# Chapitre 2 Estimation

## I. Introduction

Le problème de l'estimation statistique est le suivant :
On cherche à connaître les valeurs de certaines caractéristiques (paramètres) d'une variable aléatoire grâce à des observations réalisées sur un échantillon.

**Exemple.**

a. Quelle est la fréquence d'une certaine maladie dans une population donnée ?
b. Quel est le taux de glycémie moyen dans la population ?

Il est impossible de répondre de façon précise à ces questions. On y apporte deux types de réponses :

- On produit une valeur qui nous semble être la « meilleure » possible :
  c'est **l'estimation ponctuelle**

- On produit un intervalle de valeurs possibles, compatibles avec les observations :
  c'est **l'estimation par intervalle de confiance**

## Notations

Dans tout ce qui suit

- La variable aléatoire d'intérêt sera notée $X$
- Le paramètre à estimer est noté $\theta$
- La moyenne de la population (ou la moyenne théorique ) soit $E(X)$ sera notée $m$
- La moyenne de l'échantillon (ou la moyenne observée ) sera notée $\bar{x}$
- La variance de la population (ou la moyenne théorique) soit $V(X)$ sera notée $\sigma^2$
- La variance de l'échantillon (ou la variance observée) sera notée $S^2$
- La proportion d'un certain événement A dans la population soit $P(A)$ sera notée $p_o$
- La proportion de l'événement A dans l'échantillon ou (observée) sera notée $p$

**Rappels :**

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \quad , \quad S^2=\frac{1}{n}\sum_{i=}^{n}\left(x_i-\bar{x}\right)^2 = \quad \frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2 \quad , \quad p=\frac{|A|}{n}$$

$n$ étant la taille de l'échantillon choisi $(x_1, x_2, \ldots x_n)$.

## II. Estimation ponctuelle

### 1. Définition

A partir d'un *n*-échantillon ($X_1,\ldots,X_n$ ) de la variable aléatoire *X*, on construit une nouvelle variable aléatoire $t(X_1,\ldots,X_n$ ) dont les réalisations se rapprochent de la valeur du paramètre à estimer $\theta$.
Cette nouvelle variable notée *T* ou $T_n = t(X_1,\ldots,X_n$ ) est appelée **estimateur de $\theta$.**

**Exemple**

$T = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ est l'estimateur naturel de $m = E(X)$

La moyenne de la population $m$ est estimée par la moyenne de l'échantillon $\overline{X}$

### 2. Propriétés

Un estimateur est une fonction de l'échantillon : c'est donc une variable aléatoire qui possède une densité (au sens large), une moyenne et une variance.
Ces deux grandeurs permettent de comparer, dans une certaine mesure, plusieurs estimateurs entre eux.

***a.*** **Estimateur sans biais**

- Le biais d'un estimateur *T* est noté B(*T*). C'est la différence moyenne entre sa valeur et celle de la quantité qu'il estime.

$$B(T) = E(T-\theta) = E(T) - \theta$$

- Un estimateur *T* est dit être un estimateur sans biais du paramètre $\theta$ on écrit *T* ESB de $\theta$ si et seulement si

$$B(T) = 0 \iff E(T-\theta) = 0 \iff E(T) = \theta$$

**b. Variance minimum**

- La variance d'un estimateur *T* est définie par $V(T) = E(T-E(T))^2 = E(T^2) - E^2(T)$

- Si deux estimateurs sont sans biais, le « meilleur » d'entre eux est celui qui a la plus petite variance

**c. Erreur quadratique moyenne**

- L'erreur quadratique moyenne d'un estimateur *T* est notée *EQM(T).*

$$EQM(T) = E(T-\theta)^2 = V(T) + B^2(T)$$

- Elle permet de comparer plusieurs estimateurs entre eux qu'ils soient sans biais ou non. Le « meilleur » d'entre eux est celui qui a la plus petite erreur quadratique moyenne.

- Si *T* est un estimateur sans biais du paramètre $\theta$ alors $B(T) = 0 \iff EQM(T) = V(T)$

## 3. Exemples

### a. Estimation de la moyenne

On montre que $\overline{X}$ est le « meilleur » estimateur de *m*. C'est un estimateur sans biais et à variance minimum de *m*.
On écrit : $\hat{m}=\bar{x}$ , $\hat{m}$ est la valeur estimée de *m*.

