Chapitre 1

Rappels sur les propriétés de la loi Normale

1.1 Quelques propriétés de la loi Normale

Avant toute chose nous allons dans ce chapitre introductif rappeler les notions les plus importantes de la loi normale qui nous serviront en permanence par la suite.

Définition 1 *On appelle variable aléatoire normale ou gaussienne toute variable aléatoire absolument continue dont la densité de probabilité f est définie par*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

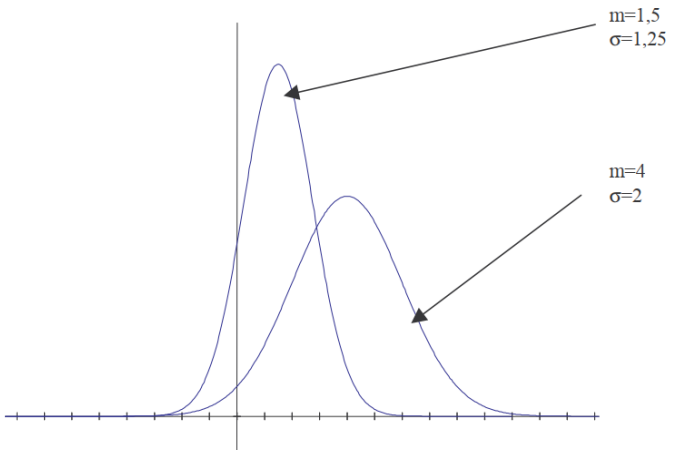


FIG. 1-1 – La courbe représentative de la loi $\mathcal{N}(m, \sigma)$

1.1.1 Fonction de répartition de la loi Normale $\mathcal{N}(m, \sigma)$

Soit Z une variable aléatoire qui suit une loi Normale de moyenne $m = \mu$ et d'écart type σ et dont la densité de probabilité est notée f et la fonction de répartition de la loi normale est donnée par :

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

1.1.2 Loi normale centrée réduite

On dit que X suit une loi normale centrée réduite et on note $X \sim \mathcal{N}(0, 1)$ si sa loi admet pour densité la fonction

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

La courbe représentative de la fonction f est donnée par la Figure 1.2. Sa densité de probabilité est la fonction f définie par

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

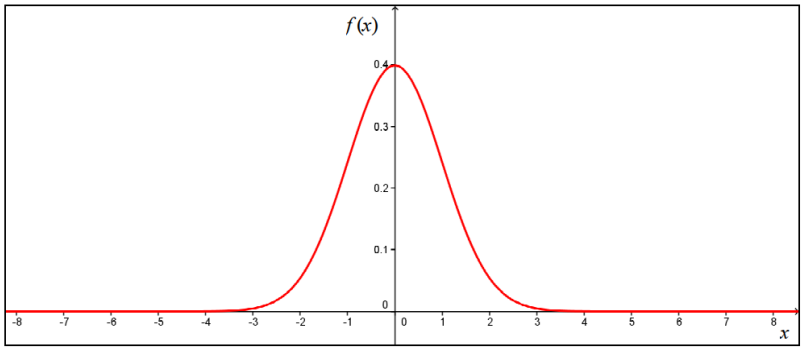


FIG. 1-2 – Représentation graphique de f

— Cette fonction f est paire, la courbe a un axe de symétrie qui est la droite des ordonnées. En $x = 0$, la fonction f vaut

$$f(0) = \frac{1}{\sqrt{2\pi}}.$$

— Les points d'inflexion de la fonction g se trouvent en $x = -1$ et $x = 1$.

— En général, les valeurs de $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ sont données à l'aide d'une table pour $x \geq 0$

Définition 2 *La fonction de répartition de la loi* $\mathcal{N}(0,1)$ *est souvent notée :*

$$\Phi(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}}\, dt$$

Théorème 3 *Si* $X \sim \mathcal{N}(m, \sigma)$ *alors* $Z = \dfrac{X - m}{\sigma} \sim \mathcal{N}(0,1)$

On utilise la lettre Z pour désigner une loi normale centrée réduite.

1.1.3 Calculs de probabilités

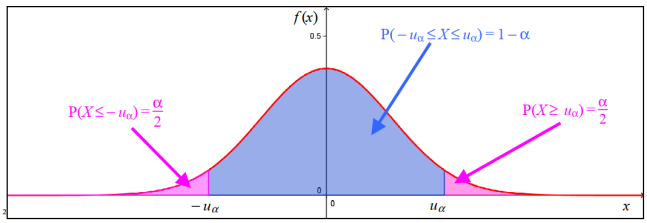
Théorème 4 *Si une variable aléatoire X suit une loi normale centrée réduite alors pour tous réels a et b tels que $a \leq b$, on a :*

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = \Phi(b) - \Phi(a)$$
$$P(X \geq a) = 1 - P(X \leq a) = 1 - \Phi(a)$$
$$P(X \leq -|a|) = 1 - \Phi(|a|)$$

On a :

$$\forall x \in \mathbb{R}_+ \ \Phi(x) = 1 - \Phi(-x)$$
$$P(|X| \leq x) = 2\Phi(x) - 1$$

- $P(X \in IR) = 1$, l'aire de la partie comprise entre l'axe des abscisses et la courbe représentative de f est égale à 1 unité d'aire.
- La symétrie de la courbe impose : $\begin{cases} P(X \leq 0) = P(X \geq 0) = 0,5 \\ P(X \leq -a) = P(X \geq a). \end{cases}$
- $(X > a)$ et $(X \leq a)$ étant des événements contraires : $P(X > a) = 1 - P(X \leq a)$



1.1.4 Probabilité d'intervalle centré en 0

Théorème 5 *X est une variable aléatoire qui suit une loi normale centrée réduite. Soit α un réel de l'intervalle $]0,1[$. Il existe un unique réel strictement positif u_α tel que :*

$$P(-u_\alpha \leq X \leq u_\alpha) = 1 - \alpha.$$

Exemple 6 *X est une variable aléatoire qui suit une loi normale centrée réduite.*

Déterminer l'intervalle I centré en 0 tel que $P(X \in I) = 0,8$.

On donnera les bornes de l'intervalle avec une précision de 10^{-2}.

Solution 7 *On a donc : $1 - \alpha = 0,8 \Leftrightarrow \alpha = 0,2$*

On doit donc avoir :

$$\Phi(u_\alpha) = 1 - \frac{\alpha}{2} = 0,9 \Leftrightarrow u_\alpha = \Phi^{-1}(0,9)$$

on trouve :

$$u_\alpha \simeq 1,28 \text{ donc } I = [-1,28, 1,28]$$

Exemple 8 *On suppose qu'une certaine variable $X \sim \mathcal{N}(0,1)$. Pour quelle proportion d'individus est-ce que $X \leq 1,56$?*

On cherche $P(X \leq 1,56)$ (rappel : on écrit aussi $\Phi(1,56)$. On cherche $1,56$ dans la table

.	...	0.6...
1.5	...	***0.9406...***
.	...	...

Donc $P(X \leq 1,56) =$***0,9406.***

Exemple 9 *On suppose qu'une certaine variable* $X \sim \mathcal{N}(0,1)$*. Pour quelle proportion d'individus est-ce que* $X \geq 1,49$*?*

On cherche $P(X \geq 1,49)$*. On écrit d'abord*

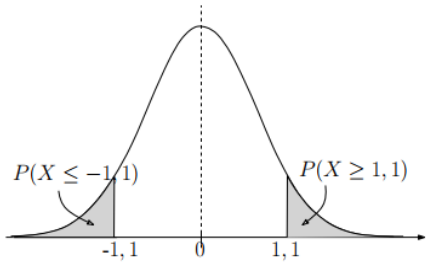
$$P\left(X \geq 1,49\right) = 1 - P\left(X \leq 1,49\right) = 1 - \Phi\left(1,49\right).$$

On cherche $\Phi\left(1,49\right)$ *dans la table.*

	...	...	0.09
	...	...	...
1.49	...	...	$\mathbf{0,9319}$
	...	...	....

