

**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

- Partie1 -

Deuxième année sciences économiques et gestion, semestre 3

**Année universitaire
2020-2021**

Le but de ce chapitre comme son nom l'indique est d'arriver à calculer des estimations des paramètres inconnus (moyenne, variance ou proportion d'une population). Par soucis de mettre le lien avec le cours de probabilité du semestre 2, le chapitre expose un rappel de quelques lois usuelles indispensables, de la convergence en probabilité ou en loi, de la loi faible des grands nombres et des approximations qui découlent du théorème central limite. Le troisième paragraphe est dédié aux **méthodes d'échantillonnages**.

Dans une deuxième partie, on va aborder la **distribution d'échantillonnage** puis l'**estimation** dite ponctuelle et on achèvera le chapitre par l'estimation par **intervalle de confiance**.

Sommaire

1. Rappel sur quelques lois usuelles	4
1.1. Loi binomiale	4
1.2. Loi de Poisson	6
1.3. Approximation d'une loi binomiale par une loi de Poisson	7
1.4. Loi exponentielle	8
1.5. Loi normale	8
4.1.1 Tabulations de la loi normale standard	9
4.1.2 Théorème d'addition de v.a. normales	11
1.6. Lois découlant de la loi normale	12
2. Convergence en probabilité et convergence en loi	12
2.1. Convergence en probabilité	13
2.2. Convergence en loi	14
2.3. Approximation par la loi normale	15
2.3.1. Cas de la loi binomiale	15
2.3.2. Cas de la loi de Poisson	16
2.3.3. Correction de continuité	17
3. Méthodes d'échantillonnage	18
3.1. Méthodes aléatoires	18
3.1.1. L'Echantillonnage aléatoire simple	18
3.1.2. L'échantillonnage stratifié	20
3.1.3. L'échantillonnage par grappes	21
3.1.4. L'échantillonnage à plusieurs degrés	22
3.1.5. L'échantillonnage à plusieurs phases	22
3.2. Méthodes empiriques ou non aléatoires	23
3.2.1. L'échantillonnage par quotas	23
3.2.2. l'échantillonnage à l'aveuglette ou de commodité	25
3.2.3. L'échantillonnage de volontaires	25
3.2.4. L'échantillonnage au jugé	25
3.3. Les erreurs	26

1. Rappel sur quelques lois usuelles

1.1. Loi binomiale

Etudier : Nombre de succès d'un événement A .

Exemple : répéter le jeu de dé n fois, on aura n épreuves de Bernoulli identiques et qui se déroulent indépendamment l'une de l'autre.

A : obtenir le résultat 4

X_i le résultat de la $i^{\text{ème}}$ épreuve, $X_i \in \{0, 1\}$

$p = P(A) = \frac{1}{6}$ si le dé n'est pas truqué

X la v.a. nombre de fois où on obtient le chiffre 4, elle est dite de loi binomiale

On écrit :

$$X = X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i$$

Définition : Soit $n \in \mathbb{N}^*$ et soit $p \in [0, 1]$, la v.a. X suit une loi binomiale de paramètres n et p si $X(\Omega) = \{0, 1, \dots, n\}$ avec

$$P_X(k) = P(X = k) = C_n^k p^k (1 - p)^{n-k} \text{ pour tout } k = 0, 1, \dots, n$$

On note $X \sim B(n, p)$

Exemple :

Pour l'exemple précédant, on a : $X \sim B\left(n, \frac{1}{6}\right)$. Pour tout $k = 0, 1, \dots, n$

$$P_X(k) = C_n^k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k} = C_n^k \frac{5^{n-k}}{6^n}$$

Si on jette le dé $n = 10$ fois, La probabilité d'obtenir 2 fois le résultat 4 est

$$P_X(2) = P(X = 2) = C_{10}^2 \frac{5^8}{6^{10}} = 0.2907 = 29.07\%$$

Propriétés : Si $X \sim B(n, p)$ alors on a les deux caractéristiques,

$$E(X) = np \text{ et } Var(X) = np(1 - p)$$

Exemple : Soit

$$X \sim B\left(10, \frac{1}{6}\right)$$
$$E(X) = \frac{10}{6} = \frac{5}{3} = 1.67 \text{ et } Var(X) = \frac{5}{3} * \frac{5}{6} = \frac{25}{18} = 1.39$$

Remarque : La loi binomiale modélise un très grand nombre de situations où il est envisagé le calcul de proportions :

- Soit p la proportion des clients qui ont l'intention d'acheter un produit
 p correspond à la probabilité qu'un client choisi au hasard ait l'intention d'acheter le produit.