### b. Estimation de la variance

De la même façon, on pourrait estimer la variance de la population $\sigma^2$ par la variance de l'échantillon $S^2$, mais $S^2$ n'est pas un estimateur sans biais de $\sigma^2$ car $E(S^2)=\frac{n-1}{n}\sigma^2 \neq \sigma^2$,

alors on lui préférera $S'^2 = \frac{n}{n-1}S^2$ qui est un estimateur sans biais de $\sigma^2$.

On écrira : $\hat{\sigma}^2 = \frac{n}{n-1}S^2$ , $\hat{\sigma}^2$ est la valeur estimée de $\sigma^2$.

On a $S'^2 = \frac{n}{n-1}S^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=}^{n}\left(x_i-\bar{x}\right)^2 = \frac{1}{n-1}\sum_{i=}^{n}\left(x_i-\bar{x}\right)^2$. $S'^2$ est appelée variance corrigée.

### c. Estimation de la proportion

On montre que *p* est le « meilleur » estimateur de $p_o$.
On écrit : $\hat{p}=p_o$ , $\hat{p}$ est la valeur estimée de $p_o$.

## 4. Proposition

Soit *T* un estimateur d'un paramètre $\theta$**,** et soit *f* une fonction réelle alors *f (T)* est un estimateur de *f* ( $\theta$ )**.**

ex : $S'^2$ est l' estimateur $\sigma^2$ alors $\sqrt{S'^2}=S'$ est l'estimateur de $\sqrt{\sigma^2}=\sigma$, $\hat{\sigma} = S' = \sqrt{\frac{n}{n-1}}S$

***Exercice 1 série5***

Dans l'analyse du sang d'un échantillon de 100 malades pris au hasard dans une population de personnes hospitalisées pour des anomalies sanguines, on a relevé le poids de calcium X et proposé les résultats suivants : $\sum x_i$ = 12000mg ; $\sum x_i^2$ = 1449900 mg$^2$

1) Estimer la moyenne, la variance et l'écart-type d'une mesure pour individu de la population

**solution**

**a.** On sait que $\hat{m}=\bar{x}$ , on calcule $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i = \frac{12000}{100}=120$ alors $\hat{m}=120\,\text{mg}$

**b.** On sait que $\hat{\sigma}^2 = \frac{n}{n-1}S^2$ , on calcule $S^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 = \frac{1449900}{100}-(120)^2 = 99\ \text{mg}^2$

alors $\hat{\sigma}^2 = \frac{100}{99}99 = 100\ \text{mg}^2$

**c.** $\hat{\sigma}^2 = 100$ alors $\hat{\sigma} = \sqrt{100} = 10$ mg

## **III.** Estimation par intervalles de confiance.

### 1. Introduction.

#### a. Définitions

- Estimer un paramètre $\theta$ par un intervalle de confiance [*a, b*] (qu'on notera IC) avec une certaine probabilité 1- $\alpha$ , est équivalent à dire que P( $\theta \in$ [*a, b*] ) = 1- $\alpha$
- Cette probabilité (1- $\alpha$ ) est appelée niveau de confiance ou taux de sécurité
- $\alpha$ est le risque ou l'erreur (de première espèce). $\alpha$ est fixé au préalable, ( $\alpha$ < 10%)
  En général on prend $\alpha$ =5%

#### b. Signification d'un intervalle de confiance

Les intervalles de confiance sont basés sur les estimateurs ponctuels du paramètre $\theta$ , qui sont des variables aléatoires, prenant donc différentes valeurs selon l'échantillon choisi.
Si on calcule les intervalles de confiance sur plusieurs échantillons, on est sûr que parmi eux, il y en aurait (1- $\alpha$ ) $\times$100% qui contiendraient la vraie valeur du paramètre à estimer $\theta$ .

### 2. 1. Intervalle de confiance pour la moyenne

On veut trouver deux réels *a* et *b* tel que $P(\ a \le m \le b\ ) = 1-\ \alpha$ (1)

#### a. $\sigma^2$ connu

##### i. $n < 30$

Dans le cas où n $< 30$, on doit supposer que la population mère est de loi normale $N(m, \sigma^2)$.