On a $\Phi\left(1,49\right) = P\left(X \leq 1,49\right) = 0,9319$*, donc :*

$$P\left(X \geq 1,49\right) = 1 - 0,9319 = 0,0681$$

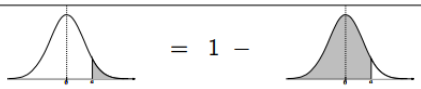
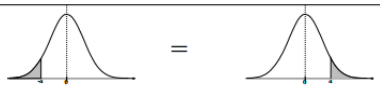
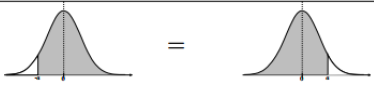


Exemple 10 *On suppose qu'une certaine variable* $X \sim \mathcal{N}(0,1)$*. Pour quelle proportion d'individus est-ce que* $X \leq -1,1$ *? On cherche* $P(X \leq -1,1)$*, c'est-à-dire*

$$\Phi(-1,1) = P(X \geq 1,1) = 1 - P(X \leq 1,1) = 1 - 0,8643 = 0,1357.$$

par exemple $\Phi(-1.1) = 1 - \Phi(1.1)$.

Pour n'importe quel $a > 0$,

I	$\mathbb{P}(X \leq a)$		$\Rightarrow$ table
II	$\mathbb{P}(X \geq a)$		$\Rightarrow$ cas I
III	$\mathbb{P}(X \leq -a)$		$\Rightarrow$ cas II
IV	$\mathbb{P}(X \geq -a)$		$\Rightarrow$ cas I

Exemple 11 *On suppose qu'une certaine variable* $X \sim \mathcal{N}(11, 2)$. *Pour quelle proportion d'individus est-ce que* $X \leq 14$?

On cherche $P(X \leq 14)$.

- $P(X \leq 14) = P(\frac{X-11}{2} \leq \frac{14-11}{2}) = P(Z \leq 1.5)$.
- On cherche $1,5$ dans la table.

On trouve finalement $P(X \leq 14) = 0,9332$.

Chapitre 2

Échantillonnage

2.1 Introduction

La théorie de l'échantillonnage étudie les liens entre une population et des échantillons de cette population.

Dans cette partie, nous allons étudier comment se comporte un échantillon (éléments pris au hasard) dans une population dont on connaît les caractéristiques statistiques (lois,...) d'une variable considérée X. Dans ce cas, prendre un échantillon aléatoire de taille n consiste à considérer n réalisations de X ou encore considérer n variables aléatoires $X_1, ..., X_n$ indépendantes, de même loi que X.

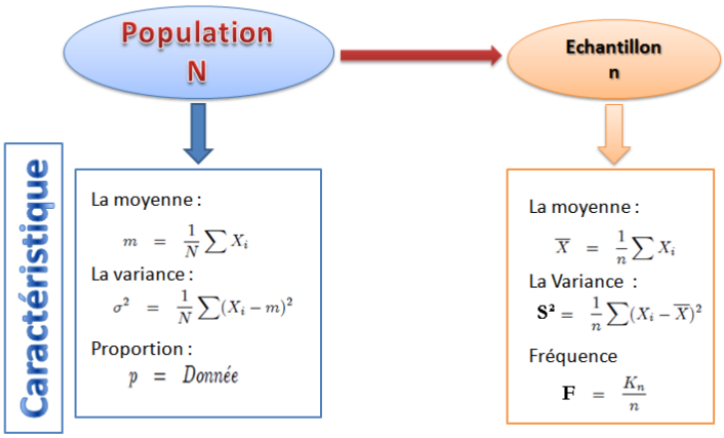
2.1.1 Avantages de l'échantillonnage

L'analyse d'un échantillon, par rapport à celle de la population, cout moindre, gain de temps et c'est la seule méthode qui donne des résultats dans le cas d'un test destructif.

2.2 Population - Echantillontion

2.2.1 Population

On appelle population la totalité des unités de n'importe quel genre prises en considération par le statisticien. Elle peut être finie ou infinie.



2.2.2 Echantillon

Soit $\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$ une population de taille N. Soit X le caractère que l'on voudrait étudier sur cette population. Avec l'échantillon aléatoire simple : soit X_k le résultat aléatoire du $k^{ième}$ tirage, c'est une variable aléatoire qui suit la même loi que X. On note x_k le résultat du $k^{ième}$ tirage et on note $(X_1, X_2, ..., X_n)$ le résultat aléatoire de ces n tirages.

Donc les n variables aléatoires indépendantes $X_1, X_2, ..., X_n$ constituent un échantillon aléatoire simple de la variable X si et seulement si

$$\begin{aligned} E(X_1) &= E(X_2) = ... = E(X_n) = E(X) = m. \\ \sigma(X_1) &= \sigma(X_2) = ... = \sigma(X_n) = \sigma(X) = \sigma^2_{pop}. \end{aligned}$$

Définition 12 *$(X_1, X_2, ..., X_n)$ sont n v.a. indépendantes et de même loi (celle de X),(Par exemple la loi de Gauss) il est appelé n-échantillon ou échantillon de taille n de X.*

Définition 13 *La réalisation unique $(x_1, ..., x_n)$ de l'échantillon $(X_1, X_2, ..., X_n)$ est l'ensemble des valeurs observées.*

Exemple 14 *On fait l'hypothsèe que la taille (en cm) des 4000 étudiants masculins d'une école de génie est une variable aléatoire X distribuée normalement, c'est-à-dire que $X \sim \mathcal{N}(\mu, \sigma)$. Un échantillon aléatoire de taille 50 de cette population est une suite de 50 variables aléatoires $X_i \sim \mathcal{N}(\mu, \sigma)$, $i = 1, 2, ..., 50$.*

Définition 15 (Définition d'une statistique) *Une statistique Y sur un échantillon $(X_1, X_2, ..., X_n)$ est une v.a., fonction mesurable des X_k; $Y = f(X_1, X_2, ..., X_n)$. Après réalisation, la v.a. Y (statistique) prend la valeur $f(x_1, ..., x_n)$.*

Exemples de statistiques :

1. La moyenne échantionnale $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. La variance échantionnale $S^2_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$.

Dans notre cours, nous allons travailler sur l'échantillonnage aléatoire simple, avec deux cas :

a) **Non Exaustif : Avec Remise** (car la taille de la population est grande).

b) **Exaustif** : **Sans Remise** : (car la taille de population est finie)

2.3 Les distributions d'échantillonnage.

Soit dans une population mère de taille N, une variable aléatoire X pour laquelle l'espérance mathématique m, la proportion P et l'écart-type σ_{pop} sont connues. De cette population sont issus k échantillons $E_1, E_2, ..., E_k$ de taille n qui auront des moyennes et des écarts-types différents. La notion de distribution d'échantillonnage peut être résumé et schématisée :

Population mère : Ω	**Echantillon** 1	**Echantillon** 2	.....	**Echantillon** k
Taille : N	Taille : n	Taille : n	.....	Taille : n
Moyenne : $m = \mu$ (connue)	Moyenne : $\bar{X}_1$	Moyenne : $\bar{X}_2$	.....	Moyenne : $\bar{X}_k$
Proportion : P (connue)	proportion : f_1	proportion : f_2	.....	proportion : f_k
Ecart-type : σ^2_{pop} (connue)	Ecart-type : σ_1	Ecart-type : σ_2	.....	Ecart-type : σ_k

2.4 Distribution échantillonnale de la moyenne $\bar{X}$

Dans une population mère de taille N, on peut tirer plusieurs échantillons de taille n :

Pour chaque échantillon, on peut calculer une moyenne :

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

et une variance

$$S_{\bar{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

La valeur de l'espérance mathématique $E\left(\bar{X}\right)$ et de la variance $S_{\bar{X}}^2$ varient d'un échantillon à l'autre.

C'est cette variation qui donne naissance à la distribution des variables aléatoires :

– **Echantillonnage de la moyenne ou moyenne d'échantillon** $\bar{X}$, caractérisée par :

$E(\bar{X})$: l'espérance mathématique des moyennes calculées sur tous les échantillons de taille n.

$S_{\bar{X}}$: l'écart type de la distribution d'échantillonnage, qui représente la dispersion de l'ensemble des moyennes d'échantillons de taille n autour de $E(\bar{X})$

– **Variance d'échantillon** $S_{\mathbf{c}}^2$ définie par

$$S_{\mathbf{c}}^2 = \frac{n}{n-1}S_{\bar{X}}^2.$$

On a bien entendu $E(S_{\mathbf{c}}^2) = \sigma_{pop}^2$.