Si X v.a. le nombre de clients qui ont l'intention d'acheter le produit parmi n clients, on associe à X la loi binomiale $B(n, p)$.

- Dans le département des achats d'une entreprise, on étudie X le nombre de bons de commande non-conforme parmi n bons de commande.

$$X \sim B(n, p)$$

p la proportion des bons de commandes non conformes dans tout l'ensemble des bons de commande du département des achats.

- Avant la tenue d'un référendum on s'intéresse à la v.a. X nombre de votants par OUI parmi n votants.

X suit une loi binomiale $B(n, p)$

p est la proportion réelle de votants par OUI dans toute la population des votants (ça correspond encore à la probabilité qu'un votant choisi au hasard vote par OUI).

Théorème d'addition de deux v.a. de lois binomiales

Soit X et Y deux v.a. indépendantes. Si $X \sim B(n_1, p)$ et $Y \sim B(n_2, p)$ alors

$$X + Y \sim B(n_1 + n_2, p)$$

Il s'agit de lois binomiales de même paramètre p qui caractérise la population dans laquelle on mène l'étude. C'est l'exemple de deux sondages effectués indépendamment dans la même population mais avec deux échantillons de tailles différentes.

Pour l'espérance et la variance de la somme, on trouve :

$$E(X + Y) = E(X) + E(Y) = (n_1 + n_2) p$$

$$Var(X + Y) = Var(X) + Var(Y) = (n_1 + n_2) p (1 - p)$$

1.2. Loi de Poisson

Cette loi s'intéresse au nombre d'apparitions (occurrences) d'un événement dans un intervalle de temps fixe.

Exemples :

- Nombre de clients qui se présentent devant un guichet chaque 20 minutes.
- Nombre d'appels téléphoniques dans une heure pour un SAV (Service Après Vente).
- Nombre de patients hospitalisés dans le service des soins intensifs chaque demi-heure.
- Nombre d'accidents sur une autoroute pendant une journée...

Soit le paramètre positif λ qui représente le **nombre moyen** d'apparitions de l'événement, (nombre moyen de clients, nombre moyen d'appels téléphoniques, nombre moyen de patients)

Définition : Soit $\lambda \in \mathbb{R}^{+*}$. La v.a. X suit une loi de Poisson de paramètre λ si $X(\Omega) = \mathbb{N}$ et pour tout $k \in \mathbb{N}$, on a :

$$P_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

On note $X \sim P(\lambda)$

Propriétés : Si $X \sim P(\lambda)$ alors on a :

$$E(X) = \lambda \text{ et } Var(X) = \lambda$$

Exemple :

X nombre de clients qui se présentent devant un guichet chaque 20 minutes, on connaît que 10 clients en moyenne se présentent devant le guichet. Alors on peut modéliser la loi de X par une loi de Poisson $P(10)$.

La probabilité que k clients se présentent est donnée par :

$$P_X(k) = e^{-10} \frac{10^k}{k!}$$

$$\begin{aligned} P(\text{Aucun client ne se présente devant le guichet}) &= P(X = 0) = e^{-10} \frac{10^0}{0!} = e^{-10} \\ &= 4.54 * 10^{-5} = 0.0000454 \end{aligned}$$

$$\begin{aligned} P(10 \text{ clients se présentent devant le guichet}) &= P(X = 10) = e^{-10} \frac{10^{10}}{10!} = 0.1251 \\ &= 12.51\% \end{aligned}$$

$$\begin{aligned}
P(\text{plus de 10 clients se présentent devant le guichet}) &= P(X \geq 10) = 1 - P(X < 10) \\
&= 1 - P(X \leq 9) \\
&= 1 - e^{-10} \left\{ 1 + \frac{10^1}{1!} + \frac{10^2}{2!} + \frac{10^3}{3!} + \frac{10^4}{4!} + \frac{10^5}{5!} + \frac{10^6}{6!} + \frac{10^7}{7!} + \frac{10^8}{8!} + \frac{10^9}{9!} \right\} \\
&= 1 - e^{-10} \{10086.57\} = 0.5421 = 54.21\%
\end{aligned}$$

Théorème d'addition de deux v.a. de lois de Poisson

Soit X et Y deux v.a. indépendantes. Si $X \sim P(\lambda_1)$ et $Y \sim P(\lambda_2)$ alors

$$X + Y \sim P(\lambda_1 + \lambda_2)$$

Et on a :

$$E(X + Y) = Var(X + Y) = \lambda_1 + \lambda_2$$

1.3. Approximation d'une loi binomiale par une loi de Poisson

Soit X la v.a. de loi binomiale $B(n, p)$, pour n assez grand et p proche de zéro, on peut approcher la loi de X par une loi de Poisson $\mathcal{P}(\lambda)$, avec $\lambda = np$.

