

**UNIVERSITE MOHAMED PREMIER**  
**FACULTE DES SCIENCES JURIDIQUES,**  
**ÉCONOMIQUES ET SOCIALES D'OUJDA**

**COURS ECHANTILLONNAGE ET ESTIMATION**

**CHAPITRE 1 : Estimation Ponctuelle et Estimation par Intervalle de Confiance**

**- Partie2 -**

**Deuxième année sciences économiques et gestion, semestre 3**

**Année universitaire**

**2020-2021**

# Sommaire

|                            |                                                                   |           |
|----------------------------|-------------------------------------------------------------------|-----------|
| <b>1.</b>                  | <b>DISTRIBUTION D'ECHANTILLONNAGE.....</b>                        | <b>3</b>  |
| 1.1.                       | INTRODUCTION .....                                                | 3         |
| 1.2.                       | ECHANTILLON ALEATOIRE .....                                       | 3         |
| 1.3.                       | DISTRIBUTION DANS LE CAS D'UN ECHANTILLON DE LOI NORMALE .....    | 6         |
| 1.4.                       | DISTRIBUTION DANS LE CAS D'UN ECHANTILLON DE LOI QUELCONQUE ..... | 7         |
| <b>2.</b>                  | <b>ESTIMATION PONCTUELLE.....</b>                                 | <b>7</b>  |
| 2.1.                       | DEFINITIONS ET PROPRIETES .....                                   | 7         |
| 2.2.                       | ESTIMATEUR PAR LA METHODE DES MOMENTS .....                       | 11        |
| 2.3.                       | EFFICACITE D'UN ESTIMATEUR.....                                   | 14        |
| <b>3.</b>                  | <b>ESTIMATION PAR INTERVALLE DE CONFIANCE .....</b>               | <b>17</b> |
| 3.1.                       | INTRODUCTION .....                                                | 17        |
| 3.2.                       | INTERVALLE DE CONFIANCE D'UNE MOYENNE .....                       | 17        |
| 3.2.1.                     | Cas où la loi de l'échantillon est inconnue.....                  | 18        |
| 3.2.2.                     | Cas où la loi de l'échantillon est normale .....                  | 19        |
| 3.3.                       | INTERVALLE DE CONFIANCE D'UNE VARIANCE.....                       | 19        |
| 3.3.1.                     | Cas où la loi de l'échantillon est inconnue .....                 | 20        |
| 3.3.2.                     | Cas où la loi de l'échantillon est normale.....                   | 20        |
| 3.4.                       | INTERVALLE DE CONFIANCE D'UNE PROPORTION .....                    | 22        |
| 3.5.                       | MARGE D'ERREUR ET TAILLE MINIMALE.....                            | 23        |
| <b>ANNEXE</b> .....        |                                                                   | <b>27</b> |
| <b>BIBLIOGRAPHIE</b> ..... |                                                                   | <b>32</b> |

## 1. Distribution d'échantillonnage

### 1.1. Introduction

Si on choisi un échantillon plutôt qu'un autre, on n'aura pas forcément la même moyenne. Considérons toutes les valeurs prises par les moyennes des échantillons possibles de même taille, on aura nécessairement des fluctuations. On veut étudier la distribution d'échantillonnage de la moyenne pour caractériser ces fluctuations.

Un paramètre est une caractéristique de la population ou de la distribution de probabilité de la variable étudiée. On peut parler alors, de moyenne ou de variance de la population (plus généralement de mesures de tendance centrale ou de dispersion de la population), mais aussi, un paramètre peut être une caractéristique de la distribution du caractère étudié comme l'espérance mathématique ou la variance de la variable aléatoire.

Si on suppose que la moyenne de la population est un paramètre inconnu, on voudrait l'approcher par la moyenne de l'échantillon. Si on a la possibilité d'obtenir tous les échantillons et de calculer leurs moyennes, on aura des moyennes d'échantillons supérieurs ou inférieurs à la moyenne de la population.

La moyenne calculée sur un échantillon se comporte comme une variable aléatoire, la connaissance de sa distribution va nous permettre de mesurer l'écart aléatoire (erreur d'échantillonnage) entre la moyenne de l'échantillon et la moyenne de la population.

### 1.2. Echantillon aléatoire

Soit  $X$  une variable aléatoire définie sur un espace de probabilité  $(\Omega, \mathcal{A}, P_\theta)$

$$X : w \in \Omega \rightarrow X(w) \in E$$

$E$  sera souvent l'espace des réels  $\mathbb{R}$ .

Si la loi  $P_\theta$  est connue, c.à.d. qu'elle correspond à une fonction explicite, prenons comme exemple une des lois de probabilités usuelles (loi Bernoulli, Poisson, normale,...) ; elle dépendra d'un paramètre inconnu  $\theta$  :

| Loi $P_\theta$   | Paramètre $\theta$                                                          |
|------------------|-----------------------------------------------------------------------------|
| $B(p)$           | $\theta = p ; \quad \theta \in [0, 1]$                                      |
| $P(\lambda)$     | $\theta = \lambda ; \quad \theta \in \mathbb{R}^{+*}$                       |
| $N(m, \sigma)$   | $\theta = (m, \sigma) ; \quad \theta \in \mathbb{R} \times \mathbb{R}^{+*}$ |
| $N(m, 2)$        | $\theta = m ; \quad \theta \in \mathbb{R}$                                  |
| $N(100, \sigma)$ | $\theta = \sigma ; \quad \theta \in \mathbb{R}^{+*}$                        |

Le paramètre  $\theta$  appartient à un ensemble  $\Theta$  dans  $\mathbb{R}$  ( ou généralement  $\mathbb{R}^d, d \in \mathbb{N}^*$  ) ; l'ensemble  $\Theta$  est confondu le plus souvent à un intervalle de  $\mathbb{R}$  ( ou  $\mathbb{R}^d$  ).

↳ Le problème de l'estimation est de chercher le « *bon* »  $\theta$  pour connaître la « *bonne loi* »  $P_\theta$ .

**Exemple :** Soit  $\theta = p$  la proportion des foyers qui ont une connexion internet chez eux.

On sélectionne au hasard un foyer et on vérifie s'il a une connexion internet.

Soit la v.a.  $X$  définie par : 
$$X = \begin{cases} 1 & \text{si oui} \\ 0 & \text{si non} \end{cases}$$

$X \sim B(p), \quad P_\theta(X = 1) = p \quad \text{et} \quad P_\theta(X = 0) = 1 - p, \quad p \in [0, 1]$

Le but de la théorie de l'estimation est de construire par un procédé ou une méthode d'estimation, une approximation du paramètre  $\theta$ , cette construction se basera sur l'observation de v.a.  $X_1, \dots, X_n$  de loi, celle de  $X$  la v.a. étudiée.

**Définition1 :** On appelle  $n$ -échantillon ou échantillon de loi  $P_\theta$ , une suite  $X_1, X_2, \dots, X_n$  de  $n$  v.a. à valeurs réelles *indépendantes* et de *même loi*  $P_\theta$ .

**Exemple :** On sélectionne  $n$  foyers, soit  $X_i$  la v.a. qui prend la valeur 1 si le  $i^{\text{ème}}$  foyer a une connexion internet à domicile et la valeur 0 sinon. On obtient un échantillon de taille  $n$  :  $X_1, \dots, X_n$  ; chaque  $X_i \sim B(p)$ .

**Remarque :** Ici la sélection de l'échantillon se fait avec un tirage aléatoire simple, on note **ES** pour **Echantillon Simple**. On distingue le tirage avec remise et le tirage sans remise :

ESAR : Echantillon Simple Avec Remise  $\rightarrow$  tirage indépendant

ESSR : Echantillon Simple Sans Remise  $\rightarrow$  tirage dépendant

**Définition2 :** On dit que  $(x_1, x_2, \dots, x_n)$  est une réalisation (ou observation) de l'échantillon  $(X_1, X_2, \dots, X_n)$  ; si  $(x_1, x_2, \dots, x_n) = (X_1(w), X_2(w), \dots, X_n(w))$  pour un  $w \in \Omega$ .

**Remarque :**  $(x_1, x_2, \dots, x_n)$  est considéré comme un point de  $\mathbb{R}^n$ , c'est une **réalisation possible** de l'échantillon (on parle aussi d'observation). C'est après la sélection des  $n$  observations  $X_i(w) = x_i$  qu'on parle de valeurs numériques non aléatoires (déterministes). En général on réserve les lettres en **MAJUSCULE** pour les variables aléatoires et les lettres en **minuscule** pour les observations ou réalisations.