Nous verrons plus tard que cela signifie que $S_{\mathbf{c}}^2$ (variance corrigée de l'échantillon) est un estimateur sans biais de σ^2

1) Cas : moyenne m et écart-type σ_{pop} de la population connus

A) Si la population est infinie ou si l'´echantillonnage est **non exhaustif** (tirage avec remise)

– L'espérance mathématique de la variable aléatiore $\bar{X}$ est égale à celle de la population mère :

$$E\left(\bar{X}\right) = \mu_{\bar{X}} = m$$

– Pour la variance $V(\bar{X})$

$$\begin{aligned} V(\bar{X}) &= \sigma_{\bar{X}}^2 = \frac{\sigma_{pop}^2}{n}. \\ \sigma_{\bar{X}} &= \frac{\sigma_{pop}}{\sqrt{n}} \end{aligned}$$

Alors dans ce cas $\bar{X} \sim \mathcal{N}\left(m, \frac{\sigma_{pop}}{\sqrt{n}}\right)$, ou bien $\bar{X}$ suit approximativement $\mathcal{N}\left(m, \frac{\sigma_{pop}}{\sqrt{n}}\right)$

(en pratique $n > 30$)

Exemple 16 *Une machine effectue l'ensachage d'un produit.*

On sait que les sacs ont un poids moyen de $250g$ *avec un écart-type de* $25g$.

Quelles sont les caractéristiques de la moyenne des poids d'un échantillon de 100 *sacs ?*

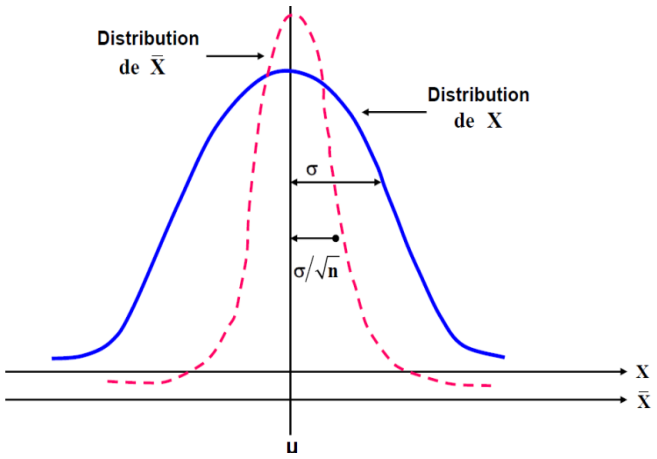
Solution 17 $(P) : m = \mu = 250$, $\sigma = 2,5$; $(E) : n = 100 > 30$

$\bar{X}$ *suit la loi normale de paramètres* $m = 250$ *et* $\dfrac{\sigma_{pop}}{\sqrt{n}} = \frac{25}{10} = 2,5$.

B) Si l'échantillonnage est **exhaustif** (tirage **sans remise**) dans une population finie : (Taille N sera donnée)

$$\begin{aligned} V(\bar{X}) &= \sigma^2_{\bar{X}} = \frac{\sigma^2_{pop}}{n}.\frac{N-n}{N-1} \\ \sigma_{\bar{X}} &= \frac{\sigma_{pop}}{\sqrt{n}}.\sqrt{\frac{N-n}{N-1}} \end{aligned}$$

Donc dans ce cas $\bar{X} \sim \mathcal{N}\left(m, \dfrac{\sigma_{pop}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$,



Remarque 18 $\sigma_{\bar{X}} = \dfrac{\sigma_{pop}}{\sqrt{n}}$ *est aussi appelé l'erreur-type de la moyenne.*

Remarque 19 *Si les échantillons sont issus d'une population mère finie et sont constituée sans remise. L'espérance mathématique de* $\bar{X}$ *est toujours égale à* m*, mais l'écart-type est corrigé par le facteur d'exhaustivité (facteur de correction)*

$\sigma_{\bar{X}} = \dfrac{\sigma_{pop}}{\sqrt{n}} \cdot \sqrt{\dfrac{N-n}{N-1}} \approx \dfrac{\sigma_{pop}}{\sqrt{n}} \cdot \sqrt{1 - \dfrac{n}{N}}$ *tel que* $\dfrac{n}{N}$ *représente le taux de sondage.*

Exemple 20 *Dans une usine textile, on utilise une machine automatique pour couper des morceaux de tissu. Lorsque la machine est correctement ajustée, la longueur des morceaux de tissu est en moyenne de* $90cm$ *avec un écart type de* 0.60 cm*.*

Pour contrôler la longueur des morceaux de tissu, on tire dans la production d'une journée un échantillon aléatoire de 200 *morceaux.*

a) Si l'on suppose que la longueur X *des morceaux de tissu suit une loi normale, calculer la probabilité que la moyenne de l'échantillon soit au plus égale à* 89.90 cm*, ceci dans* 2 *cas :*

—Production de la journée : 10000 *morceaux*

—Production de la journée : 2000 *morceaux.*

b) Déterminer la même probabilité sans faire l'hypothèse que X *soit distribuée normalement.*

Solution 21 *a) Production journalière* $= N = 10000$*, Taille de l'échantillon* $= n = 200$*,* $\dfrac{n}{N} = 0.02$

Même si l'échantillonnage est exhaustif, ce n'est pas la peine de tenir compte du coefficient d'exhaustivité.

Dans ce cas $E\left(\bar{X}\right) = 90$ *cm et* $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.6}{\sqrt{200}} = 0.042.$

Comme $X \sim \mathcal{N}(90, 0.6) \longrightarrow \bar{X} \sim \mathcal{N}((90, 0.042))$

$P(\bar{X} \leq 89.9) = P\left(T \leq \dfrac{89.9 - 90}{0.042}\right) = P(T \leq -2.38) = 1 - \Phi(2.38) = 0.0087 \longrightarrow 0.87\%$

Production journaliére $= N = 2000 \longrightarrow \dfrac{n}{N} = 0.1 \longrightarrow$ *on doit tenir compte du coefficient d'exhaustivité*

$s_{\bar{X}} = \dfrac{\sigma^2_{pop}}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}} = \dfrac{0.6}{\sqrt{200}}\sqrt{\frac{2000-200}{2000-1}} = 0.04$

$\bar{X} \sim \mathcal{N}(90, 0.04)$

$P(\bar{X} \leq 89.9) = P\left(T \leq \dfrac{89.9 - 90}{0.04}\right) = P(T \leq -2.5) = 1 - \Phi(2.5) = 0.0062 \longrightarrow 0.62\%$

Même si l'on ne fait plus l'hypothèse que X *soit une variable normale, comme* $n = 200 > 30$*, le théorème central limite permet de dire quel* $\bar{X} \backsim \mathcal{N}(90, 0.042)$ *pour* $N = 10000$*. On trouvera donc la même probabilité*

$P(\bar{X} \leq 89.9) = 0.0087 \longrightarrow 0.87\%.$

Définition 22 *On appelle variance empirique de l'échantillon* $(X_1, X_2, ..., X_n)$ *de* X *, la statistique*

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

Sa réalisation est $s^2_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (qui est la variance de l'échantillon), aussi appelée variance observée.

$$E\left(S^2_{\bar{X}}\right) = \frac{n-1}{n}\sigma^2_{pop}.$$

Calculons $E\left(S^2_{\bar{X}}\right)$

$$\begin{aligned}
E\left(S^2_{\bar{X}}\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) = \frac{1}{n}\sum E\left(X_i^2\right) - E\left(\bar{X}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[V(X_i) + (E(X_i))^2\right] - \left[V(\bar{X}) + \left(E(\bar{X})\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[V(X) + (E(X))^2\right] - \frac{1}{n}\sigma^2_{pop} - m^2 \\
&= V(X) + (E(X))^2 - \frac{1}{n}\sigma^2_{pop} - m^2 = \sigma^2 + m^2 - \frac{1}{n}\sigma^2_{pop} - m^2 \\
&= \left(1 - \frac{1}{n}\right)\sigma^2_{pop} = \frac{n-1}{n}\sigma^2_{pop}
\end{aligned}$$

Conclusion 23 *La moyenne des variances d'échantillon n'est pas la variance de la population, mais la variance de la population multipliée par* $\frac{n-1}{n}$. *Bien sûr, si n est très grand, ces deux nombres seront très proches l'un de l'autre.*