En pratique on considère que cette approximation est bonne si on a les conditions suivantes,

$$\begin{cases} n \geq 30 \\ p \leq 0,1 \\ np < 15 \end{cases}$$

Exemple :

Soit $X \sim B(100; 0.06)$, on a $n = 100 \geq 30$; $p = 0.06 \leq 0,1$ et $np = 6 < 15$

On peut approcher la loi de X par une loi de Poisson $P(6)$.

k	0	3	6	10
Binomiale	0,002	0,086	0,166	0,040
Poisson	0,002	0,089	0,161	0,041

Remarque :

On peut toujours approcher la loi binomiale par la loi de Poisson pour d'autres conditions :

$$\begin{cases} n > 20 \\ p \leq 0,1 \\ np \leq 5 \end{cases}$$

1.4. Loi exponentielle

Elle modélise les phénomènes liés au temps, par exemple, temps d'attente devant une caissière dans un supermarché.

Le domaine du contrôle et de la fiabilité modélise, avec cette loi, les durées de vie des produits ou composants, par exemple, durée de vie d'une batterie ou tout autre composant électronique introduit dans un Smartphone ou un ordinateur.

Définition : Une v.a. X est dite de loi exponentielle de paramètre λ ($\lambda > 0$) si sa densité est

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

On note $X \sim \mathcal{E}(\lambda)$ ou $\text{Exp}(\lambda)$

Proposition : Soit $X \sim \text{Exp}(\lambda)$, on a :

$$E(X) = \frac{1}{\lambda} \quad \text{et} \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

1.5. Loi normale

Dite encore loi de Laplace-Gauss ou loi gaussienne, elle est considérée comme la plus connue et la plus utilisée des lois de probabilités, elle a des applications en économie, gestion, météorologie, médecine, physique, biologie, etc.

Le résultat le plus connu évoquant la loi normale est le théorème central limite qu'on va rappeler dans cette partie du chapitre.

Définition : Une v.a. X suit une loi normale de paramètre (μ, σ) (avec $\sigma > 0$) si sa densité est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \text{pour } x \in \mathbb{R}$$

On note $X \sim \mathcal{N}(\mu, \sigma)$

Proposition : Soit $X \sim \mathcal{N}(\mu, \sigma)$, on a :

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

Cas de la loi normale standard :

Une v.a. X est de loi normale standard si ses paramètres sont $\mu = 0$ et $\sigma = 1$, sa densité sera

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad \text{pour } x \in \mathbb{R}$$

et on a $X \sim \mathcal{N}(0, 1)$

En pratique, pour toute v.a. X de loi normale, $X \sim \mathcal{N}(\mu, \sigma)$, on associe la v.a. normale standard

$$Z = \frac{X - \mu}{\sigma}$$

Alors $Z \sim \mathcal{N}(0, 1)$.

4.1.1 Tabulations de la loi normale standard

La tabulation est disponible, elle est utilisée pour toute autre loi normale.

Soit $Z \sim \mathcal{N}(0, 1)$ et ϕ la fonction de répartition de Z :

$$\phi(z) = P(Z \leq z) = \alpha$$

On note z_α le nombre qui vérifie $\phi(z) = \alpha$ et on le nommera **quantile d'ordre α** .

Lecture de la table :

Si on connaît $z \in \mathbb{R}$, on cherche la probabilité $\alpha \in [0, 1]$

La tabulation est faite pour $z \geq 0$. Si z est négatif, on utilise la formule $\phi(z) = 1 - \phi(-z)$

Si on connaît la probabilité α , on peut déterminer le nombre z qui lui correspond.

On **résume** :

Si $z_\alpha \geq 0$ on détermine le α qui lui correspond.

Si $z_\alpha < 0$ on détermine le β qui correspond à $-z_\alpha$ et on pose $\alpha = 1 - \beta$.