**Définition3 :** On appelle moyenne empirique d'un échantillon  $X_1, \dots, X_n$  la v.a. définie par

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

**Remarque :** La moyenne empirique est aussi appelée moyenne échantillonnale, on notera  $\bar{X}_n$  ou  $\bar{X}$ .

**Propriétés :** Si  $E(X_1) = \mu$  et  $Var(X_1) = \sigma^2$ , alors on a

$$E(\bar{X}_n) = \mu \quad \text{et} \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

**Preuve :** Voir Annexe

### Cas d'une population finie

Quand l'échantillonnage s'effectue avec un tirage sans remise (ESSR) à partir d'une population finie de taille  $N$  (le tirage est dit exhaustive), on corrige  $Var(\bar{X}_n)$ , c.à.d. qu'on multiplie par le « coefficient d'exhaustivité » :

$$\frac{N-n}{N-1}$$

$$Var(\bar{X}_n) = \frac{\sigma^2}{n} \frac{N-n}{N-1} \cong \frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)$$

$\frac{n}{N}$  est le taux de sondage, il est négligeable si  $N \rightarrow +\infty$ , ou si il est très petit.

**Remarque :** On va considérer que le facteur de correction peut être ignoré si le taux de sondage est inférieur à 5%. Par exemple, si  $N = 1000$  et  $n = 30$ , alors

$$\frac{N-n}{N-1} = \frac{970}{999} = 0,97 \approx 1$$

$$\left(\frac{n}{N} = \frac{30}{1000} = 0.03 < 0.05\right)$$

**Définition :** On appelle variance empirique d'un échantillon  $X_1, \dots, X_n$  la v.a. définie par

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

**Propriétés :** Si  $E(X_1) = \mu$  et  $Var(X_1) = \sigma^2$ , alors on a

$$E(S_n^2) = \sigma^2 \quad \text{et} \quad Var(S_n^2) = \frac{n}{(n-1)^2} \left( \mu_4 - \frac{n-1}{n-3} \sigma^4 \right)$$

Avec  $\mu_4 = E(X^4)$ .

**Preuve :** Voir Annexe pour le premier résultat

**Remarque :** On peut utiliser de même d'autres variances empiriques :

- $\hat{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ , dans ce cas là on vérifie que

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

- $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ , ici on a  $E(S_n^2) = \sigma^2$

Maintenant on passe à la distribution (loi) de probabilité des deux caractéristiques empiriques de l'échantillon :  $\mathcal{L}(\bar{X}_n)$  ?  $\mathcal{L}(S_n^2)$  ?

### 1.3. Distribution dans le cas d'un échantillon de loi normale

On suppose qu'une v.a.  $X$  est de loi normale  $N(\mu, \sigma)$  ; si on échantillonne, on obtient  $n$  v.a.  $X_1, \dots, X_n$  i.i.d. de loi  $N(\mu, \sigma)$ . On admet les résultats suivants :

1. La loi exacte de

$$\sqrt{n} \left( \frac{\bar{X}_n - \mu}{\sigma} \right)$$

est une loi normale  $N(0,1)$

2. Les deux v.a.  $\bar{X}_n$  et  $S_n^2$  sont indépendantes, ce qui permet d'obtenir le résultat suivant

3. Si la variance est inconnue, on remplace  $\sigma$  par  $S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$

La loi exacte de

$$\sqrt{n} \left( \frac{\bar{X}_n - \mu}{S_n} \right)$$

est une loi de Student à  $n - 1$  degrés de liberté  $t(n - 1)$

4. La loi exacte de

$$\left( \frac{n-1}{\sigma^2} \right) S_n^2$$

est une loi de khideux à  $n - 1$  degrés de liberté  $\chi^2(n - 1)$

5. Si la moyenne  $\mu$  est connue, on considère la variance empirique suivante :

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

La loi exacte de

$$\frac{n}{\sigma^2} S_n^2$$

est une loi de khideux à  $n$  degrés de liberté  $\chi^2(n)$

## 1.4. Distribution dans le cas d'un échantillon de loi quelconque

Dans cette partie on ne considère que des échantillons de grande taille, on utilise principalement le théorème centrale limite.

Soit l'échantillon  $X_1, \dots, X_n$  de loi inconnue de moyenne  $\mu$  et de variance  $\sigma^2$ . On admet les résultats suivants :

### 1. TCL

$$\sqrt{n} \left( \frac{\bar{X} - \mu}{\sigma} \right) \sim N(0,1) \text{ pour } n \text{ assez grand}$$

### 2. Si la variance est inconnue, on aura

$$\sqrt{n} \left( \frac{\bar{X} - \mu}{S_n} \right) \sim N(0,1) \text{ pour } n \text{ assez grand}$$

### 3. On vérifie la convergence en probabilité :

$$\begin{aligned} S_n^2 &\xrightarrow{P} \sigma^2 \\ \bar{X} &\xrightarrow{P} \mu \end{aligned}$$

### 4. La variance de $S_n^2$ est égale à

$$\text{Var}(S_n^2) = \frac{n}{(n-1)^2} \left( \mu_4 - \frac{n-1}{n-3} \sigma^4 \right) \approx \frac{1}{n} (\mu_4 - \sigma^4)$$

Et on a le résultat suivant :

$$\sqrt{n} \frac{S_n^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \sim N(0,1) \text{ pour } n \text{ assez grand}$$

## 2. Estimation ponctuelle

### 2.1. Définitions et propriétés

**Définition1** : on appelle **statistique** sur  $(X_1, \dots, X_n)$  toute v.a.  $T$  qui est obtenue par une fonction des v.a.  $X_1, \dots, X_n$  :  $T = f(X_1, \dots, X_n)$

On distingue deux types de statistiques, celles qui ont le but d'estimer un paramètre inconnu et celles qui nous permettent de construire un test d'hypothèse.

**Définition2** : On appelle **estimateur**  $\hat{\theta}_n$  de  $\theta$ , toute statistique qui estime  $\theta$ .

Ainsi l'obtention de l'estimateur  $\hat{\theta}_n$  de  $\theta$  est liée à une fonction des observations  $X_1, \dots, X_n$  ;

$$\hat{\theta}_n = f(X_1, \dots, X_n)$$

**Définition3** : Soit un estimateur  $T_n$  de  $\theta$ . On appelle **estimation** de  $\theta$  la valeur observée de  $T_n$ , qu'on peut noter  $t_n$  ou  $t_{n,obs}$ .

$$t_n = T_n(w) = f(X_1(w), \dots, X_n(w)) = f(x_1, \dots, x_n)$$

**Exemple :** on peut proposer des estimateurs de la moyenne  $\mu$  par différentes formules sans se préoccuper de la manière par laquelle ils sont obtenus !

| Estimateur                                    | Estimation                                    |
|-----------------------------------------------|-----------------------------------------------|
| $\frac{1}{n} \sum_{i=1}^n X_i$                | $\frac{1}{n} \sum_{i=1}^n x_i$                |
| $\frac{\text{Max}(X_i) - \text{Min}(X_i)}{2}$ | $\frac{\text{Max}(x_i) - \text{Min}(x_i)}{2}$ |
| $\frac{2}{n(n+1)} \sum_{i=1}^n i X_i$         | $\frac{2}{n(n+1)} \sum_{i=1}^n i x_i$         |

D'où la nécessité de critères de sélection d'un estimateur parmi d'autres.

**Définition4 :** Soit un estimateur  $\hat{\theta}_n$  de  $\theta$ . On appelle **biais** de  $\hat{\theta}_n$ , le nombre réel

$$B(\hat{\theta}_n) = E[\hat{\theta}_n - \theta] = E(\hat{\theta}_n) - \theta$$

- On dit que  $\hat{\theta}_n$  est un **estimateur sans biais** (e.s.b.) de  $\theta$  si  $B(\hat{\theta}_n) = 0$  c.à.d. que  $E(\hat{\theta}_n) = \theta$
- On dit que  $\hat{\theta}_n$  est un estimateur asymptotiquement sans biais de  $\theta$  si  $B(\hat{\theta}_n) \rightarrow 0$  quand  $n \rightarrow +\infty$  ce qui veut dire que  $E(\hat{\theta}_n) \rightarrow \theta$  quand  $n \rightarrow +\infty$

**Définition5 :** On appelle **risque quadratique** ou **erreur quadratique moyenne** de  $\hat{\theta}_n$ , le nombre réel positif

$$MSE(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2]$$

*MSE = Mean Square Error*

**Propriété :** Soit un estimateur  $\hat{\theta}_n$  de  $\theta$ . Alors on a

$$MSE(\hat{\theta}_n) = Var(\hat{\theta}_n) + [B(\hat{\theta}_n)]^2$$

**Remarque :** Si  $\hat{\theta}_n$  appartient à la famille des e.s.b. de  $\theta$ , alors son risque quadratique coïncide avec sa variance :  $MSE(\hat{\theta}_n) = Var(\hat{\theta}_n)$