2) Cas écart-type σ de la population inconnu

A) Cas des grands échantillon $(n \geq 30)$

-Si la variance est **inconnue**, un grand échantillon permet de déduire une valeur fiable pour σ^2_{pop} en calculant la variance de l'échantillon S^2 et en posant

$$\sigma^2_{pop} = \frac{n}{n-1} S^2_{\bar{X}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

$$\text{Donc si } \begin{cases} n \geq 30 \\ \sigma_{pop} \text{ inconnue} \end{cases}, \quad \bar{X} \text{ suit } \mathcal{N}\left(m, \sqrt{\frac{n}{n-1}} S\right)$$

B) Cas des petits échantillons **:** $(n < 30)$

On considère exclusivement le cas où X suit une loi normale dans la population.

$$\text{Si : } \begin{cases} n < 30 \\ \sigma_{pop} \text{ inconnue} \end{cases}, \; T = \frac{\bar{X} - m}{\frac{S}{\sqrt{n-1}}} \text{ suit une loi de Student à } n-1 \text{ degrés de liberté, notée } T_{n-1}.$$

Exercice 24 *Le responsable d'une entreprise a accumulé depuis des années les résultats à un test d'aptitude à effectuer un certain travail. Il semble plausible de supposer que les résultats au test d'aptitude sont distribués suivant une loi normale de moyenne* $m = 150$ *et de variance* $\sigma^2 = 100$ *: On fait passer le test à* 25 *individus de l'entreprise. Quelle est la probabilité que la moyenne de l'échantillon soit entre* 146 *et* 154?

Solution 25 *On considére la variable aléatoire* $\bar{X}$ *moyenne d'échantillon pour les échantillons de taille* $n = 25$ *: On cherche à déterminer* $P(146 < \bar{X} < 154)$

Nous sommes en présence d'un petit échantillon $(n < 30)$ *et heureusement dans le cas où la variable X suit une loi normale. De plus,* σ_{pop} *est connu. Donc* $\bar{X}$ *suit* $\mathcal{N}\left(m, \frac{\sigma_{pop}}{\sqrt{n}}\right) = \mathcal{N}\left(150, \frac{10}{5}\right)$. *On en déduit que* $Z = \frac{\bar{X} - 150}{2}$ *suit* $\mathcal{N}(0,1)$.

La table donne

$$\begin{aligned} P(146 &< \bar{X} < 154) = P\left(\frac{146-150}{2} < Z < \frac{154-150}{2}\right) = P\left(-2 < Z < 2\right). \\ &= 2P\left(0 < Z < 2\right) = 2 \times \left(P\left(Z < 2\right) - P\left(Z < 0\right)\right) = 2 \times \left(0,9772 - 0,5\right). \\ &= 2 \times 0,4772 = 0,9544. \end{aligned}$$

2.4.1 Distribution de la variance d'échantillon $S^2_{\bar{X}}$

Supposons que X suit une loi normale.

On considère la variable

$$Y = \frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} = \frac{n}{\sigma^2_{pop}} \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma^2_{pop}}\right)^2.$$

D'aprés la décomposition de $S^2_{\bar{X}}$:

$$\sum_{i=1}^{n} \left(X_i - m\right)^2 = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 + n\left(\bar{X} - m\right)^2.$$

En divisant par σ^2 :

$$\begin{aligned} \sum_{i=1}^{n} \left(\frac{X_i - m}{\sigma^2_{pop}}\right)^2 &= \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma^2_{pop}}\right)^2 + \frac{n}{\sigma^2_{pop}} \left(\bar{X} - m\right)^2. \\ &= \frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} + \left(\frac{\bar{X} - m}{\frac{\sigma_{pop}}{\sqrt{n}}}\right)^2. \end{aligned}$$

On a : $\dfrac{X_i - m}{\sigma_{pop}} \sim \mathcal{N}\left(0,1\right) \Rightarrow \left(\dfrac{X_i - m}{\sigma_{pop}}\right)^2 \sim \chi^2_1$ (Khi-deux 1 degret de libérté).

D'ou $\displaystyle\sum_{i=1}^{n} \left(\frac{X_i - m}{\sigma_{pop}}\right)^2 \sim \chi^2_n$ (Comme somme de n carrés de variables aléatoires indépendantes normales centrées réduite)

$$\frac{\bar{X} - m}{\frac{\sigma_{pop}}{\sqrt{n}}} \sim \mathcal{N}\left(0,1\right) \Rightarrow \left(\frac{\bar{X} - m}{\frac{\sigma_{pop}}{\sqrt{n}}}\right)^2 \sim \chi^2_1.$$

D'où, on en déduit :

$$\begin{aligned} Y &= \frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} \sim \chi^2_{n-1} \\ \text{c'est à dire } S^2_{\bar{X}} &\sim \frac{\sigma^2_{pop}}{n}\chi^2_{n-1} \quad \chi^2_{n-1} \text{ (Khi-deux } n-1 \deg \text{ret de libérté)} \end{aligned}$$

$$E(S^2_{\bar{X}}) = \frac{\sigma^2_{pop}}{n}(n-1) \text{ et } Var\left(S^2_{\bar{X}}\right) = 2(n-1)\frac{\sigma^4_{pop}}{n^2}.$$

Exercice 26 *On prélève* 25 *pièces dans une production industrielle. Une étude préalable a montré que le diamètre de ces pièces suivait une loi gaussienne de moyenne* $10mm$ *et d'écart-type* $2mm$*. Entre quelles valeurs a-t-on* 85% *de chances de trouver l'écart-type de ces pièces?*

Solution 27 *Pour commencer, il faut déterminer* α *et* β *t.q.*

$$0.85 = P(\alpha < \frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} < \beta) = P(\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} < \beta) - P(\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} < \alpha)$$

$$\begin{aligned} &= 1 - P(\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} > \beta) - [1 - P(\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} > \alpha)] \\ &= P(\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} > \alpha) - P(\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} > \beta). \end{aligned}$$

On sait que $\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} \sim \mathcal{X}^2_{25-1} = \mathcal{X}^2_{24}$ *et alors on cherche dans la table du* $\mathcal{X}^2$ *à* 24 *degrés de liberté les valeurs* α *et* β *comme suit :*

$$\begin{cases} P(\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} > \alpha) = 0.90 \\ P(\frac{nS^2_{\bar{X}}}{\sigma^2_{pop}} > \beta) = 0.05 \end{cases} \quad \textit{d'après la table de } \mathcal{X}^2$$

on trouve $\alpha = 15,659$ *et* $\beta = 36,415$

et alors :

$$P\left(15.659 < \frac{25S_{\bar{X}}^2}{2^2} < 36.415\right) = 0.85.$$
$$P(2.5054 < S^2 < 5.8264) = 0.85.$$
$$P(1.58 < S < 2.41) = 0.85.$$

2.4.2 Distribution d'échantillonnage d'une proportion F

Soit une population comportant deux modalités A et B. Soit p la proportion d'individus de la population possédant la modalité A. $1-p$ est donc la proportion des individus de la population possédant la modalité B.

On extrait de la population un échantillon de taille n. Soit X la v.a qui représente le nombre d'individus dans l'échantillon ayant la modalité A.

Définition 28 *La v.a.* $F = \dfrac{X}{n}$ *s'appelle fréquence empirique. Sa réalisation* f *est la proportion d'individus dans l'échantillon ayant la modalité* A*. Où* X *est le nombre de fois où le caractère apparaît dans le* $n-$*échantillon.*

Par définition X suit $\mathcal{B}(n,p)$. Donc $E(X) = np$ et $Var(X) = npq$.

i) Si la population est infinie ou si l'échantillonnage est non exhaustif (tirage avec remise), on montre que :

$$\begin{cases} E(F) = p \\ V(F) = S_F^2 = \dfrac{p\,(1-p)}{n},\ S_F = \sqrt{\dfrac{p\,(1-p)}{n}} \end{cases}$$

– **Loi de probabilité pour** F

Si n est grand $n \geq 30$ et $np \geq 15$, $nq \geq 15$, on peut approcher la loi binomiale par la loi normale de même espérance et de même écart-type. Donc F suit approximativement $\mathcal{N}\left(p, \sqrt{\dfrac{p\,(1-p)}{n}}\right)$, et la variable $T = \dfrac{F-p}{\sqrt{\dfrac{p\,(1-p)}{n}}}$, suit alors approximativement la loi $\mathcal{N}(0,1)$.

$$\text{Et on écrit } F \sim \mathcal{N}\left(p, \sqrt{\frac{p\,(1-p)}{n}}\right)$$

ii) Si l'échantillonnage est exhaustif (tirage sans remise) dans une population finie $(n > 0,05N)$