Si $\alpha \geq 0.5$ on détermine z_α qui lui correspond.

Si $\alpha < 0.5$ on détermine $z_{1-\alpha}$ puis on pose $z_\alpha = -z_{1-\alpha}$.

Exemples :

Cherchons les probabilités associées aux quantiles suivants $-1.56, 2.22$.

$$-1.56 = z_{1-0.9406} = z_{0.0594} \quad (\text{lecture sur la table : } 1.56 = z_{0.9406}).$$

$$2.22 = z_{0.9868} \quad (\text{lecture direct sur la table}).$$

Cherchons les quantiles suivants $z_{0.01}$ et $z_{0.95}$

Directement sur la table de loi normale $\mathcal{N}(0, 1)$, on a :

$$z_{0.95} \approx z_{0.9505} = 1.65$$

Indirectement mais en utilisant la table de la loi, on a :

$$z_{0.01} = -z_{1-0.01} = -z_{0.99} \approx -z_{0.9901} = -2.33$$

Soit la v.a. $X \sim \mathcal{N}(5, 2)$, la v.a.

$$Z = \frac{X - 5}{2} \sim \mathcal{N}(0, 1)$$

Calculons les probabilités :

$$\begin{aligned} P(X < 4) &= P\left(\frac{X - 5}{2} < \frac{4 - 5}{2}\right) = P(Z < -0.5) = \phi(-0.5) = 1 - \phi(0.5) = 1 - 0.6915 \\ &= 0.3085. \end{aligned}$$

$$\begin{aligned} P(X > 5.5) &= 1 - P(X \leq 5.5) = 1 - P\left(Z \leq \frac{5.5 - 5}{2}\right) = 1 - P(Z < 0.25) = 1 - 0.5987 \\ &= 0.4013. \end{aligned}$$

$$P(4 < X < 5.5) = P(X < 5.5) - P(X \leq 4) = 0.5987 - 0.3085 = 0.2902 = 29.02\%.$$

$$\begin{aligned} P(2 < X < 6) &= P\left(\frac{2 - 5}{2} < Z < \frac{6 - 5}{2}\right) = P(-1.5 < Z < 0.5) = \phi(0.5) - \phi(-1.5) \\ &= \phi(0.5) + \phi(1.5) - 1 = 0.6915 + 0.9332 - 1 = 0.6247 = 62.47\%. \end{aligned}$$

Pour la même loi $\mathcal{N}(5, 2)$ on cherchera les valeurs x et y tel que :

$$P(X \leq x) = 0.8554 \quad \text{et} \quad P(X \geq y) = 87.9\%$$

$$P(X \leq x) = P\left(Z < \frac{x - 5}{2}\right) = 0.8554 \quad \Rightarrow \quad \frac{x - 5}{2} = 1.06 \quad \Rightarrow \quad x = 1.06 * 2 + 5 = 7.12.$$

$$P(X \geq y) = 1 - P\left(Z < \frac{y - 5}{2}\right) = 1 - \phi\left(\frac{y - 5}{2}\right) = \phi\left(\frac{5 - y}{2}\right) = 87.9\% = 0.879$$

$$\Rightarrow \frac{5 - y}{2} = 1.17 \quad \Rightarrow \quad y = 5 - 1.17 * 2 = 2.66.$$

4.1.2 Théorème d'addition de v.a. normales

Soit (X, Y) un couple de v.a. indépendantes. On a le résultat suivant :

Si $X \sim N(\mu_1, \sigma_1)$ et $Y \sim N(\mu_2, \sigma_2)$ alors

$$X + Y \sim N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Ce résultat se généralise à n v.a. indépendantes, de lois normales :

Si $X_i \sim N(\mu_i, \sigma_i)$ pour tout $i = 1, \dots, n$ alors

$$\sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sqrt{\sum_{i=1}^n \sigma_i^2}\right)$$

On peut introduire des paramètres $\alpha_i \in \mathbb{R}$ dans la somme de n v.a. :

Si $X_i \sim N(\mu_i, \sigma_i)$ et $\alpha_i \in \mathbb{R}$ pour tout $i = 1, \dots, n$ alors

$$\sum_{i=1}^n \alpha_i X_i \sim N\left(\sum_{i=1}^n \alpha_i \mu_i, \sqrt{\sum_{i=1}^n \alpha_i^2 \sigma_i^2}\right)$$

