubes fabriqués par deux usines U_1 et U_2.*

On supposera que la résistance à la compression des tubes provenant de chaque usine peut être considérée comme la réalisation de variables aléatoires suivant des lois normales.

Les résultats à la compression en kg/cm^2 sont résumés dans le tableau suivant.

3.9.1 Application

si lors d'un sondage d'une population vous observez la réponse de 100 personnes tirées au hasard dans la population et que vous observez que 42 parmi elles aiment la marque X, que pouvez vous dire de la probabilité p qu'une personne tirée au hasard dans la population aime la marque X ?

Une première façon de répondre à cette question serait de proposer, comme dans les paragraphes précédents, une estimation ponctuelle de cette probabilité p, donc pour cet exemple, $\widehat{p} = 0.42$.

Examinons cela de près : notons $X_1, ..., X_n$ les variables de Bernoulli $B(1, p)$ indépendantes modélisant la réponse éventuelle de chaque individu parmi les n tirés au hasard dans la population

L'intervalle de confiance au niveau 95% est :

sous R :

```
> 0.42 - qnorm(0.975)*sqrt(0.42*0.58)/sqrt(100)

[1] 0.3232643

> 0.42 + qnorm(0.975)*sqrt(0.42*0.58)/sqrt(100)
```

```
[1] 0.5167357
```

intervalle de confiance ne recouvre pas toujours la vraie valeur du paramètre p.

sous R :

```
m <- 50 ; n<-20 ; Pi <- .5 ; # 20 pièces à lancer 50 fois

p <- rbinom(m,n,Pi)/n # simuler et calculer l'estimation

ET<- sqrt(p*(1-p)/n) # estimer l'écart-type

alpha = 0.10 # seuil de signification

zstar <-qnorm(1-alpha/2)

matplot(rbind(p-zstar*ET, p+zstar*ET),rbind(1 :m,1 :m),type="l",lty=1)

abline(v=Pi) # trace la verticale en Pi
```

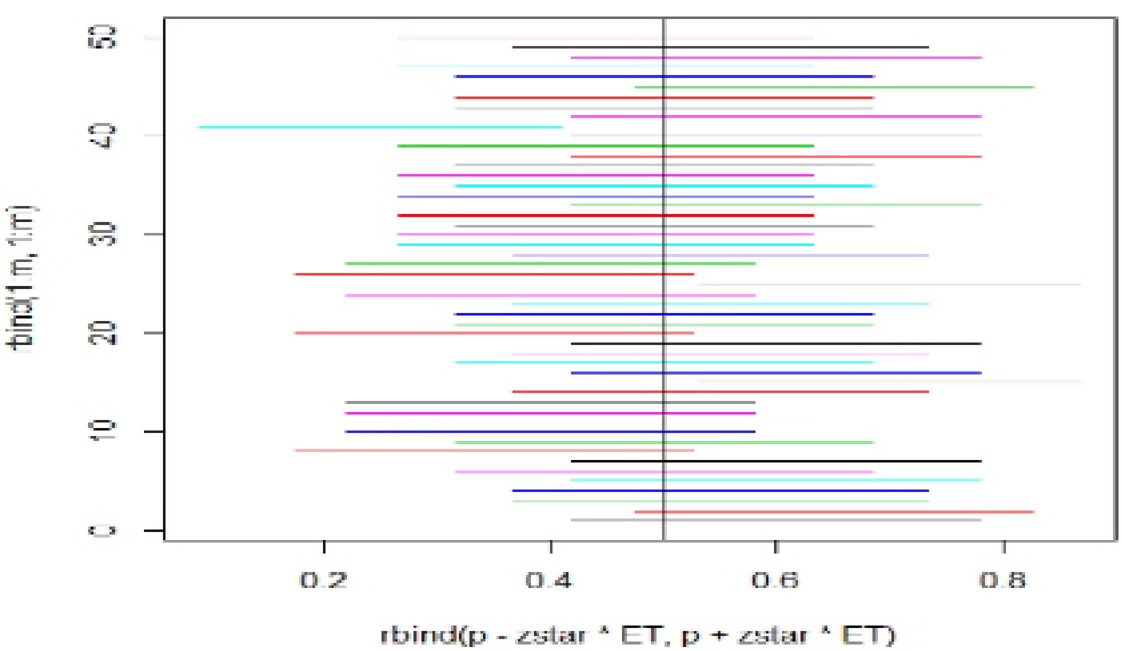

FIG. 3.2 – Combien d'intervalles de confiance de niveau 80 % recouvrent-ils la vraie proportion $p = 0.5$?.