**Définition6 :** on dit que  $\hat{\theta}_n$  est un estimateur **consistant** (ou **convergeant**) de  $\theta$  si  $\hat{\theta}_n$  converge en probabilité vers  $\theta$ .

$$\hat{\theta}_n \xrightarrow{P} \theta$$

Pour tout  $\varepsilon > 0$

$$\lim_{n \rightarrow +\infty} P(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0$$

**Remarque :** Une condition suffisante pour que  $\hat{\theta}_n$  soit un estimateur consistant est que son risque quadratique converge vers 0. En effet, d'après l'inégalité de Markov on a :

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) \leq \frac{E[(\hat{\theta}_n - \theta)^2]}{\varepsilon^2} = \frac{MSE(\hat{\theta}_n)}{\varepsilon^2}$$

Si  $MSE(\hat{\theta}_n) \rightarrow 0$  quand  $n \rightarrow +\infty$  alors de même

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) \rightarrow 0$$

### Application1 : estimation de la moyenne d'une population

Soit  $X$  une v.a. de moyenne  $\mu$  et de variance  $\sigma^2$ ; le plus naturel pour estimer  $\mu$  est de considérer un échantillon de  $X$  et de calculer sa moyenne empirique. ( $\theta = \mu$  et  $\hat{\theta}_n = \bar{X}$ )

**Propriétés :** La moyenne empirique  $\bar{X}$  est un estimateur de  $\mu$  qui a les propriétés suivantes,

- C'est un e.s.b. de  $\mu$  :  $E(\bar{X}) = \mu$
- Son risque quadratique converge vers 0 quand  $n$  augmente :

$$MSE(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0 \text{ quand } n \rightarrow \infty$$

- C'est un estimateur consistant de  $\mu$

(Loi faible des grands nombre ou d'après la remarque précédente)

- Sa loi asymptotique correspond à la loi normale  $N(\mu, \frac{\sigma}{\sqrt{n}})$

### Application2 : estimation de la variance d'une population

Un estimateur de  $\sigma^2$  est la variance empirique  $S_n^2$ , on donne ses propriétés.

**Propriétés :** La variance empirique a les propriétés suivantes,

- C'est un estimateur sans biais :  $E(S_n^2) = \sigma^2$
- Son risque quadratique converge vers 0 :

$$MSE(S_n^2) = \frac{n}{(n-1)^2} \left( \mu_4 - \frac{n-1}{n-3} \sigma^4 \right) \rightarrow 0$$

- Donc c'est un estimateur consistant de  $\sigma^2$

### Application3 : estimation de la proportion d'une population

C'est le cas d'une population où l'on peut observer un événement particulier  $A$  :

$$X(w) = \begin{cases} 1 & \text{si } A \text{ est observé} \\ 0 & \text{si } A \text{ n'est pas observé} \end{cases}$$

La v.a.  $X$  suit une loi Bernoulli de paramètre  $p$ , avec  $p$  la proportion d'individus de la population avec qui on a observé l'événement  $A$ .

La proportion  $p$  correspond à la moyenne de  $X$  :

$$p = E(X)$$

Tandis que la variance de  $X$  est

$$Var(X) = p(1 - p)$$

Soit  $X_1, \dots, X_n$  un échantillon de  $X$ . Un estimateur de  $p$  est la moyenne empirique qu'on notera :  $\bar{p}_n$ , elle correspond à une fréquence ou pourcentage. On donne les propriétés de cet estimateur.

**Propriétés :** L'estimateur  $\bar{p}_n$  de  $p$  a les propriétés suivantes,

- C'est un estimateur sans biais :  $E(\bar{p}_n) = p$ .
- Son risque quadratique converge vers 0 :

$$MSE(\bar{p}_n) = \frac{p(1-p)}{n} \xrightarrow{n \rightarrow \infty} 0$$

- C'est un estimateur convergent de  $p$  :  $\bar{p}_n \xrightarrow{P} p$
- C'est un estimateur qui est asymptotiquement de loi normale (TCL) :

$$\sqrt{n} \left( \frac{\bar{p}_n - p}{\sqrt{p(1-p)}} \right) \xrightarrow{\mathcal{L}} N(0,1)$$

Qu'on peut écrire sous la forme

$$\mathcal{L}(\bar{p}_n) \approx N \left( p, \sqrt{\frac{p(1-p)}{n}} \right)$$

**Remarque :** on utilise le terme « **erreur-type** » pour désigner l'écart-type d'un estimateur. On note ainsi  $\sigma_{\bar{X}}$  l'écart-type de l'estimateur  $\bar{X}$  qui dépend de l'écart-type de la population :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

De même on retrouve l'erreur-type de la proportion  $\bar{p}_n$  :

$$\sigma_{\bar{p}_n} = \sqrt{\frac{p(1-p)}{n}}$$

Si on veut citer les méthodes d'estimations, on citera deux qui sont très connues, la méthode des moments et la méthode du maximum de vraisemblance. Dans ce chapitre on se restreindra à la méthode des moments.

## 2.2. Estimateur par la méthode des moments

Soit  $(X_1, \dots, X_n)$  un échantillon d'une variable  $X$ . Le principe de cette méthode est de considérer que toute transformation de l'échantillon de la forme  $(f(X_1), \dots, f(X_n))$  avec  $f$  une fonction continue sur  $X(\Omega)$ , aboutit à un échantillon de la variable  $f(X)$ .

En effet, l'indépendance est conservée entre les  $(f(X_i))$  et on a la même loi, celle calculée sur la transformée  $f(X)$ . La fonction  $f$  sera le plus souvent une fonction puissance  $f: x \mapsto x^k$ .

### Proposition :

Soit  $(X_1, \dots, X_n)$  un échantillon de  $X$ ,  $k \in \mathbb{N}^*$ , avec  $E(X^k) < +\infty$ . Alors

$$\overline{X_n^k} = \frac{1}{n} \sum_{i=1}^n X_i^k$$

est un estimateur sans biais et consistant (convergent) de  $E(X^k) = m_k$  (moment d'ordre  $k$ ).

**Preuve :** Il est clair qu'il est s.b. Pour la convergence en probabilité on peut utiliser la loi faible des grands nombres.

**Définition :** Soit  $g$  une fonction continue de  $\mathbb{R}$  vers  $\mathbb{R}$ ,  $k \in \mathbb{N}^*$ . On peut estimer  $g(m_k)$  par

$$g(\overline{X_n^k}) = g\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right)$$

Plus généralement, si  $g$  est continue de  $\mathbb{R}^p$  vers  $\mathbb{R}$ ,  $p \in \mathbb{N}^*$ , on peut estimer  $g(m_1, m_2, \dots, m_p)$  par

$$g(\overline{X_n}, \overline{X_n^2}, \dots, \overline{X_n^p})$$

On dit que  $g(\overline{X_n^k})$  (resp.  $g(\overline{X_n}, \overline{X_n^2}, \dots, \overline{X_n^p})$ ) est un estimateur de  $g(m_k)$  (resp.  $g(m_1, m_2, \dots, m_p)$ ) par la méthode des moments (EMM).

**Remarque :** Puisque la fonction  $g$  est continue, tous les estimateurs par la méthode des moments sont convergents, cependant on peut facilement perdre la propriété -sans biais-.

## Applications sur des lois usuelles

### a) Loi Bernoulli $\mathcal{B}(p)$

On cherche à estimer la proportion  $p$  et en même temps, on a besoin de déterminer l'estimateur de la variance.

On connaît que  $p = E(X)$  et  $Var(X) = p(1 - p) = E(X)(1 - E(X))$ , ce qui donne les estimateurs par la méthode des moments :

- Pour la proportion  $p$ , on a comme estimateur la moyenne empirique  $\bar{X}_n$  qu'on notera ici par  $\bar{p}_n$
- Pour la variance, on a comme estimateur  $\bar{p}_n(1 - \bar{p}_n)$ , ce deuxième estimateur est biaisé, néanmoins il sera utilisé dans l'estimation par intervalle de confiance de la proportion  $p$ .

### b) Loi de Poisson $\mathcal{P}(\lambda)$

On connaît que  $E(X) = Var(X) = \lambda$ .