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{p\,(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

Exercice 29 *Selon une étude sur le comportement du consommateur,* 25% *d'entre eux sont influencés par la marque, lors de l'achat d'un bien. Si on interroge* 100 *consommateurs pris au hasard, quelle est la probabilité pour qu'au moins* 35 *d'entre eux se d´eclarent influencés par la marque ?*

Solution 30 *Appelons* F *la variable aléatoire :"proportion d'échantillon dans un échantillon de taille* 100*". Il s'agit ici de la proportion de consommateurs dans l'échantillon qui se d´eclarent influencés par la marque. On cherche à calculer* $P\,(F > 0.35)$*.*

Il nous faut donc déterminer la loi de F*. Or* $np = 100 \times 0.25 = 25$ *et* $nq = 100 \times 0.75 = 75$*. Ces deux quantités étant supérieures à* 15*, on peut considérer que* F *suit*

$$\mathcal{N}\left(p, \sqrt{\frac{p\,(1-p)}{n}}\right) = \mathcal{N}\,(0.25, 0.0433)\,.$$

On utilise la variable $T = \dfrac{F - 0.25}{0.0433}$ *qui suit la loi* $\mathcal{N}\,(0,1)$*. Il vient*

$$P\,(F > 0.35) = P\,(T > 2.31) = 0.5 - P\,(0 < T < 2.31) = 0.5 - 0.4896 = 0.0104.$$

Conclusion 31 *Il y a environ une chance sur* 100 *pour que plus de* 35 *consommateurs dans un* $100-$ *échantillon se disent influencés par la marque lorsque l'ensemble de la population contient* 25% *de tels consommateurs.*

Exercice 32 *Le directeur financier d'une société sait par expérience que* 12% *des factures émises ne sont pas réglées dans les* 10 *jours ouvrables suivant l'échéance. Il fait prélever un échantillon aléatoire de* 500 *factures.*

Exemple 33 *Quelle est la probabilité qu'au moins* 70 *factures ne sont pas réglées dans le délais, sachant quel'ensemble des factures pouvant être étudiées est de plusieurs dizaines de milliers*

Solution 34 *Soit* $F =$ *"proportion d'échantillon dans un échantillon de taille* 500*".*

$P\left(F \geq \dfrac{500}{70}\right)$ =?- *Distribution d'échantillonnage d'une proportion* F*, échantillonnage exhaustif (tirage sans remise) dans une population finie,*

mais $n < 0,05N$*, donc il ne faut pas tenir compte du facteur d'exhaustivité.*

Ici $p = 0,12$, $q = 1 - p = 1 - 0,12 = 0,88$.

Comme $n = 500 > 30$, $np = 500 \times 0;12 = 60 > 15$, $nq = 500 \times 0,88 = 440 > 15 \Rightarrow$ *approximation de la loi binomiale par la loi normale*

$$1)\ F \sim \mathcal{N}\left(p, \sqrt{\frac{p\,(1-p)}{n}}\right) = \mathcal{N}\left(0,12, \sqrt{\frac{0,12 \times 0,88}{500}}\right) = \mathcal{N}\,(0,12, 0,015)$$

$$\begin{aligned} 2)\ P\left(F \geq \frac{500}{70}\right) &= P\left(Z > \frac{0,139 - 0,12}{0,015}\right) \\ &= 1 - P\left(Z < \frac{0,019}{0,015}\right) = 1 - P(Z < 1,27) = 1 - \Phi(1,27) = 1 - 0,8997 \approx 0,1 \\ &\approx 10\% \text{ de chances pour que plus de } 70 \text{ factures dans un } 500 \text{ échantillon soient non réglées dans le délais.} \end{aligned}$$

Chapitre 3

Estimation de Paramètres

3.1 Introduction

L'estimation consiste à donner des valeurs approchées aux paramètres d'une population (p, μ, σ^2) ou (proportion, moyenne, variance) à partir des données de l'échantillon (f, $\bar{x}$, s^2). Dans les problémes d'estimation, on cherche à se faire une idée de la valeur d'un paramètre inconnu de la population mère à partir de données observées dans un échantillon - induction du particulier au général.

L'objectif est d'obtenir une bonne estimation de $m = \mu$, p et σ à partir de $\bar{x}$, f et s, compte tenu de l'existence d'une dispersion dans la distribution d'échantillonnage.

On supposera vérifiée l'hypothèse d'échantillonnage aléatoire simple. La statistique inférentielle peut se résumer par le schéma suivant :

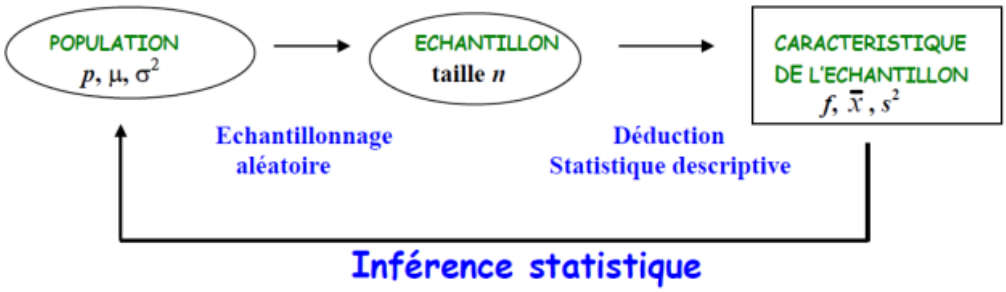


FIG. 3-1 – Statistiques Inférentielle

Dans cette section nous allons définir un estimateur et nous allons étudier ses propriétés statistiques. Les méthodes utilisées pour déterminer un estimateur d'un parametre seront présentées dans le chapitre suivant

Les méthodes d'estimation se divisent en 2 grandes catégories :

1. |L'estimation **ponctuelle** : on estime la valeur du paramètre inconnu de la population mère par un seul nombre à partir de l'information fournie par l'échantillon.
2. |L'estimation **par intervalle de confiance** : on estime un paramètre d'une population donnée par deux nombres qui forment un intervalle à l'intérieur du quel le paramètre de la population a de grandes chances de se trouver.

Les estimations par intervalles indiquent la précision d'une estimation et sont donc préférables aux estimations ponctuelles.

3.2 Définition d'un estimateur

Soit $(X_1, ..., X_n)$ un échantillon aléatoire indépendant et identiquement distribué de même loi que X, $X \longrightarrow L(\theta)$ ou la loi L est supposé connu et le paramètre θ est inconnu

Définition 35 *Un estimateur de θ, noté $\hat{\theta}$, est une statistique $\hat{\theta} = T(X_1, X_2, ..., X_n)$.*

Toute valeur $T(x_1, x_2, ..., x_n)$ de cet estimateur $\hat{\theta}$ est appelée une estimation de θ

- Il importe de faire la distinction entre l'estimateur de θ (qui est une variable aléatoire réelle) et l'estimation de θ qui est une grandeur numérique.

- On désignera souvent un estimateur quelconque de θ par le symbole $\hat{\theta}$ (ou par $\hat{\theta}_n$ pour rappeler que la taille de l'échantillon est n)

Exemple 36 1) $\hat{\theta} = \bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ (*Estimateur de* μ), $\hat{\theta} \sim \mathcal{N}\left(\theta, \dfrac{\sigma}{\sqrt{n}}\right)$

2) $\hat{\mu} = \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ (*Estimation*)

3.3 Propriétés d'un estimateur

3.3.1 Estimateur sans biais :

Soit $\hat{\theta}$ un estimateur de θ.

Définition 37 *On dit que* $\hat{\theta}$ *est un estimateur sans biais du paramètre* θ *si* $E[\hat{\theta}] = \theta$

- $E[\hat{\theta}]$: esperance mathematique de. $\hat{\theta}$
- Un estimateur de θ est dit asymptotiquement sans biais si $\lim\limits_{n\longrightarrow\infty} E[\hat{\theta}] = \theta$

3.3.2 Estimateur convergent :

Théorème 38 *On dit que* $\hat{\theta}$ *est un estimateur convergent de ssi*

$$\begin{cases} \lim\limits_{n\longrightarrow\infty} E[\hat{\theta}] = \theta \\ \lim\limits_{n\longrightarrow\infty} V[\hat{\theta}] = 0 \end{cases}$$

Définition 39 $\hat{\theta}$ *est dit biaisé si* $E[\hat{\theta}] \neq \theta$ *et* $b\left(\hat{\theta}\right) = E[\hat{\theta}] - \theta$ *s'appelle le biais de* $\hat{\theta}$.