Dans ce cas $\overline{X} \rightarrow N(m, \frac{\sigma^2}{n})$ et $Z = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1)$

On cherchera alors deux réels $z_1$ et $z_2$ tels que $P(\ z_1 \le Z \le z_2) = 1-\ \alpha$

On choisira $z_1$ et $z_2$ symétriques : $z_1 = -\ z_2$, de façon à avoir et $P(Z \le z_1) = \frac{\alpha}{2} = P(Z \ge z_2)$

L'intervalle de confiance pour la moyenne au risque $\alpha$ s'écrit donc comme :

$$IC_{(\alpha)}(m) = \left[ \overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \ ,\ \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] = \left[ \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

$z_{\frac{\alpha}{2}}$ et $z_{1-\frac{\alpha}{2}}$ sont respectivement les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi $N(0,1)$

##### ii. $n \ge 30$

D'après le TCL, on sait que $\overline{X} \rightarrow N(m, \frac{\sigma^2}{n})$ on retrouve le même intervalle de confiance.

#### b. $\sigma^2$ inconnu

donc estimé par : $\hat{\sigma} = \sqrt{\frac{n}{n-1}}\, S$ . On pose $T = \frac{\overline{X} - m}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{\overline{X} - m}{\frac{S}{\sqrt{n-1}}} = \frac{\overline{X} - m}{\frac{S'}{\sqrt{n}}} \rightarrow St_{(n-1)}$

**i.** $n < 30$

On cherchera deux réels $t_1$ et $t_2$ tels que $P(t_1 \le T \le t_2) = 1-\alpha$ et $P(T \le t_1) = \frac{\alpha}{2} = P(T \ge t_2)$

Alors on obtient

$$IC_{(\alpha)}(m) = \left[ \overline{X} \pm t_{(n-1)}(\alpha) \frac{S}{\sqrt{n-1}} \right] = \left[ \overline{X} \pm t_{(n-1)}(\alpha) \frac{S'}{\sqrt{n}} \right]$$

**ii**. $n \ge 30$

$T$ est de loi de Student à ($n$-1) degré de liberté qu'on ne peut pas approximer par une $N(0,1)$ .

$$IC_{(\alpha)}(m) = \left[ \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \right] = \left[ \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S'}{\sqrt{n}} \right]$$

**On donne quelques valeurs particulières de** $z_{1-\frac{\alpha}{2}}$

| $\alpha$ | 1% | 5% | 10% |
|---|---|---|---|
| $z_{1-\frac{\alpha}{2}}$ | 2.58 | 1.96 | 1.64 |

**Remarque :** L'intervalle de confiance pour la moyenne est centré en $\overline{X}$

### *2.2. Exercice 1 série 6*

2) Donner l'intervalle de confiance à 95%, puis à 99% du poids moyen de calcium pour l'ensemble des malades et comparer ces deux intervalles

**Solution**

2) $IC_{(\alpha)}(m) =$ ?

On est dans le cas où $n$=100>30 , $\sigma^2$ inconnue .

**Attention :** Ce qu'on connaît (calculé) est la variance $S^2$ la variance estimée et de l'échantillon alors $IC_{(\alpha)}(m) = \left[ \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \right] = \left[ \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S'}{\sqrt{n}} \right]$

AN (Application Numérique)

- $1-\alpha = 95\% \Rightarrow \alpha = 5\% \Rightarrow z_{1-\frac{\alpha}{2}} = 1.96$

$$IC_{(5\%)}(m) = \left[ \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \right] \left[ 120 \pm 1.96 \sqrt{\frac{99}{100-1}} \right] = [120 \pm 1.96] = [118.04 \ , 121.96]$$

ou bien $$IC_{(5\%)}(m) = \left[ \overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S'}{\sqrt{n}} \right] = \left[ 120 \pm 1.96 \sqrt{\frac{100}{100}} \right] = [118.04 \ , 121.96]$$

- $1-\alpha = 99\% \Rightarrow \alpha = 1\% \Rightarrow z_{1-\frac{\alpha}{2}} = 2.58$

$$\text{donc } IC_{(1\%)}(m) = \left[120 \pm 2.58\sqrt{\frac{99}{100-1}}\right] = [120 \pm 2.58] = [117.42 \ , 122.58]$$

Comparaison des deux intervalles :

L'IC au risque de 1% est plus large que celui au un risque de 5% $\Leftrightarrow$ $IC_{(5\%)} \subset IC_{(1\%)}$

D'une manière générale, plus le risque $\alpha$ est grand, moins large est l'intervalle de confiance.

$$IC_{(\alpha_1)}(m) \subset IC_{(\alpha_2)}(m) \ \forall \alpha_1 > \alpha_2$$

### 2.3 Précision d'un l'intervalle de confiance

#### a. Définition

On définit la précision comme étant la demi longueur de l'intervalle de confiance.