- $\lambda = E(X)$  ce qui donne l'estimateur par la méthode des moments du paramètre  $\lambda$  :

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

De même, on a :

- $\lambda = Var(X) = E(X^2) - (E(X))^2$  ce qui donne un deuxième estimateur par la méthode des moments du paramètre  $\lambda$  :

$$\overline{X_n^2} - (\bar{X}_n)^2$$

Les deux estimateurs sont consistants, le premier est sans biais, le deuxième est biaisé. Cependant pour choisir entre les deux estimateurs, la propriété s.b. ne tranche pas, il faudra aller jusqu'au calcul des risques quadratiques des deux estimateurs, en fait le premier a le risque quadratique minimum, et il est préférable au deuxième.

### c) Loi exponentielle $\mathcal{E}(\mu)$

On a :

$$\begin{aligned} E(X) &= \frac{1}{\mu} \Rightarrow \mu = \frac{1}{E(X)} \\ Var(X) &= \frac{1}{\mu^2} \Rightarrow \mu = \frac{1}{\sqrt{Var(X)}} \end{aligned}$$

- Un premier estimateur par la méthode des moments du paramètre  $\mu$  sera :

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

- Un deuxième estimateur de  $\mu$  est donné par la formule :

$$\frac{1}{\sqrt{\bar{X}_n^2 - (\bar{X}_n)^2}}$$

Les deux estimateurs sont consistants, mais les deux sont biaisés. Pour comparer ces estimateurs on doit calculer le risque quadratique de chacun.

#### d) Loi Normale $N(\mu, \sigma^2)$

L'espérance et la variance de cette loi sont les paramètres  $\mu$  et  $\sigma^2$ .

$\mu = E(X)$  et  $\sigma^2 = Var(X)$ , ce qui donne les deux estimateurs moyenne empirique  $\bar{X}_n$  et variance empirique non corrigée

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

comme estimateurs par la méthode des moments.

#### Exercice d'application :

Soit  $X$  une v.a. de densité :

$$f_{\theta}(x) = \begin{cases} \frac{2x}{\theta^2} & \text{si } 0 \leq x \leq \theta \\ 0 & \text{sinon} \end{cases}$$

Où  $\theta$  est un paramètre positif inconnu

1. Déterminer un estimateur  $T_n$  de  $\theta$  par la méthode des moments.
2. Est-ce que  $T_n$  est un estimateur s.b., consistant de  $\theta$ .
3. Calculer  $Var(X)$  et déduire  $Var(T_n)$ .

On considère maintenant  $M_n$  une autre v.a. qui dépend de  $(X_1, \dots, X_n)$  où les  $X_i$  sont des v.a. indépendantes et de même loi que  $X$ .

On suppose que la fonction de densité de  $M_n$  est

$$g_{\theta}(x) = \begin{cases} 2n \frac{x^{2n-1}}{\theta^{2n}} & \text{si } 0 \leq x \leq \theta \\ 0 & \text{sinon} \end{cases}$$

4. Calculer  $E(M_n)$  et  $Var(M_n)$ .
5. D duire un autre estimateur  $T'_n$  s.b. de  $\theta$  qui est fonction de  $M_n$ .
6. Calculer  $Var(T'_n)$  et la comparer avec  $Var(T_n)$ . Que peut-on conclure ?

**Solution :** Voir Annexe

**Remarque :** La reproduction des estimateurs par la m thode des moments repose sur le r sultat de la loi faible des grands nombres. Et si la taille de l' chantillon n'est pas grande, on risque d'avoir des estimateurs qui ont tendance   s' loigner du vrai param tre.

La m thode fournit des estimations faciles   calculer, elles sont souvent utilis es comme point de d part pour une am lioration de l'estimation, en utilisant d'autres m thodes comme la m thode du maximum de vraisemblance, qui donne de meilleurs r sultats.

### 2.3. Efficacit  d'un estimateur

**D finition1 :** un estimateur  $T_{n,1}$  de  $\theta$  est optimal si quelque soit  $T_{n,2}$  estimateur de  $\theta$  on a :

$$Var(T_{n,1}) \leq Var(T_{n,2})$$

On dira que  $T_{n,1}$  est   variance minimale.

**D finition2 :** soit deux estimateurs  $T_{n,1}$  et  $T_{n,2}$  de  $\theta$ , on dira que l'estimateur  $T_{n,1}$  est plus efficace que l'estimateur  $T_{n,2}$  si

$$MSE(T_{n,1}) < MSE(T_{n,2})$$

Si on calcule le  $MSE$  d'un estimateur  $T_n$  de  $\theta$  on aura :

$$MSE(T_n) = Var(T_n) + [E(T_n) - \theta]^2$$

Si on veut am liorer le  $MSE(T_n)$ , on a deux voies :

- Chercher dans la famille des e.s.b. l'estimateur optimal, celui qui a la variance minimale.
- Chercher parmi les estimateurs biais s une am lioration du  $MSE(T_n)$  : une l g re modification au niveau du biais  $E(T_n) - \theta$  peut am liorer globalement le  $MSE(T_n)$  si la variance diminue significativement

### Cas de la famille des estimateurs sans biais :

$$MSE(T_n) = Var(T_n)$$

On peut montrer le résultat suivant :

Si  $T_n$  est un e.s.b. de  $\theta$  alors

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

La borne inférieure

$$B_{CR} = \frac{1}{I_n(\theta)}$$

est dite borne de *Cramér-Rao* ou borne de *Fréchet-Darmois-Cramér-Rao* notée  $B_{FDCR}$

La quantité positive  $I_n(\theta)$  est dite information de *Fisher*, elle est par définition :

$$I_n(\theta) = E \left( \left[ \frac{L'_\theta(X_1, \dots, X_n)}{L_\theta(X_1, \dots, X_n)} \right]^2 \right)$$

$L_\theta(X_1, \dots, X_n)$  est la vraisemblance associée à  $\theta$  calculée sur l'échantillon :

- Si  $X$  est une v.a. discrète, alors on pose :

$$L_\theta(x_1, \dots, x_n) = \prod_{i=1}^n P_\theta(X_i = x_i) = P_\theta(X_1 = x_1) P_\theta(X_2 = x_2) \dots P_\theta(X_n = x_n)$$

- Si  $X$  est une v.a. continue de densité  $f_\theta$ , alors on pose :

$$L_\theta(x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i) = f_\theta(x_1) f_\theta(x_2) \dots f_\theta(x_n)$$

La fonction  $L_\theta(X_1, \dots, X_n)$  est la vraisemblance de l'échantillon  $X_1, \dots, X_n$ , elle est obtenue en remplaçant les valeurs observées  $x_1, \dots, x_n$  par les v.a.  $X_1, \dots, X_n$ . La vraisemblance est une variable aléatoire.

Si on utilise le log-vraisemblance  $l_\theta(X_1, \dots, X_n) = \log[L_\theta(X_1, \dots, X_n)]$ , on aura :

$$I_n(\theta) = E \left( \left[ \frac{\partial}{\partial \theta} l_\theta(X_1, \dots, X_n) \right]^2 \right)$$

Ce qui donne ensuite l'expression qu'on utilisera dans l'exemple qui va suivre :

$$I_n(\theta) = -E \left( \frac{\partial^2}{\partial \theta^2} l_\theta(X_1, \dots, X_n) \right)$$

### Définition d'un estimateur efficace :

Si  $T_n$  est un e.s.b. de  $\theta$ , il est dit efficace si

$$Var(T_n) = \frac{1}{I_n(\theta)}$$

**Remarque :** Si un e.s.b. de  $\theta$  est efficace, il est optimal dans la famille des estimateurs s.b.

**Exemple :** Soit  $X \sim B(p)$ , on veut estimer  $p$  par l'estimateur  $\bar{p}_n$

La v.a.  $X \in E = \{0,1\}$ , si  $p \neq 0$  et  $p \neq 1$  alors on peut écrire pour  $x \in E$

$$P(X = x) = p^x (1 - p)^{1-x}$$

La vraisemblance  $L_p$  est une application de  $E^n$  vers  $[0, 1]$  avec

$$L_p(x_1, \dots, x_n) = \prod_{i=1}^n p^{x_i} (1 - p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i}$$

On aura enfin la v.a. vraisemblance

$$L_p(X_1, \dots, X_n) = p^{\sum_{i=1}^n X_i} (1 - p)^{n - \sum_{i=1}^n X_i} = p^{n\bar{X}} (1 - p)^{n(1-\bar{X})}$$

Le log-vraisemblance sera

$$l_p(X_1, \dots, X_n) = n\bar{X} \log(p) + n(1 - \bar{X}) \log(1 - p)$$

On aura

$$\begin{aligned} \frac{\partial}{\partial p} l_p(X_1, \dots, X_n) &= \frac{n\bar{X}}{p} - \frac{n(1 - \bar{X})}{1 - p} \\ \frac{\partial^2}{\partial p^2} l_p(X_1, \dots, X_n) &= -\frac{n\bar{X}}{p^2} - \frac{n(1 - \bar{X})}{(1 - p)^2} \end{aligned}$$

L'information de Fisher sera

$$I_n(p) = -E\left(\frac{\partial^2}{\partial p^2} l_p(X_1, \dots, X_n)\right) = \frac{n}{p^2} E(\bar{X}) + \frac{n}{(1 - p)^2} (1 - E(\bar{X})) = \frac{n}{p^2} p + \frac{n}{(1 - p)^2} (1 - p)$$

$$I_n(p) = \frac{n}{p} + \frac{n}{1 - p} = \frac{n}{p(1 - p)}$$

Si on calcule la variance de l'estimateur, elle coïncide avec la borne de Cramér-Rao

$$Var(\bar{p}_n) = \frac{p(1 - p)}{n} = \frac{1}{I_n(\theta)}$$

On conclut que  $\bar{p}_n$  est un estimateur efficace.