Définition 40 *Soient* $\hat{\theta}_1$ *et* $\hat{\theta}_2$ *deux estimateurs sans biais de* θ. $\hat{\theta}_1$ *est dit plus efficace que* $\hat{\theta}_2$ *si* $V\left(\hat{\theta}_1\right) \leq V\left(\hat{\theta}_2\right)$

► **On utilise souvent** S_c^2 pour estimer σ^2.

Perte quadratique : C'est l'´ecart au carré entre le paramètre et son estimateur :

$$l\left(\theta,\hat{\theta}\right)=\left(\theta-\hat{\theta}\right)^2$$

Risque d'un estimateur :C'est la moyenne des pertes

$$R\left(\hat{\theta},\theta\right)=E\left[\left(\theta-\hat{\theta}\right)^2\right].$$

$R\left(\hat{\theta},\theta\right)=E\left[\left(\theta-\hat{\theta}\right)^2\right]$ est le risque quadratique moyen.

3.4 Estimation ponctuelle des paramètres usuels

On souhaite estimer un paramètre θ d'une population (cela peut être sa moyenne μ, son écart-type σ, une proportion p).

3.4.1 Estimation ponctuelle de la moyenne de la population

Soit $(X_1, X_2, ..., X_n)$ indépendantes et identiquement distribuées $(i.i.d.)$ n observations de $X \sim \mathcal{N}(\mu,\sigma)$ ou grand échantillon $(n \geq 30)$. $\forall i = \overline{1,n}$ $E(X_i)=\mu$, $V(X_i)=\sigma^2$.

Théorème 41 *La variable aléatoire $\bar{X}$ définie par*

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

est un estimateur convergent et sans biais de μ

Preuve. 1) Estimateur sans biais

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

La moyenne empirique calculée sur un échantillon est une bonne estimation de la moyenne dans la population.

2) Estimateur convergent en probabilité :

-Cas : population infinie ou tirage non exhaustif :

$$\begin{aligned} V\left(\bar{X}\right) &= S_{\bar{X}}^2 = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \underset{n\longrightarrow\infty}{\frac{\sigma^2}{n}} \longrightarrow 0 \\ &\Rightarrow V\left(\bar{X}\right) = \underset{n\longrightarrow\infty}{S_{\bar{X}}^2} \longrightarrow 0 \end{aligned}$$

-Cas : population finie et tirage exhaustif (sans remise) :

$$V\left(\bar{X}\right) = S_{\bar{X}}^2 = \frac{\sigma^2}{n}\frac{N-n}{N-1}, \frac{N-n}{N-1} \approx 1 - \frac{n}{N} \approx 1,\ V\left(\bar{X}\right) = \underset{n\longrightarrow\infty}{S_{\bar{X}}^2} \longrightarrow 0$$

■

3.4.2 Estimation ponctuelle de la variance et de l'´ecart-type de la population

• Cas : μ connue

Soient $X_1, X_2, ..., X_n$ n v.a indépendantes de même loi de moyenne μ et de variance σ^2.

Théorème 42 *La variable aléatoire S_{ech}^2 d´efinie par*

$$S_{ech}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$$

est un estimateur convergent et sans biais de σ^2 uniquement si μ est connue.

Preuve. En effet

1- S_{ech}^2 est sans biais ;

$$\begin{aligned} E\left(S_{ech}^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\frac{1}{n}\sum_{i=1}^{n}\mu X_i + \mu^2\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n}X_i^2\right) - 2\frac{\mu}{n}\sum_{i=1}^{n}E\left(X_i\right) + \mu^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}E\left(X_i^2\right)\right) - \mu^2 \end{aligned}$$

Or par définition :

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

Donc

$$E\left(X^2\right) = \sigma^2 + \mu^2$$

Finalement :

$$E\left(S_{ech}^2\right) == \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - \mu^2 = \sigma^2$$

Donc si μ est connue alors S_{ech}^2 est un estimateur sans biais de σ^2.

2-Estimateur convergent

$$\begin{aligned} V\left(S_{ech}^2\right) &= V\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n^2}\sum_{i=1}^{n}V\left[(X_i-\mu)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left[E\left((X_i-\mu)^4\right) - \left[E(X_i-\mu)^2\right]^2\right] \\ &= \frac{1}{n}\left[E\left((X-\mu)^4\right) - \left[E(X-\mu)^2\right]^2\right] \longrightarrow 0 \text{ lorsque } n \longrightarrow \infty. \end{aligned}$$

■

Cas : μ inconnue

En général on ne connaît pas, μ on le remplace par son estimateur sans biais : $\bar{X}$ et on introduit la variance

empirique associée

$$S_{\bar{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 : \text{ la variance calculée sur l'échantillon.}$$

Théorème 43 *La variance empirique $S_{\bar{X}}^2$ est un estimateur biaisé et convergent de σ^2. Il est asymptotiquement sans biais. Avec*

$$E\left(S_{\bar{X}}^2\right) = \left(\frac{n-1}{n}\right)\sigma^2.$$

Théorème 44 *La variance empirique corrigée :*

$$S_{\mathbf{c}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \left(\frac{n}{n-1}\right) S_{\bar{X}}^2.$$

est un estimateur sans biais et convergent de la variance de la population σ^2

- Si n grand, $E\left(S_{\bar{X}}^2\right) \approx E\left(S_{\mathbf{c}}^2\right)$ et on préfère

$$S_{\bar{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

- Si n petit, on préfére

$$S_{\mathbf{c}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

Exemple 45 *Les prix d'un article en* 5 *différents marchés d'une région donnée sont*

i	1	2	3	4	5
x_i	75	82	83	78	80

Calculer les estimations ponctuelles de la moyenne et de l'´ecart-type

Solution 46 *L'effectif* $n = 5$ *de l'échantillon est inférieur à* 30 *et la moyenne* μ *et la variance* σ^2 *de la population sont inconnus. On utilise les expressions d'estimation ponctuelle les suivantes :*

Moyenne : $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{3985}{5} = 79,6$

Ecart-type : $\hat{\sigma}_{pop} = \sqrt{\frac{n}{n-1}} S_{\bar{X}} = \sqrt{\frac{\sum_{i=1}^{5} x_i^2 - 5\bar{x}^2}{4}}$

On ajoute encore une ligne à la table :

i	1	2	3	4	5	*Total*
x_i	75	82	83	78	80	398
x_i^2	5625	6724	6889	6084	6400	31722

$\Rightarrow \hat{\sigma}_{pop} = \sqrt{\frac{31722 - 54 \cdot 6336,16}{4}} = 3.21$

3.4.3 Estimation ponctuelle de la proportion de la population :

Soit une population comportant deux modalité A et B. Soit p la proportion d'individus possédant la modalité A. $1-p$ est donc la proportion des individus possédant la modalité B.

On extrait de la population un échantillon de taille n. Soit la variable aléatoire Y : nombre d'individus dans l'échantillon ayant la modalité A

Exemple 47 *Par exemple, à l'issue d'une chaîne de fabrication, un article est défectueux avec la probabilité p, non défectueux avec la probabilité $1-p=q$. Pour chaque article fabriqué, on peut définir la variable aléatoire X : lorsque l'article est défectueux, X prend la valeur 1 et $P(\{X=1\})=p$, lorsque l'article n'est pas défectueux, X prend la valeur 0 et $P(\{X=0\})=q$. X suit une loi de Bernoulli de paramètre p. Considérons un n-échantillon aléatoire simple de cette variable X soit $X_1, X_2, ..., X_n$ de réalisation $x_1, x_2, ..., x_n$. Ces n variables aléatoires indépendantes suivent toutes la même loi, celle de X, c'est-à-dire $\mathcal{B}(p)$. Leur somme $Y = X_1 + X_2 + ... + X_n$ suit une loi binomiale de paramètres n et p,*

$$Y \sim \mathcal{B}(n,p) \text{ et } P(\{Y=k\}) = C_n^k p^k (1-p)^{n-k},$$

pour k variant de 0 à n

Définition 48 *La variable aléatoire $F = \dfrac{Y}{n} = \dfrac{X_1 + ... + X_n}{n}$ s'appelle fréquence empirique, sa réalisation f est la proportion d'individus dans l'´echantillon ayant la modalité A.*

Cette variable fréquence est une variable aléatoire dont l'univers image est $\left\{0, \frac{1}{n}, \frac{2}{n}, ..., 1\right\}$ dont on connaît la distribution de probabilité : $P(\{Y=k\}) = C_n^k p^k q^{n-k}$, $q = 1-p$

La réalisation de cette variable fréquence est $f = x_1 + ... + x_n$, c'est-à-dire la fréquence de l'échantillon ou encore fréquence (ou pourcentage) des articles défectueux dans l'échantillon.