Conclusion

En conclusion,il existe deux types d'estimation proposés :l'estimation ponctuelle et l'estimation par l'intervalle de confiance.

L'estimation ponctuelle d'un paramètre consiste à évaluer la valeur du paramètre de la population à l'aide d'une valeur unique prise dans un échantillion.

L'intervalle de confiance evaluer la confiance que l'on peut avoir en une estimation ,il est nécessaire de lui associer un intervalle qui contient, avec une certaine probabilité, la vraie valeur du paramètre.

Bibliographie

[1] Veysseyre, R. (2014). Aide-mémoire-Statistique et probabilités pour les ingénieurs. Dunod..

[2] JACQUES, J. Statistiques inférentielles..

[3] L Grammont.2003.Cours de statistique inférentielles.licence d'économie et de gestion.https ://dossier.univ-st-etienne.fr/fac-sciences-maths/www/stat03.pdf.

[4] I, Gannaz. Introduction à la statistique.http ://math.univ-lyon1.fr/~gannaz/Cours/cours_stat.pdf.

[5] P Dusart.2018.Cours de Statistiques inférentielles.https ://www.unilim.fr/pages_perso/pierre.dusart/Probas/cours_stat_S4.pdf.

[6] L Berdjoudj.S Bouraine.2013/2014.https ://cours-examens.org/images/An_2017_1/Etudes_superieures/statistiques/Alg%C3%A9rie/mi_lessons06_stat_inferentielle.pdf

[7] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip

[8] Velenik, Y. (2011). Probabilités et statistique. Université de Geneve.

[9] Lejeune, M. (2004). Statistique : La théorie et ses applications. Springer Science & Business Media..

[10] Hubler, J. (2007). Statistique descriptive appliquée à la gestion et à l'économie. Editions Bréal..

Annexe A : Logiciel *R*

3.10 Qu'est-ce-que le langage R ?

• Le langage **R** est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• **R** a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

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

$v.a$	:	Variable aléatoire.
$i.i.d$	:	Indépendantes identiqument distribuées.
$c.\grave{a}.d$ ou $i.e$	:	C'est à dire.
$E(X)$ ou μ	:	Espérance mathématique ou moyenne di v.a X.
$V(X)$	:	Variance du v.a X.
$\overline{X}$	:	Moyenne empirique.
IC	:	Intervalle de confaince.
σ	:	Ecart type.
$\xrightarrow{p}$	:	Converge en probabilité.
S^2	:	Variance empirique.
$L(x,\theta)$ou $L_\theta(x)$	:	Fonction de la vraismeblance.
$f(x,\theta)$	:	Densité de probabilité.
$\mathbb{P}_\theta$	:	Loi de paramètre θ.
$b_\theta(T)$	:	Biais de T.
$I_n(\theta)$	:	Inforamtion de fisher.

$\mu_r^{'}$	:	Moment d'ordre r.
EMV	:	Estimation du maximum de vraisemblance.
EMM	:	Estimation des moments.
$LG(0;1)$ ou $N(0;1)$	:	Loi normale centré réduite.
$t(n-1)$	:	Loi student de degré $n-1$.
$\chi^2(n)$	:	Loi de khi-deux de degré n.
$\mu_{1-\frac{\alpha}{2}}$	:	Fractile d'ordre $1-\frac{\alpha}{2}$.
$N(\mu,\sigma^2)$	:	Loi normale de moyenne μ et de variance σ^2.
α	:	Risque que l'intervalle de confiance.
$\xrightarrow{L}$	:	Converge en loi.