### 3. Estimation par intervalle de confiance

#### 3.1. Introduction

Soit  $\hat{\theta}_n$  un estimateur du paramètre  $\theta$ , supposons que  $\hat{\theta}_n$  est convergent en probabilité :

$$\hat{\theta}_n \xrightarrow{P} \theta ,$$

càd

$$\forall a > 0, \quad P(|\hat{\theta}_n - \theta| \leq a) \xrightarrow{n \rightarrow +\infty} 1$$

$$\Leftrightarrow$$

$$\forall a > 0, \quad P(\hat{\theta}_n - a \leq \theta \leq \hat{\theta}_n + a) \xrightarrow{n \rightarrow +\infty} 1 .$$

**Question :** peut-on déterminer un intervalle  $I$  autour de  $\hat{\theta}_n$  pour lequel on garantie que la probabilité  $P(\theta \in I)$  soit connue ?

Si on pose  $\alpha = P(\theta \notin I)$ ,  $\alpha$  sera le niveau de signification.

On appelle  $1 - \alpha$  **le niveau de confiance** (par exemple pour  $\alpha = 5\%$ ,  $1 - \alpha = 95\%$ ) et on note  $I(\theta)$  l'intervalle de confiance (on le note aussi  $I_{1-\alpha}(\theta)$  ),

$$I(\theta) = [\hat{\theta}_n - a; \hat{\theta}_n + a] .$$

↪ Le niveau de confiance est associé au degré de crédibilité qu'on donne à l'intervalle  $I(\theta)$  de contenir le paramètre inconnu  $\theta$ .

Pour un niveau de confiance  $1 - \alpha$  connue (ou fixé au préalable), on doit chercher  $a$  tel que

$$P(\theta \in I(\theta)) = P(|\hat{\theta}_n - \theta| \leq a) = 1 - \alpha .$$

Cela ne serait possible que si on connaît la loi de  $\hat{\theta}_n$ .

#### 3.2. Intervalle de confiance d'une moyenne

Soit  $(X_1, \dots, X_n)$  un échantillon de la v.a.  $X$  de moyenne  $\mu$  et de variance  $\sigma^2$ . On cherchera un intervalle de confiance autour de l'estimateur  $\bar{X}$  de  $\mu$ .

La loi de l'estimateur  $\bar{X}$  dépend de celle de la v.a.  $X$ . Si la loi de la variable  $X$  est inconnue et c'est le cas le plus courant, il est difficile de donner une réponse si la taille de l'échantillon est petite.

Dans le paragraphe qui suit, on va supposer que la taille de l'échantillon est grande, cela va nous permettre d'utiliser une approximation par le théorème centrale limite (T.C.L) de la loi de  $\bar{X}$  par celle de la loi normale.

### 3.2.1. Cas où la loi de l'échantillon est inconnue

On peut distinguer deux cas, le cas où  $\sigma^2$  est connue et celui où elle est inconnue.

**$\sigma^2$  connue :**

D'après le T.C.L. on a :

$$\sqrt{n} \left( \frac{\bar{X} - \mu}{\sigma} \right) \xrightarrow{\mathcal{L}} Z$$

avec  $Z$  suit la loi  $N(0,1)$ .

La loi  $N(0,1)$  est symétrique, on choisira  $a > 0$  tel que  $P(|Z| \leq a) = 1 - \alpha \Leftrightarrow \Phi(a) = 1 - \frac{\alpha}{2}$ , avec  $\Phi$  est la fonction de répartition de la loi  $N(0,1)$ .

(On dit que  $a$  est le quantile d'ordre  $1 - \frac{\alpha}{2}$  qu'on peut noter  $z_{1-\frac{\alpha}{2}}$ . On lit directement la valeur  $a$  dans la table de la loi normale  $N(0,1)$ , par exemple, si  $\alpha = 5\%$  alors  $a = z_{0,975} = 1,96$ ).

D'après le T.C.L. on approche

$$P\left(\sqrt{n} \left| \frac{\bar{X} - \mu}{\sigma} \right| \leq a\right)$$

par

$$P(|Z| \leq a)$$

Ce qui nous donne l'intervalle de confiance  $I(\mu)$  au niveau de confiance  $1 - \alpha$  :

$$I(\mu) = \left[ \bar{X} - a \frac{\sigma}{\sqrt{n}}; \bar{X} + a \frac{\sigma}{\sqrt{n}} \right]$$

**$\sigma^2$  inconnue :**

On remplace  $\sigma^2$  par son estimateur sans biais et consistant, la variance empirique

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

et on a toujours le résultat

$$\sqrt{n} \left( \frac{\bar{X} - \mu}{S} \right) \xrightarrow{\mathcal{L}} Z$$

avec  $Z \sim N(0,1)$ . (ce résultat est dû au T.C.L. et au fait que  $S \xrightarrow{P} \sigma$ )

L'intervalle de confiance  $I(\mu)$  au niveau de confiance  $1 - \alpha$  est :

$$I(\mu) = \left[ \bar{X} - a \frac{S}{\sqrt{n}}; \bar{X} + a \frac{S}{\sqrt{n}} \right]$$

avec  $a$  est le quantile d'ordre  $1 - \frac{\alpha}{2}$  càd  $P(|Z| \leq a) = 1 - \alpha$ .

**Remarque :** On peut utiliser ces approximations même si on connaît la loi de l'échantillon, il est plus pratique pour une taille d'échantillon assez grande de se restreindre à la loi normale comme loi de l'estimateur, néanmoins on doit vérifier certaines conditions préalables pour avoir une bonne approximation, en général on doit au moins avoir la condition  $n > 30$ .

### 3.2.2. Cas où la loi de l'échantillon est normale

Supposons que  $X$  suit une loi normale  $N(\mu, \sigma^2)$  (cas d'un petit échantillon), là aussi on peut distinguer deux cas :

**$\sigma^2$  connue :**

On a la loi exacte de l'estimateur  $\bar{X}$ ,

$$\sqrt{n} \left( \frac{\bar{X} - \mu}{\sigma} \right) \sim N(0,1)$$

On obtient l'intervalle de confiance  $I(\mu)$  au niveau de confiance  $1 - \alpha$  :

$$I(\mu) = \left[ \bar{X} - a \frac{\sigma}{\sqrt{n}}; \bar{X} + a \frac{\sigma}{\sqrt{n}} \right]$$

avec  $a = z_{1-\frac{\alpha}{2}}$ .

**$\sigma^2$  inconnue :**

On remplace  $\sigma^2$  par son estimateur sans biais et consistant  $S^2$  et on a la loi exacte de l'estimateur

$$\sqrt{n} \left( \frac{\bar{X} - \mu}{S} \right) \sim t(n-1)$$

Puisque la loi de Student est symétrique, on peut choisir un intervalle  $I(\mu)$  symétrique :

$$I(\mu) = \left[ \bar{X} - a \frac{S}{\sqrt{n}}; \bar{X} + a \frac{S}{\sqrt{n}} \right]$$

$I(\mu)$  est l'intervalle de confiance, au niveau de confiance  $1 - \alpha$  (niveau de signification  $\alpha$ ), le scalaire  $a$  correspond au quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une loi de Student  $t(n-1)$  (qu'on doit lire sur la table de cette loi).

### 3.3. Intervalle de confiance d'une variance

Là aussi on va distinguer le cas où la loi est inconnue et le cas où la loi est normale.

### 3.3.1. Cas où la loi de l'échantillon est inconnue

On supposera ici que la taille de l'échantillon est grande. La moyenne de la population peut être connue ou pas.