−Espérance de la variable fréquence F : on sait que $F = \dfrac{Y}{n}$ et $E(Y) = E(X_1) + ... + E(X_n) = np$ donc

$$E(F) = p$$

−Variance de la variable F : on a $V(Y) = npq$ donc

$$V(F) = S_F^2 = \frac{1}{n^2} V(Y) = \frac{p(1-p)}{n}$$

Donc F est un estimateur sans biais et convergent de p.

1. **Pour toute $n \geq 30$ la distribution des fréquence suit approximativement la loi normale $\mathcal{N}\left(p, \sqrt{\dfrac{p\,(1-p)}{n}}\right)$.**
2. **La proportion f observée sur l'échantillon est une estimation ponctuelle de la proportion p de la population**

3.4.4 Déterminer s_F , lorsque la proportion p de la population mère n'est pas connue.

▶ L'idée donc est de remplacer p par son estimateur sans biais f ,on obtient :

$$\begin{cases} v\,(F) = s_F^2 = \dfrac{f\,(1-f)}{n} \text{ tirage avec remise} \\ v\,(F) = s_F^2 = \dfrac{f\,(1-f)}{n}\dfrac{N-n}{N-1} \text{ tirage sans remise } (n \geq 0.05N) \end{cases}$$

Exercice 49 *Dans une population d'´etudiants AES, on a prélevé ind´ependamment* 2 *échantillons de taille* $n_1 = 120$, $n_2 = 150$. *On constate que* 48 *étudiants du* 1*-er échantillon et* 66 *du* 2*-éme ont une formation scientifique secondaire. Soit* p *la proportion d'´etudiants ayant suivi une formation scientifique. Calculer* 3 *estimations ponctuelles de* p.

Solution 50 $F = \dfrac{Y}{n}$, $f_1 = \dfrac{48}{120} = 0.4$, $f_2 = \dfrac{66}{150} = 044$, $f_3 = \dfrac{48+66}{120+150} = 0.422$

3.5 Méthode d'estimation par intervalle de confiance

Dans le cadre de l'estimation ponctuelle, on associe un nombre, une estimation à un paramètre dont la valeur est inconnue. La précision de cette estimation peut être déterminée en calculant un intervalle de confiance pour ce paramètre, c'est-à-dire un intervalle contenant la valeur inconnue du paramètre avec une grande probabilité donnée.

Soit X une variable aléatoire dont la loi dépend d'un paramètre θ inconnu. Soit $(X_1, X_2, ..., X_n)$ un n-échantillon issu de X et $\alpha \in]0, 1[$.

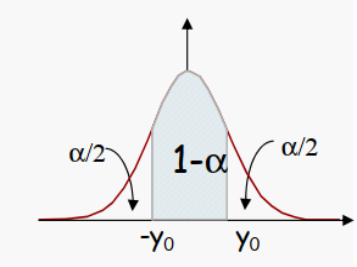
Définition 51 *On dit que* $[C_1, C_2]$ *est un* ***intervalle de confiance****, de niveau* $1-\alpha$*, (ou de*

seuil α) du paramètre θ si on a

$$P\left(\{C_1 \leq \theta \leq C_2\}\right) = 1 - \alpha$$

Les bornes de l'intervalle C_1 et , C_2 sont les statistiques basées sur l'échantillon aléatoire. À priori C_1 et C_2 sont des variables aléatoires, une fois les réalisations de l'échantillon obtenues, on dispose des valeurs numériques $x_1, x_2, ..., x_n$. On remplace C_1 et C_2 par leurs réalisations et on obtient les bornes de l'intervalle recherché. Cet intervalle est une réalisation de l'intervalle de confiance $[C_1, C_2]$.

- 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



Remarque 52 1) *Les niveaux de confiance les plus fréquemment utilisés sont* $90\%, 95\%, 99\%$

2) α *est appelé le seuil (le risque), on choisira dans la plupart des cas un risques symétriques, c'est-à-dire tels que*

$$P(\{\theta < C_1\}) = P(\{\theta > C_2\}) = \frac{\alpha}{2}.$$

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

3.5.1 Intervalle de confiance pour une moyenne

Supposons qu'une variable aléatoire X suit la loi normale de moyenne $m = E(X)$ inconnue et de **variance connue** σ^2 **:** $X \hookrightarrow \mathcal{N}(m, \sigma^2)$ m est donc le paramètre inconnu que l'on cherche à estimer.

On cherche à estimer m et l'encadrer entre deux valeurs à un certain niveau de confiance $1-\alpha$ tel que :

$$P(\{C_1 \leq m \leq C_2\}) = 1 - \alpha$$

Soient $X_1, X_2, ..., X_n$, n variables aléatoires de lois normales et indépendantes.

La statistique utilisée pour construire cet intervalle de confiance et pour trouver les valeurs C_1 et C_2 est

$$Z = \frac{\bar{X} - E(\bar{X})}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}.$$

Avec $\bar{X} = \dfrac{X_1 + ... + X_n}{n}$ Son choix est justifié par le fait que $\bar{X}$ est le bon estimateur

ponctuel de m (estimateur sans biais, convergent et efficace) et que

$$\bar{X} \sim \mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right) \text{ et } Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

On a donc :

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Updownarrow$$

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Avec $z_{\frac{\alpha}{2}}$ est le fractile de la loi $\mathcal{N}(0,1)$ d'ordre $1-\alpha$.

On obtient alors le théorème suivant :

Théorème 53 (IC pour l'espérance d'un échantillon gaussien, σ^2 connue) *Si $X_1, X_2 \ldots\ldots X_n$ est un échantillon gaussien d'espérance inconnue $m = \mu$ et de variance connue $\sigma^2 > 0$, alors*

$$\left[\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right],$$

est un intervalle de confiance de niveau $1-\alpha$ où la valeur $z_{\frac{\alpha}{2}}$ est lue dans la table normale centrée réduite $\mathcal{N}(0,1)$ telle que $\Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$

Exemple 54 *Soit $n = 100$, $\sigma = 2.5$ et $\bar{x} = 11.5$*

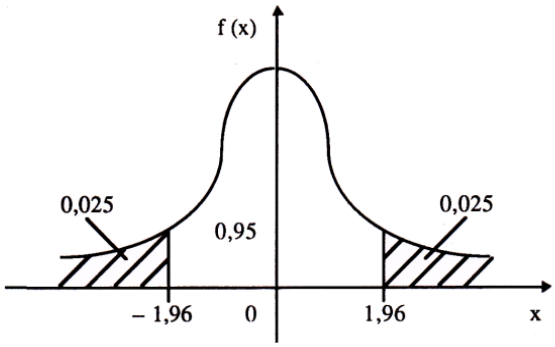
Donner un intervalle de confiance de niveau 0.95 pour m

Ici, $\alpha = 0.05$ et $1 - \frac{\alpha}{2} = 0.975$. Le quantile d'ordre 0.975 de la loi $N(0,1)$

est $z_{\frac{\alpha}{2}} = 1,96$. *L'intervalle de confiance est :*

$$\left[11.5 - 1.96\frac{2.5}{100}, 11.5 + 1.96\frac{2.5}{100}\right]$$

$$\textit{donc } m \in [11.01, 11.99]$$



– **Cas : σ^2 inconnue**

Pour des échantillons de taille $n < 30$ extraits d'une population suivant une loi normale d'écart-type inconnu, on utilise la distribution t de Student pour déterminer l'intervalle de confiance de la moyenne.

Lorsque la variance σ^2 est inconnue on doit d'abord estimer la moyenne m pour estimer σ^2

– **Estimateur sans biais** : $S_{\mathbf{c}}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

– **Estimation** : $s_{\mathbf{c}}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

Dans ce cas, la distribution d'échantillonnage de $\bar{X}$ a pour moyenne $E\left(\bar{X}\right) = m$ et de variance estimée

$$Var\left(\bar{X}\right) = \frac{S_{\mathbf{c}}^2}{n}.$$

La statistique de test : $T = \dfrac{\bar{X}-m}{\frac{S_c}{\sqrt{n}}} \sim T_{(n-1)}$ Student.