Si $IC_{(\alpha)}(\theta) = [a, b]$ alors la précision notée $h = \frac{b-a}{2} = \frac{L}{2}$, L étant la longueur de l'intervalle de confiance.

Le « meilleur » intervalle, celui qui est le plus précis est celui qui a la plus petite longueur (ou demi longueur).

#### b. Exemple

Dans le cas de la moyenne, on a vu que l'intervalle de confiance pour la moyenne est centré en $\overline{X}$ alors il s'écrit toujours comme

$$IC_{(\alpha)}(m) = \left[\overline{X} - h \ , \ \overline{X} + h\right]$$

On a $P(\overline{X} - h \le m \le \overline{X} + h) = 1 - \alpha$, donc plus la longueur *h* est petite , plus $\overline{X}$ se rapproche de *m* et « meilleure » est la précision de l'estimation.

#### c. Risque et précision

On a vu que $IC_{(\alpha_1)}(m) \subset IC_{(\alpha_2)}(m) \quad \forall \alpha_1 > \alpha_2$

**D'une manière générale, plus le risque $\alpha$ est grand, moins large est l'intervalle de confiance, donc plus précis il est.**

#### d. Taille de l'échantillon

1) On veut connaître la taille *n* de l'échantillon nécessaire pour avoir une précision de $h_0$
2) On veut connaître la taille minimale $n_0$ de l'échantillon, nécessaire pour avoir une précision **d'au moins** $h_0$

**Exemple :**

Choisissons le cas le plus fréquent dans le calcul des intervalles de confiance pour la moyenne ***n* > 30 et $\sigma^2$ inconnu.**

$$IC_{(\alpha)}(m) = \left[\overline{X} \pm h\right] = \left[\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}\right]$$

1) $n$ = ? tel que $h = h_0$

On a $h = z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} = h_0 \Rightarrow n = \left(\frac{z_{1-\frac{\alpha}{2}}}{h_0} S\right)^2 + 1$

2) $n$ = ? tel que $h \le h_0$

On a $h = z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \le h_0 \Rightarrow n \ge \left(\frac{z_{1-\frac{\alpha}{2}}}{h_0} S\right)^2 + 1 \Rightarrow n_0 = \text{E}\left[\left(\frac{z_{1-\frac{\alpha}{2}}}{h_0} S\right)^2 + 1\right] + 1$

où E[ . ] désigne la partie entière

***Exercice 6 de la série 6:***

On a mesuré la pression sanguine chez 32 chèvres. L'intervalle de confiance à 95% de la pression moyenne est [4,7 ; 5,5] en cm de mercure (Hg).

1) Déterminer la pression moyenne correspondant aux 32 chèvres.
2) Quel est le nombre minimal de chèvres qu'il aurait fallu utiliser pour obtenir une précision d'au moins 2% ? On supposera que la moyenne et la variance restent inchangées.

**Solution**

1) on a $n$ =32 > 30 $IC_{(5\%)}(m) = [4{,}7 ; 5{,}5]$ $\bar{x}$ = ?

L'intervalle de confiance pour la moyenne est centré en $\bar{x}$ alors $\bar{x} = \frac{4.7+5.5}{2} = 5.1$(Hg).

2) $n$ = ? tel que $h \le h_0 = 0.02$

On sait que $h = z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}$ car la variance de la population $\sigma^2$ est inconnue**.**

On doit calculer $S$ en supposant que dans ce nouvel échantillon la moyenne et la variance restent inchangées.

$h = z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \Rightarrow S = \frac{h\sqrt{n-1}}{z_{1-\frac{\alpha}{2}}}$

A.N $\alpha = 5\% \Rightarrow z_{1-\frac{\alpha}{2}} = 1.96$ , $h = \frac{5.5-4.7}{2} = 0.4 \Rightarrow S = \frac{0.4\sqrt{32-1}}{1.96} = 1.14$

$h = z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \le h_0 \Rightarrow n \ge \left(\frac{z_{1-\frac{\alpha}{2}}}{h_0} S\right)^2 + 1 \Rightarrow n \ge \left(\frac{1.96}{0.02} 1.14\right)^2 + 1 = 12482.36$

La taille minimale pour avoir une précision d'au moins 2% est $n_0$ = E[12482.36 ] +1= 12483

$n_0$ est le plus petit entier supérieur à 12482.36 $n_0$ = 12483

***Exercice 4 de la série 6:***

On a mesuré chez 50 adultes le taux d'acide urique et on a obtenu une moyenne de 47,3 mg/l et un écart-type de 1,85 mg/l.