L'estimateur de  $\sigma^2$  est  $S^2$  de variance

$$Var(S^2) = \frac{n}{(n-1)^2} \left( \mu_4 - \frac{n-1}{n-3} \sigma^4 \right) \underset{+\infty}{\sim} \frac{1}{n} (\mu_4 - \sigma^4)$$

$S^2$  est un estimateur s.b de  $\sigma^2$  :  $E(S^2) = \sigma^2$

On a le résultat de la convergence en loi vers la loi normale :

$$\sqrt{n} \frac{S^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{\mathcal{L}} N(0, 1)$$

On remplace dans le dénominateur un estimateur convergent de  $\mu_4 - \sigma^4$  :

$$\hat{\mu}_4 - S^4$$

avec

$$\hat{\mu}_4 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^4$$

est l'estimateur du moment centré d'ordre 4 :  $E(X - \mu)^4$

On aura toujours la convergence en loi normale

$$\sqrt{n} \frac{S^2 - \sigma^2}{\sqrt{\hat{\mu}_4 - S^4}} \xrightarrow{\mathcal{L}} N(0, 1)$$

Pour un niveau de confiance égale  $1 - \alpha$ , soit  $a = z_{1-\frac{\alpha}{2}}$  le quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une loi normale  $N(0, 1)$ .

L'intervalle de confiance  $I(\sigma^2)$  sera

$$I(\sigma^2) = \left[ S_n^2 - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\mu}_4 - S_n^4}{n}} ; S_n^2 + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\mu}_4 - S_n^4}{n}} \right]$$

### 3.3.2. Cas où la loi de l'échantillon est normale

On distinguera là aussi le cas où la moyenne de la population est connue et le cas où elle est inconnue.

**$\mu$  est inconnue :**

Un estimateur sans biais et consistant de  $\sigma^2$  est la variance empirique

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

La loi de  $S^2$  est donnée par :

$$(n-1) \frac{S^2}{\sigma^2} \sim \chi^2(n-1)$$

Fixons un niveau de confiance égale  $1 - \alpha$ , étant donné que la loi de  $\chi^2(n-1)$  n'est pas symétrique, on doit déterminer  $a$  et  $b$  tel que  $0 < a < b$ ,

$$P(Z < a) = \frac{\alpha}{2} \text{ et } P(Z \leq b) = 1 - \frac{\alpha}{2}$$

avec  $Z$  suit la loi de  $\chi^2(n-1)$ . ( $a$  et  $b$  sont lus dans la table de cette loi).

L'intervalle de confiance  $I(\sigma^2)$  vérifiant  $P(\sigma^2 \in I(\sigma^2)) = 1 - \alpha$  est donné par

$$I(\sigma^2) = \left[ \left( \frac{n-1}{b} \right) S^2 ; \left( \frac{n-1}{a} \right) S^2 \right]$$

**$\mu$  est connue :**

Un estimateur sans biais et consistant de  $\sigma^2$  est la variance empirique

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

La loi de  $S^2$  est donnée par :

$$n \frac{S^2}{\sigma^2} \sim \chi^2(n)$$

Fixons un niveau de confiance égale  $1 - \alpha$ , soit  $a$  le quantile d'ordre  $\frac{\alpha}{2}$  et  $b$  le quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une loi de  $\chi^2(n)$ .

L'intervalle de confiance  $I(\sigma^2)$  vérifiant  $P(\sigma^2 \in I(\sigma^2)) = 1 - \alpha$  est donné par

$$I(\sigma^2) = \left[ \frac{n}{b} S^2 ; \frac{n}{a} S^2 \right]$$

**Exemple :** (Examen du 28/04/2018)

Un courtier en biens immobiliers s'intéresse au prix des maisons de type A qui présentent certaines caractéristiques et qui se situent dans la ville d'Oujda. Le courtier engage un premier expert qui doit évaluer 5 maisons de type A à Oujda. Les résultats fournis par ce premier expert sont :

| Maison $i$                      | 1  | 2  | 3  | 4  | 5  | Total |
|---------------------------------|----|----|----|----|----|-------|
| Evaluation $x_i$ (en $10^4$ DH) | 65 | 72 | 68 | 63 | 82 | 350   |

$$\sum_{i=1}^5 (x_i - \bar{x})^2 = 226$$

On veut connaître le prix moyen  $\mu_p$  et la variance  $\sigma_p^2$  de l'ensemble des biens immobiliers catégorie « maison de type A ».

a. L'intervalle de confiance de  $\mu_p$  au niveau de confiance 95% est

$$I(\mu_p) = \left[ \bar{x} - t_{0,975;4} \frac{s}{\sqrt{5}}; \bar{x} + t_{0,975;4} \frac{s}{\sqrt{5}} \right]$$

$$s = \sqrt{\frac{226}{4}} = \sqrt{56,5} = 7,52 \text{ et } \bar{x} = \frac{350}{5} = 70$$

$$I(\mu_p) = \left[ 70 - 2,776 \frac{7,52}{\sqrt{5}}; 70 + 2,776 \frac{7,52}{\sqrt{5}} \right] = [60,66 ; 79,34]$$

b. L'intervalle de confiance de  $\sigma_p^2$  au niveau de confiance 95% est

$$I(\sigma_p^2) = \left[ \left( \frac{4}{\chi_{0,975;4}} \right) s^2; \left( \frac{4}{\chi_{0,025;4}} \right) s^2 \right] = \left[ \frac{4 * 56,5}{11,143}; \frac{4 * 56,5}{0,484} \right]$$

$$I(\sigma_p^2) = [20,28 ; 466,94]$$

On en déduit un intervalle de confiance de  $\sigma_p$  :

$$I(\sigma_p) = [\sqrt{20,28}; \sqrt{466,94}] = [4,5 ; 21,6]$$

### 3.4. Intervalle de confiance d'une proportion

Dans une population d'électeurs, on s'intéresse à la proportion  $p$  d'électeurs favorables à notre candidat. On voudrait construire un intervalle de confiance de la proportion  $p$ .

Pour se faire on a pris un échantillon de taille  $n$  d'électeurs et on a noté  $X_i$  la  $i^{\text{ème}}$  réponse correspondant au  $i^{\text{ème}}$  électeur, si il est favorable à notre candidat on note 1 sinon 0.

La v.a.  $X_i$  suit une loi de Bernoulli  $\mathcal{B}(p)$ , de moyenne  $E(X_i) = p$  et de variance  $Var(X_i) = p(1 - p)$ . Un estimateur consistant de  $p$  est la moyenne empirique

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

et on a sa loi asymptotique :

$$\sqrt{n} \left( \frac{\bar{p}_n - p}{\sqrt{p(1-p)}} \right) \xrightarrow{\mathcal{L}} Z$$

avec  $Z$  suit une loi  $N(0,1)$ .

Puisque la variance  $p(1-p)$  est inconnue on peut l'estimer par la méthode des moments pour obtenir un estimateur consistant mais pas nécessairement sans biais  $\bar{p}_n(1-\bar{p}_n)$ . On garde la même convergence :

$$\sqrt{n} \left( \frac{\bar{p}_n - p}{\sqrt{\bar{p}_n(1-\bar{p}_n)}} \right) \xrightarrow{\mathcal{L}} Z$$

Pour un niveau de confiance  $1 - \alpha$ , on lit dans la table de la loi  $N(0,1)$  la valeur de  $a$  tel que  $P(|Z| \leq a) = 1 - \alpha$  c'est-à-dire  $a = z_{1-\frac{\alpha}{2}}$ .

On obtient l'intervalle de confiance  $I(p)$  :

$$I(p) = \left[ \bar{p}_n - a \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}}; \bar{p}_n + a \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}} \right]$$

**Exemple :** Sur un échantillon de 200 clients d'un café, 120 d'entre eux préfèrent le café "Arabica" au café "Robusta". Une estimation de la proportion  $p$  des clients qui préfèrent l'arabica au robusta est

$$\bar{p}_n = \frac{120}{200} = 0,6$$

L'intervalle de confiance à 95% de la proportion  $p$  sera :

$$I(p) = \left[ 0,6 - 1,96 \sqrt{\frac{0,24}{200}}; 0,6 + 1,96 \sqrt{\frac{0,24}{200}} \right] = [0,6 - 0,0679; 0,6 + 0,0679]$$

Ainsi

$$I(p) = [53,2\%; 66,79\%]$$

### 3.5. Marge d'erreur et taille minimale

Dans l'estimation de la moyenne inconnue  $\mu$  par l'estimateur  $\bar{X}$ , on définit la marge d'erreur comme la différence en valeur absolue entre  $\mu$  et  $\bar{X}$  :  $|\bar{X} - \mu|$ . Elle mesure l'erreur attribuable aux fluctuations d'échantillonnage.