On peut alors écrire

$$P\left(-t_{\frac{\alpha}{2}} \leq \frac{\bar{X} - m}{\frac{S_c}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

On trouve l'intervalle de confiance de niveau $1-\alpha$ pour μ :

$$\left[\bar{X} - t_{\frac{\alpha}{2}} \frac{S_c}{\sqrt{n}},\ \bar{X} + t_{\frac{\alpha}{2}} \frac{S_c}{\sqrt{n}}\right]$$

Avec $t_{\frac{\alpha}{2}}$ est le fractile de la loi de student à $n-1$ *ddl* d'ordre $1-\frac{\alpha}{2}$

Alors nous avons le théorème suivant :

Théorème 55 (Espérance d'un échantillon gaussien, variance inconnue) *Si $X_1, X_2, ..., X_n$ est un échantillon gaussien d'espérance inconnue m et de variance inconnue $\sigma^2 > 0$, alors*

$$\left[\bar{X} - t_{\frac{\alpha}{2}} \frac{S_c}{\sqrt{n}},\ \bar{X} + t_{\frac{\alpha}{2}} \frac{S_c}{\sqrt{n}}\right]$$

est un intervalle de confiance au niveau $1-\alpha$ de m où S_c est un estimateur de σ et la valeur $t_{\frac{\alpha}{2}}$ est lue dans la table de Student à $k = n-1$ dégrés de liberté (ddl) $1-\alpha$

Exemple 56 *Un examen de probabilité est organisé pour promotion très nombreuse on extrait un échantillon de 4 notes*

12.5, 10 , 14.5, 14.

Déterminer l'intervalle de confiance à 95% pour la moyenne de tout la promotion

Solution 57 *$n = 4 < 30$, en utilisant la distribution t de student on a :*

$$m \in \left[\bar{x} - t_{\frac{0.05}{2}} \frac{S_c}{\sqrt{n}}, \bar{x} + t_{\frac{0.05}{2}} \frac{S_c}{\sqrt{n}}\right],$$

et niveau de confiance $1-\alpha = 0.95 \Rightarrow \alpha = 0.05$ et $ddl = k = 3$, on a

$$t_{\frac{0.05}{2}} = t_{0.025} = 3.182$$

et $\bar{x} = 12.75$, $S_c^2 = 4.08$ est une estimation de la valeur inconnue σ^2, donc

$$\begin{aligned} m &\in \left[12.753 - 3.182\frac{\sqrt{4.08}}{\sqrt{4}}, 12.753 + 3.182\frac{\sqrt{4.08}}{\sqrt{4}}\right]. \\ m &\in [9.535, 15.964] \end{aligned}$$

Exemple 58 *On suppose que le taux de cholestérol X d'un individu choisi au hasard dans une population donnée suit une loi normale. Sur un échantillon de 20 individus, on constate la moyenne des taux observés est $\bar{x} = 1.55$(gr pour millr). On constate aussi une variance empirique $S_c^2 = 0.25$.*

Donner un intervalle de confiance pour la moyenne m au niveau de confiance 0.95?

Remarque 59 *Lorsque la population est distribuée normalement, que σ n'est pas connu et que l'échantillon est de faible taille ($n < 30$), on se référe à la loi de Student Fisher.*

***Approximation** : si la taille de l'échantillon est grande ($n \geq 30$) alors on peut remplacer la valeur du fractile $t_{\frac{\alpha}{2}}$ de Student à $(n-1)$ ddl : par celle du fractile $z_{\frac{\alpha}{2}}$ de la loi normale centrée-réduite $\mathcal{N}(0,1)$.*

3.5.2 Intervalle de confiance pour une proportion p inconnue

Pour construire l'intervalle de confiance d'une proportion p (inconnue) des individus possédant un certain caractère appartenant à une population infinie (ou finie si le tirage s'effectue avec remise), on utilise f , proportion calculée dans un échantillon de taille n. Si n est faible on utilise les tables de la loi binomiale puisque : $nF \sim \mathcal{B}(n,p)$, ou on utilise l'abaque.

Si n est suffisamment grand généralement $n > 30$: $F \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

où $F = \frac{1}{n}\sum_{i=1}^{n} X_i$ avec $X_i \sim \mathcal{B}(p)$

et $X_i = \begin{cases} 1 \text{ si succés avec } p \\ 0 \text{ si echec avec } 1-p \end{cases}$

$E(F) = p$ et la variance $V(F) = \frac{p(1-p)}{n}$ estimée par son estimateur $\frac{f(1-f)}{n}$

– **La statistique de test** : $\frac{F-f}{\sqrt{\frac{f(1-f)}{n}}} \sim \mathcal{N}(0,1)$

– On peut alors écrire :

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{F-f}{\sqrt{\frac{f(1-f)}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

– On en déduit l'intervalle asymptotique de confiance de niveau $(1-\alpha)$ de p :

$$p \in \left[f - z_{\frac{\alpha}{2}} \sqrt{\frac{f\,(1-f)}{n}}, f + z_{\frac{\alpha}{2}} \sqrt{\frac{f\,(1-f)}{n}} \right].$$

Cette approximation de la loi Binomiale par la loi Normale n'est valable que lorsque
$n > 30$, $np > 5$, $nq > 5$

Remarque 60 *On peut élargir cet intervalle, en remarquant que* $f\,(1-f)$ *est maximal pour* $f = 1/2$ *et vaut* $1/4$. *On aura :*

$$\left[f - \frac{1}{2n} z_{\frac{\alpha}{2}}, f + \frac{1}{2n} z_{\frac{\alpha}{2}} \right]$$

Est un intervalle de confiance de p *au seuil de confiance* $1-\alpha$

(cet élargissement n'est acceptable que si $0,3 \leq f \leq 0,7$)

3.5.3 Marge d'erreur dans l'estimation de p

Cas de l'échantillon indépendant (non exhaustif)

La formule donnant la taille "n" minimum de l'échantillon est la suivante

$$n = \frac{z_{\frac{\alpha}{2}}^2 f\,(1-f)}{e^2},$$

Sa réciproque (**Marge d'erreur dans l'estimation de** p)

$$E = z_{\frac{\alpha}{2}} \sqrt{\frac{f\,(1-f)}{n}}$$

Remarque 61 E *est un pourcentage c'est la marge d'erreur qu'on se donne*

Exemple 62 *Les responsables d'une étude de marché ont choisi au hasard* 500 *femmes dans une grande ville et ont constaté que* 35% *des femmes retenues dans l'échantillon préférent utiliser*

une marque de lessive A plutôt que les autres. Ils veulent déterminer l'intervalle de confiance à 95% *de la proportion des femmes de cette ville qui préférent la marque de lessive A*

Solution 63 *$f = 0 : 35$) $s = \sqrt{\dfrac{0.35 \times 0.65}{500}} = 0.021331$:*

$$P(0.35 - 1.96 \times 0.02133 < p < 0.35 + 1.96 \times 0.02133) = 0.95$$

L'intervalle de confiance est donc $[0.3082, 0.3918]$. *Il y a donc entre* 30.82% *et* 39.18% *des femmes de cette ville qui préférent la marque de lessive A (avec un risque de* 5% *de se tromper).*

Exemple 64 *Les responsables d'une étude de marché ont choisi au hasard* 500 *femmes dans une grande ville et ont constaté que* 35% *des femmes retenues dans l'échantillon préférent utiliser une marque de lessive A plutôt que les autres.*

Supposons qu'avant de tirer l'échantillon, les responsables de l'étude aient d´ecidé d'estimer la proportion p à $\pm 2\%$ *près.*

Quelle devrait être dans ce cas la taille minimale de l'échantillon à tirer, en désirant toujours avoir un intervalle de confiance à 95% *et en considérant que* $f = 0.35$

Solution 65 *Pour avoir la proportion à* 2% *près, il faut que :*

$$\begin{aligned} & 1.96\sqrt{\frac{0.35 \times 0.65}{n}} = 0.02 \\ & \Rightarrow (1.96)^2 \frac{0.35 \times 0.65}{n} = (0.02)^2 \\ & \Rightarrow n = \frac{(1.96)^2 \times 0.35 \times 0.65}{(0.02)^2} = 2184.91 \simeq 2185 \end{aligned}$$