1) Donner les intervalles de confiance aux niveaux 95% et 90% pour le taux moyen d'acide urique dans la population. Comparer ces deux intervalles.
2) On suppose que la variance dans la population est égale à 2,5 $m/g^2$. Donner le nombre d'adultes n que l'on doit examiner pour que l'intervalle de confiance soit [46,9 ; 48,1] au risque de 5%, puis au risque de 1%. Conclure.

**Solution**

1) voir solution exercice 1

2) $\sigma^2$ est connue, $\sigma^2 = 2.5$. $n$ = ?

On sait que $h = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \Rightarrow n = \left( \frac{z_{1-\frac{\alpha}{2}}}{h} \right)^2 \sigma^2$

AN :

- $\alpha = 5\% \Rightarrow z_{1-\frac{\alpha}{2}} = 1.96$ , $\sigma^2 = 2.5$, $h = \frac{48.1 - 46.9}{2} = 0.6 \Rightarrow n = \left( \frac{1.96}{0.6} \right)^2 2.5 = 26.67 \approx 27$
- $\alpha = 1\% \Rightarrow z_{1-\frac{\alpha}{2}} = 2.58 \Rightarrow n = \left( \frac{2.58}{0.6} \right)^2 2.5 = 46.23 \approx 47$

### 3. Intervalle de confiance pour la proportion

**<u>Remarques</u>**

- Les intervalles de confiance pour la proportion ne se calculent que pour les grands échantillons ( $n > 30$).
- Pour les petits échantillons, il existe des abaques qui donnent les intervalles de confiances pour la proportion pour certaines valeurs de $\alpha$ et de $n$

**Calcul de l'intervalle de confiance**

On veut trouver deux réels $a$ et $b$ tel que $P(a \leq p_0 \leq b) = 1 - \alpha$

En vérifiant toujours les conditions de validité : $n > 30$, $np_0 \geq 5$ et $nq_0 \geq 5$

On a vu que la fréquence $P \rightarrow N\left( p_0 , \frac{p_0(1-p_0)}{n} \right)$ ( cf cours échantillonnage)

On pose $Z = \frac{P - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ . $Z$ est de loi $N(0,1)$ (cf cours échantillonnage).

On se retrouve dans les mêmes dispositions que lors du calcul de l'intervalle de confiance pour la moyenne. Alors

$$P[\ p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}} \le p_0 \le p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0}{n}}\ ] = 1-\alpha$$

En estimant $p_0$ sous le radical par $p$, on obtient

$$IC_{(\alpha)}(p_0) = \left[\ p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\ ,\ p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\ \right] = \left[\ p \pm\ z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\ \right]$$

**Remarque :**

$IC_{(\alpha)}(p_0)$ est centré en $p \Leftrightarrow IC_{(\alpha)}(p_0) = [p-h,\ p+h] = [p \pm h\ ]$

***Exercice 7 de la série 6:***

Une étude sur un échantillon de 100 souris d'une certaine race a montré que la présence de cancers spontanés est de 25%

1) Donner l'intervalle de confiance au risque de 5% pour la proportion de cancers spontanés dans la population.
2) Quelle taille d'échantillon minimale doit-on prendre pour avoir une précision d'au moins 1% ?

**Solution**

1) On vérifie les conditions de validité : n > 30 , $n$> 30, $np \ge 5$ et $nq \ge 5$

donc on peut calculer $IC_{(\alpha)}(p_0)$ $\qquad IC_{(\alpha)}(p_0) = \left[\ p \pm\ z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\ \right]$

AN : Quand $\alpha$ n'est pas donné, prendre toujours $\alpha = 5\%$

$$IC_{(0.05)}(p_0) = \left[\ 0.25 \pm\ 1.96\sqrt{\frac{0.25\times 0.75}{100}}\ \right] = [0.25\ \pm\ 0.0849] = [0.1651\ ,0.3349]$$

2) $n$ = ? tel que $h \le h_0 = 0.01$

$n$ = ? tel que $h = z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le h_0 \ \Rightarrow\ n \ge \left(\frac{z_{1-\frac{\alpha}{2}}}{h_0}\right)^2 pq$

AN : $n \ge \left(\frac{1.96}{0.01}\right)^2 0.25\times 0.75 = 7203$ , la taille minimale = 7203