L'intervalle de confiance de  $\mu$  au niveau de confiance de  $1 - \alpha$  pour une variance  $\sigma^2$  connue s'écrit,

$$I(\mu) = \left[ \bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

La marge d'erreur est au plus égale

$$E = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Pour une taille d'échantillon  $n$  fixé, cette marge d'erreur dépend du niveau de confiance. En augmentant le niveau de confiance (ou en diminuant le niveau de signification  $\alpha$ ), la marge d'erreur augmente.

En pratique avant de réaliser un sondage, on peut fixer la marge d'erreur  $E$  qu'on voudrait ne pas dépasser et déterminer la taille minimale  $n$  de l'échantillon nécessaire.

$$E = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left[ \frac{z_{1-\frac{\alpha}{2}} \sigma}{E} \right]^2$$

#### Exemple :

On veut effectuer un sondage auprès des clients d'une entreprise, on s'intéresse à l'âge des clients, l'écart-type est connu égal à 6ans. On veut construire un intervalle de confiance qui contient l'âge moyen de la population avec un niveau de confiance de 95% et dont la marge d'erreur  $E$  n'excède pas 2ans.

Le niveau de confiance  $1 - \alpha = 95\%$ ,  $z_{1-\frac{\alpha}{2}} = z_{97.5\%} = z_{0.975} = 1.96$

$$n = \left[ \frac{z_{1-\frac{\alpha}{2}} \sigma}{E} \right]^2 = \left[ \frac{1.96 * 6}{2} \right]^2 = 34.57$$

On a besoin d'au moins une taille d'échantillon de 35 individus pour garantir que la moyenne calculée sur l'échantillon s'écarte de la moyenne espérée de moins de 2ans.

- Dans le cas où l'écart-type est inconnu, on doit le remplacer par une estimation  $s$  (sondage préliminaire), la marge d'erreur s'écrit,

$$E = z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

et la taille de l'échantillon nécessaire sera,

$$n = \left[ \frac{z_{1-\frac{\alpha}{2}} s}{E} \right]^2$$

- Dans le cas d'une population finie de taille  $N$ , si le tirage se fait sans remise (ESSR), alors

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} \cong \frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)$$

La marge d'erreur devient

$$E = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

Cette dernière formule nous permet de calculer la taille de l'échantillon nécessaire,

$$n = \frac{N * z_{1-\frac{\alpha}{2}}^2 * \sigma^2}{N * E^2 + z_{1-\frac{\alpha}{2}}^2 * \sigma^2}$$

### Marge d'erreur dans le cas de l'estimation d'une proportion $p$

Si on veut estimer  $p$  par  $\bar{p}_n$ , l'intervalle de confiance pour un niveau de confiance de  $1 - \alpha$  sera,

$$I(p) = \left[ \bar{p}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}}; \bar{p}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}} \right]$$

La marge d'erreur dépend de l'estimation  $\bar{p}_n$ ,

$$E = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}}$$

La taille de l'échantillon nécessaire sera

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 \bar{p}_n(1-\bar{p}_n)}{E^2}$$

On doit remplacer  $\bar{p}_n$  par une valeur, on a deux voies,

- Soit on effectue un sondage préliminaire, dont la taille de l'échantillon soit d'environ 30, pour calculer une valeur provisoire de  $\bar{p}_n$ , puis déterminer la taille  $n$  pour un deuxième sondage qui donnera plus de précision dans l'estimation.
- Soit remplacer  $\bar{p}_n$  par une valeur arbitraire entre 0 et 1.

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 p(1-p)}{E^2}, \quad p \in ]0, 1[$$

On remarquera pour cela que la fonction  $f(p) = p(1 - p)$  définie sur  $]0, 1[$  atteint son maximum pour  $p = 0.5$ ,

$$p(1 - p) \leq 0.5(1 - 0.5) = 0.5^2 = 0.25 \text{ pour tout } p \in ]0, 1[$$

C'est le cas où on choisit la plus grande dispersion de  $\bar{p}_n$ .

La taille de l'échantillon sera,

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 0.25}{E^2} = \frac{z_{1-\frac{\alpha}{2}}^2}{4E^2}$$

**Exemple :** (extrait de l'examen du 08/01/2020)

Dans une étude menée par une entreprise, on veut connaître le pourcentage  $p$  de clients qui sont satisfaits d'un nouveau produit. Le niveau de signification est fixé à 5%.

- Calculons la taille de l'échantillon à prévoir pour garantir une erreur d'estimation de  $\pm 5\%$ .

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 p(1 - p)}{E^2}$$

$p \in ]0, 1[$ , on prendra la valeur centrale  $p = 0.5$ , on aura ainsi

$$n = \frac{z_{1-\frac{\alpha}{2}}^2}{4 E^2} = \frac{1.96^2}{4 * 0.05^2} = 384.16$$

On prendra  $n = 385$ . On doit garantir l'observation de 385 individus.

- Sur un échantillon de 400 clients choisis au hasard, il y a 22% qui ne sont pas satisfaits du nouveau produit contre 78% qui sont satisfaits.

Une estimation par intervalle de confiance de la proportion  $p$  est

$$\begin{aligned} I(p) &= \left[ \bar{p}_n \pm z_{1-\frac{\alpha}{2}} * \frac{\sqrt{\bar{p}_n(1 - \bar{p}_n)}}{\sqrt{n}} \right] = \left[ 0.78 \pm 1.96 * \frac{\sqrt{0.78 * 0.22}}{\sqrt{400}} \right] = [0.78 \pm 0.0406] \\ &= [73.94\% ; 82.06\%] \end{aligned}$$

Si on veut réaliser un deuxième sondage, la taille de l'échantillon à prévoir pour garantir toujours une erreur d'estimation de  $\pm 5\%$  est

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 \bar{p}_n(1 - \bar{p}_n)}{E^2} = \frac{1.96^2 0.78 * 0.22}{0.05^2} = 263.69$$

On prend  $n = 264$ .

## Annexe

### Solution de l'exercice d'application, page 13 :

1. Calculons l'espérance mathématique :

$$\begin{aligned} E(X) &= \int_{\mathbb{R}} x f_{\theta}(x) dx = \int_0^{\theta} \frac{2x^2}{\theta^2} dx \\ &= \frac{2}{\theta^2} \left[ \frac{x^3}{3} \right]_0^{\theta} = \frac{2}{\theta^2} * \frac{\theta^3}{3} = \frac{2}{3} \theta \\ \theta &= \frac{3}{2} E(X) \Rightarrow T_n = \frac{3}{2} \bar{X} \text{ avec } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \end{aligned}$$

2. Calculons  $E(T_n)$  :

$$E(T_n) = E\left(\frac{3}{2} \bar{X}\right) = \frac{3}{2} E(\bar{X}) = \frac{3}{2} E(X) = \frac{3}{2} * \frac{2}{3} \theta = \theta$$

$T_n$  est un estimateur sans biais.

$$\bar{X} \xrightarrow{P} E(X) \Rightarrow T_n = \frac{3}{2} \bar{X} \xrightarrow{P} \frac{3}{2} E(X) = \theta$$

$T_n$  est un estimateur consistant

3. Calculons le moment d'ordre 2,  $E(X^2)$

$$\begin{aligned} E(X^2) &= \int_{\mathbb{R}} x^2 f_{\theta}(x) dx = \int_0^{\theta} \frac{2x^3}{\theta^2} dx \\ &= \frac{2}{\theta^2} \left[ \frac{x^4}{4} \right]_0^{\theta} = \frac{2}{\theta^2} * \frac{\theta^4}{4} = \frac{\theta^2}{2} \\ \text{Var}(X) &= \frac{\theta^2}{2} - \left(\frac{2}{3} \theta\right)^2 = \frac{\theta^2}{18} \end{aligned}$$

On peut déduire la variance de  $T_n$

$$\text{Var}(T_n) = \text{Var}\left(\frac{3}{2} \bar{X}\right) = \frac{9}{4} * \frac{\theta^2}{18} * \frac{1}{n} = \frac{\theta^2}{8n}$$

4. Calculons  $E(M_n)$

$$\begin{aligned} E(M_n) &= \int_{\mathbb{R}} x g_{\theta}(x) dx = \int_0^{\theta} \frac{2nx^{2n}}{\theta^{2n}} dx \\ &= \frac{2n}{\theta^{2n}} \left[ \frac{x^{2n+1}}{2n+1} \right]_0^{\theta} = \frac{2n}{2n+1} * \frac{\theta^{2n+1}}{\theta^{2n}} = \frac{2n}{2n+1} \theta \end{aligned}$$