En particulier si on prend

$$\alpha_i = \frac{1}{n} \text{ pour tout } i = 1, \dots, n$$

On aura le résultat :

Si $X_i \sim N(\mu_i, \sigma_i)$ pour tout $i = 1, \dots, n$; si on note $\bar{X}_n$ la moyenne arithmétique des n v.a. X_i alors

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\frac{\sum_{i=1}^n \mu_i}{n}, \sqrt{\frac{\sum_{i=1}^n \sigma_i^2}{n}}\right)$$

En particulier, s'il s'agit de la même loi normale, c.à.d. qu'on a : $\mu_i = \mu$ et $\sigma_i = \sigma$ pour tous $i = 1, \dots, n$, alors

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Remarques :

- La moyenne ou la somme de n v.a. indépendantes de lois normales est toujours une loi normale, reste à déterminer les deux paramètres de cette loi.

- Si X et Y sont deux lois normales avec $Cov(X, Y) = 0$ alors X et Y sont indépendantes. La non corrélation entre X et Y est une condition suffisante pour que X et Y soient indépendants, ce qui n'est pas vrai pour les autres lois de probabilité.

1.6. Lois découlant de la loi normale

Loi de khi-deux

Soit $X_1, X_2, \dots, X_n$, n v.a. indépendantes de même loi normale standard $N(0,1)$, alors la v.a.

$$Y = \sum_{i=1}^n X_i^2$$

suit une loi de probabilité dite de khi-deux à n degrés de liberté.

On note $Y \sim \chi^2(n)$ ou $Y \sim \chi_n^2$.

Cette loi est utilisée en statistique dans l'estimation de la variance σ^2 inconnue d'une population.

Loi de Student

Soit deux v.a. X et Y tel que $X \sim N(0,1)$ et $Y \sim \chi^2(n)$, alors la v.a.

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} = \sqrt{n} \frac{X}{\sqrt{Y}}$$

suit une loi de probabilité dite de Student à n degrés de liberté.

On note $T \sim t(n)$.

Cette loi sera utilisée en statistique dans l'estimation de la moyenne μ inconnue d'une population dont la variance σ^2 est aussi inconnue.

2. Convergence en probabilité et convergence en loi

On donne dans ce paragraphe les deux convergences qui vont caractériser les estimateurs, on parle ici de convergence en probabilité et de convergence en loi. On donne aussi la loi faible des grands nombres (LFGN) et le fameux théorème central limite (TCL).

Deux applications du TCL sont prévues pour les deux lois discrètes binomiale et de Poisson, dans ce cas une correction de continuité est envisageable.

2.1. Convergence en probabilité

Définition : Une suite de variables aléatoires (X_n) converge en probabilité vers une variable aléatoire X si et seulement si, pour tout $\varepsilon > 0$

$$\lim_{n \rightarrow +\infty} P(|X_n - X| > \varepsilon) = 0$$

On note $X_n \xrightarrow{P} X$

Pour tout écart fixé ε , quand n devient très grand, il est de moins en moins probable d'observer des écarts supérieurs à ε .

Un exemple très apprécié de la convergence en probabilité est la loi faible des grands nombres.

Théorème de la LFGN

Soit X une v.a. et (X_n) une suite de v.a. de même loi que X et deux à deux indépendantes

Alors, pour tout $\varepsilon > 0$

$$\lim_{n \rightarrow +\infty} P(|\bar{X}_n - E(X)| > \varepsilon) = 0$$

Remarques :

- La condition deux à deux indépendantes peut être allégée par la condition de la non corrélation deux à deux ; c.à.d.

$$\text{Cov}(X_i, X_j) = 0 \text{ si } i \neq j.$$

- D'après la LGN, on est presque sûr que les valeurs observées de la moyenne d'une série de données issues de la même loi de probabilité et qui sont au moins non corrélées se trouvent au voisinage de l'espérance, et cela est d'autant vrai quand n est assez grand.

$$\begin{aligned} \lim_{n \rightarrow +\infty} P(|\bar{X}_n - E(X)| > \varepsilon) = 0 &\Leftrightarrow \lim_{n \rightarrow +\infty} P(|\bar{X}_n - E(X)| \leq \varepsilon) = 1 \\ &\Leftrightarrow \lim_{n \rightarrow +\infty} P(\bar{X}_n \in [E(X) - \varepsilon ; E(X) + \varepsilon]) = 1 \end{aligned}$$