Calculons  $E(M_n^2)$ ,

$$\begin{aligned}
 E(M_n^2) &= \int_{\mathbb{R}} x^2 g_{\theta}(x) dx = \int_0^{\theta} \frac{2nx^{2n+1}}{\theta^{2n}} dx \\
 &= \frac{2n}{\theta^{2n}} \left[ \frac{x^{2n+2}}{2n+2} \right]_0^{\theta} = \frac{2n}{2n+2} * \frac{\theta^{2n+2}}{\theta^{2n}} = \frac{2n}{2n+2} \theta^2 = \frac{n}{n+1} \theta^2 \\
 Var(M_n) &= \frac{n}{n+1} \theta^2 - \left( \frac{2n}{2n+1} \theta \right)^2 = \frac{n(2n+1)^2 - (n+1) * 4n^2}{(n+1)(2n+1)^2} \theta^2 \\
 Var(M_n) &= \frac{n}{(n+1)(2n+1)^2} \theta^2
 \end{aligned}$$

5. On veut maintenant construire un estimateur de  $\theta$ , on va utiliser l'expression de  $E(M_n)$

$$E(M_n) = \frac{2n}{2n+1} \theta \Rightarrow \theta = \frac{2n+1}{2n} E(M_n)$$

Un estimateur sans biais de  $\theta$  sera

$$T'_n = \frac{2n+1}{2n} M_n \text{ et on verifie bien } E(T'_n) = \theta, \quad T'_n \text{ est un e.s.b.}$$

6. Calculons  $Var(T'_n)$ ,

$$\begin{aligned}
 Var(T'_n) &= Var\left(\frac{2n+1}{2n} M_n\right) = \left(\frac{2n+1}{2n}\right)^2 Var(M_n) \\
 Var(T'_n) &= \frac{(2n+1)^2}{(2n)^2} * \frac{n}{(n+1)(2n+1)^2} \theta^2 \\
 Var(T'_n) &= \frac{1}{4n(n+1)} \theta^2
 \end{aligned}$$

$T_n$  et  $T'_n$  sont deux e.s.b. de  $\theta$ , pour les comparer on doit comparer leurs variances :

$$\begin{aligned}
 Var(T_n) &= \frac{\theta^2}{8n} \quad \text{et} \quad Var(T'_n) = \frac{1}{4n(n+1)} \theta^2 \\
 \frac{Var(T_n)}{Var(T'_n)} &= \frac{\theta^2}{8n} * \frac{4n(n+1)}{\theta^2} = \frac{n+1}{2} \\
 n \geq 1 &\Leftrightarrow n+1 \geq 2 \Leftrightarrow \frac{n+1}{2} \geq 1
 \end{aligned}$$

On aura ainsi :  $Var(T_n) \geq Var(T'_n)$

**Conclusion :** L'estimateur  $T'_n$  est plus efficace que l'estimateur  $T_n$ .

[Retour au paragraphe p.14](#)

**Propriétés :** Soit un échantillon  $X_1, X_2, \dots, X_n$  de  $n$  v.a. à valeurs réelles *i. i. d.*

Si  $E(X_1) = \mu$  et  $Var(X_1) = \sigma^2$ , alors on a

$$E(\bar{X}_n) = \mu \quad \text{et} \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

**Preuve :** On a :

$$E(\bar{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{n \mu}{n} = \mu$$

$$Var(\bar{X}_n) = Var\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{n \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$Var\left(\sum_{i=1}^n X_i\right) = cov\left(\sum_{i=1}^n X_i; \sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sum_{j=1}^n cov(X_i, X_j) = \sum_{i=1}^n Var(X_i)$$

En effet si  $i \neq j$ , les deux v.a.  $X_i$  et  $X_j$  sont indépendantes et une condition nécessaire pour cette indépendance est que leur covariance soient nulle  $cov(X_i, X_j) = 0$

$$cov(X_i, X_j) = \begin{cases} 0 & \text{si } i \neq j \\ var(X_i) & \text{si } i = j \end{cases}$$

**Propriétés :** Soit un échantillon  $X_1, X_2, \dots, X_n$  de  $n$  v.a. à valeurs réelles *i. i. d.*

Si  $E(X_1) = \mu$  et  $Var(X_1) = \sigma^2$ , alors on a

$$E(S_n^2) = \sigma^2 \quad \text{et} \quad Var(S_n^2) = \frac{n}{(n-1)^2} \left( \mu_4 - \frac{n-1}{n-3} \sigma^4 \right)$$

Avec  $\mu_4 = E(X^4)$ .

**Preuve :** Montrons le premier résultat. On a :

$$E(S_n^2) = E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2\right) = \frac{1}{n-1} \sum_{i=1}^n E[(X_i - \bar{X}_n)^2]$$

$$\begin{aligned} E[(X_i - \bar{X}_n)^2] &= E\left[(X_i - \mu) + (\mu - \bar{X}_n)\right]^2 \\ &= E[(X_i - \mu)^2] + E[(\mu - \bar{X}_n)^2] + 2 E[(X_i - \mu)(\mu - \bar{X}_n)] \end{aligned}$$

$$\begin{aligned}
E[(X_i - \bar{X}_n)^2] &= \text{Var}(X_i) + \text{Var}(\bar{X}_n) + 2 E[(X_i - \mu) (\mu - \bar{X}_n)] \\
&= \sigma^2 + \frac{\sigma^2}{n} + 2 E[(X_i - \mu) (\mu - \bar{X}_n)]
\end{aligned}$$

$$\begin{aligned}
E[(X_i - \mu) (\mu - \bar{X}_n)] &= E[X_i \mu - X_i \bar{X}_n - \mu^2 + \mu \bar{X}_n] = \mu E(X_i) - E(X_i \bar{X}_n) - \mu^2 + \mu E(\bar{X}_n) \\
&= \mu^2 - E(X_i \bar{X}_n) - \mu^2 + \mu^2 = \mu^2 - E(X_i \bar{X}_n)
\end{aligned}$$

$$E(X_i \bar{X}_n) = \frac{1}{n} \sum_{j=1}^n E(X_i X_j) = \frac{1}{n} ((n-1) \mu^2 + \sigma^2 + \mu^2) = \mu^2 + \frac{\sigma^2}{n}$$

$$E[(X_i - \mu) (\mu - \bar{X}_n)] = \mu^2 - \mu^2 - \frac{\sigma^2}{n} = -\frac{\sigma^2}{n}$$

$$\begin{aligned}
E[(X_i - \bar{X}_n)^2] &= \text{Var}(X_i) + \text{Var}(\bar{X}_n) + 2 E[(X_i - \mu) (\mu - \bar{X}_n)] = \sigma^2 + \frac{\sigma^2}{n} - 2 \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} \\
&= \sigma^2 \left( \frac{n-1}{n} \right)
\end{aligned}$$

$$E(S_n^2) = \frac{1}{n-1} \sum_{i=1}^n \sigma^2 \left( \frac{n-1}{n} \right) = \sigma^2 \frac{n-1}{n-1} = \sigma^2$$

## Notations et abréviations

|                  |                                                                  |
|------------------|------------------------------------------------------------------|
| ES               | <i>Echantillon Simple</i>                                        |
| ESSR             | <i>Echantillon Simple Sans Remise</i>                            |
| ESAR             | <i>Echantillon Simple Avec Remise</i>                            |
| v.a.             | <i>variable aléatoire</i>                                        |
| <i>i. i. d.</i>  | <i>indépendant identiquement distribué</i>                       |
| MSE              | <i>Mean Square Error : erreur quadratique moyenne</i>            |
| $X, Y, Z, \dots$ | <i>Les majuscules sont réservées aux variables aléatoires</i>    |
| $x, y, z, \dots$ | <i>Les minuscules sont réservées aux observations de la v.a.</i> |
| e.s.b            | <i>estimateur sans biais</i>                                     |

## Bibliographie

- D.R. Anderson, D.J. Sweeney, T.A. Williams, (2010) *Statistiques pour l'économie et la gestion*, De Boeck.
- P.-C. Pupion (2008) *Statistiques pour la gestion, applications avec Excel et SPSS*, Gestion Sup, Dunod
- G. Baillargeon (2004), *Outils statistiques et analyse de données*, SMG.
- J.-M. Martel (1972) *Décision et inférence statistique en affaires*, Les Presses de l'université Laval.
- B. Grais, *Méthodes statistiques*, Dunod.
- T.H. Wonnacott, R.J. Wonnacott, *Statistique, économie-gestion-sciences-médecine*, ECONOMICA.
- B. Tribout, *Statistique pour économistes et gestionnaire*, Pearson.
- C. Hurlin, *Statistique et probabilités en économie-gestion*, Dunod.