- La LFGN s'écrit

$$\bar{X}_n \xrightarrow{P} E(X)$$

Exemple : Si on joue n fois avec une pièce de monnaie bien équilibrée, on pose X_i la v.a. qui prend la valeur 1 si le résultat du $i^{\text{ème}}$ lancé est Pile, alors $X_i \sim \mathcal{B}\left(\frac{1}{2}\right)$; la moyenne $\bar{X}_n$ (qu'on note aussi $\bar{p}_n$) est la v.a. fréquence (pourcentage) de l'événement Pile, si on a supposé que

le dé n'est pas truqué, on s'attend à une fréquence de $0.5 = 50\%$, cette fréquence n'est autre que l'espérance de X_i :

$$E(X_i) = \frac{1}{2} = 0.5$$

Cela s'exprime par la LGN, pour un $\varepsilon > 0$ assez petit, on a :

$$\lim_{n \rightarrow +\infty} P(\bar{X}_n \in [0.5 - \varepsilon; 0.5 + \varepsilon]) = 1$$

D'une autre manière, si on prend des valeurs différentes de ε , il suffit d'augmenter la taille n pour être presque sûr que la fréquence observée $\bar{x}_n$ est proche de la valeur espérée.

Si $\varepsilon = 0.1$ on écrit $\lim_{n \rightarrow +\infty} P(\bar{X}_n \in [0.4; 0.6]) = 1$

Si $\varepsilon = 0.01$ on écrit $\lim_{n \rightarrow +\infty} P(\bar{X}_n \in [0.49; 0.51]) = 1$

Si $\varepsilon = 0.001$ on écrit $\lim_{n \rightarrow +\infty} P(\bar{X}_n \in [0.499; 0.501]) = 1$

2.2. Convergence en loi

Définition : Une suite de v.a. (X_n) de fonction de répartition F_n converge en loi vers une v.a. X de fonction de répartition F si et seulement si, pour tout $x \in \mathbb{R}$

$$\lim_{n \rightarrow +\infty} F_n(x) = F(x)$$

On note $X_n \xrightarrow{\mathcal{L}} X$

Posons :

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

Il s'agit de la moyenne de n v.a. de même loi de probabilité, de même espérance $E(X_1)$ et de même variance $Var(X_1)$.

Théorème central limite (TCL)

Soit n v.a. $X_1, X_2, \dots, X_n$ indépendantes et de même loi de probabilité, de moyenne $E(X_1) = \mu$ et de variance $Var(X_1) = \sigma^2$. Alors, on a :

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

Si on centre et on réduit la v.a. $\bar{X}_n$, sa loi sera approximativement une loi normale centrée réduite.

$$E(\bar{X}_n) = \mu,$$

$$Var(\bar{X}_n) = \frac{\sigma^2}{n} \quad \Rightarrow \quad \mathcal{L} \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right) = \mathcal{L} \left(\sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \right) \cong N(0,1)$$

On remplace pour n assez grand la fonction de répartition de la v.a. $\sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right)$ par la fonction de répartition de la loi normale $N(0,1)$.

En général, on considère que pour $n \geq 30$, on aura une bonne approximation par le TCL.

Si on considère la somme de n v.a. indépendantes et de même loi, $S_n = \sum_{i=1}^n X_i$ alors

$$\mathcal{L}(S_n) \cong N(n\mu, \sqrt{n} \sigma)$$

Rappelons que les paramètres de la loi normale ci-dessus sont l'espérance et l'écart-type de S_n , $E(S_n) = n\mu$; $Var(S_n) = n\sigma^2$ et $\sigma(S_n) = \sqrt{n} \sigma$.

2.3. Approximation par la loi normale

2.3.1. Cas de la loi binomiale

Soit n v.a. $X_1, X_2, \dots, X_n$ indépendantes et de même loi Bernoulli $B(p)$, on a :

$$E(X_1) = p \quad \text{et} \quad Var(X_1) = p(1-p)$$

La v.a. $S = \sum_{i=1}^n X_i$ suit une loi binomiale $B(n, p)$, on a :

$$E(S) = np \quad \text{et} \quad Var(S) = np(1-p)$$

Pour n assez grand, la loi de probabilité de S est approximativement une loi normale :

$$\mathcal{L}(S) \cong N(np, \sqrt{np(1-p)})$$

Si on centre et on réduit la v.a. S , on écrit :

$$\mathcal{L} \left(\frac{S - np}{\sqrt{np(1-p)}} \right) \cong N(0,1) \quad \text{pour } n \text{ assez grand}$$

On rappelle qu'on parle toujours de la convergence en loi :

$$\frac{S - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} N(0,1)$$

A l'origine de ce résultat était De Moivre (1738) et Laplace (1812), qu'on tire du nom du théorème de Moivre-Laplace. (De Moivre a démontré le résultat pour le cas d'équiprobabilité $p = 0.5$ et Laplace l'a généralisé pour n'importe quelle valeur de p).

Remarque : Cette approximation est d'autant meilleure si on a les conditions suivantes,

$$\begin{cases} n \geq 30 \\ p < 0,1 \\ np(1-p) > 5 \end{cases}$$

Si on veut que p prenne n'importe quelle valeur entre 0 et 1, on peut se restreindre à la seule condition suivante

$$np(1-p) \geq 10$$

2.3.2. Cas de la loi de Poisson

Soit n v.a. $X_1, X_2, \dots, X_n$ indépendantes et de même loi de Poisson $\mathcal{P}(\lambda)$, d'espérance $E(X_1) = \lambda$ et de variance $Var(X_1) = \lambda$. La v.a. $S = \sum_{i=1}^n X_i$ suit une loi Poisson $\mathcal{P}(n\lambda)$, on a :

$$E(S) = n\lambda \text{ et } Var(S) = n\lambda$$

Pour n assez grand, la loi de probabilité de S est approximativement une loi normale $N(n\lambda, \sqrt{n\lambda})$.

$$\mathcal{L}\left(\frac{S - n\lambda}{\sqrt{n\lambda}}\right) \cong N(0,1) \text{ pour } n \text{ assez grand}$$

On va considérer que cette approximation est valable si $n\lambda \geq 10$.

Si on prend le cas particulier où $\lambda = 1$, on aura :

$$\mathcal{L}\left(\frac{S - n}{\sqrt{n}}\right) \cong N(0,1)$$

Cette approximation est valable pour $n \geq 10$.

2.3.3. Correction de continuité

Dans le cas de l'utilisation du TCL pour approcher une loi discrète X par la loi continue normale, on devra corriger cette approximation en considérant des intervalles continus centrés sur les valeurs prises par X , par exemple si $X \in \mathbb{N}$, l'intervalle $[1, 2]$ sera remplacé par l'intervalle $[0,5 ; 2,5]$.

On a pour toute valeur $k \in \mathbb{N}$

$$P(X = k) = P(k - 0.5 < X < k + 0.5)$$

$$P(X < k) = P(X < k - 0.5) \text{ et } P(X > k) = P(X > k + 0.5)$$

$$P(X \leq k) = P(X < k + 0.5) \text{ et } P(X \geq k) = P(X > k - 0.5)$$

Pour $k > l$ on aura :

$$P(l \leq X < k) = P(l - 0.5 < X < k - 0.5)$$

$$P(l < X < k) = P(l + 0.5 < X < k - 0.5)$$

Exemple :

Si $X \sim B(100; 0,06)$, on veut calculer $P(3 \leq X \leq 7)$.

- Exactement, on aura la calcul suivant :

$$P(3 \leq X \leq 7) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) = 69,17\%$$

- En utilisant le TCL sans correction de continuité, on aura le résultat suivant :

$$\begin{aligned} P(3 \leq X \leq 7) &\approx P\left(\frac{-3}{\sqrt{5,64}} \leq Z \leq \frac{1}{\sqrt{5,64}}\right) = \phi(0,42) - \phi(-1,26) \\ &= \phi(0,42) + \phi(1,26) - 1 = 55,90\% \end{aligned}$$

- En utilisant le TCL avec correction de continuité, on aura :

$$\begin{aligned} P(3 \leq X \leq 7) &= P(2,5 \leq X \leq 7,5) \approx P\left(\frac{-3,5}{\sqrt{5,64}} \leq Z \leq \frac{1,5}{\sqrt{5,64}}\right) = \phi(0,63) - \phi(-1,47) \\ &= 66,49\% \end{aligned}$$

On peut conclure que sans correction de continuité, la différence entre la probabilité $P(3 \leq X \leq 7)$ et son approximation par la loi normale est égale à 0,1327 ; cette différence s'est réduite à 0,0268 quand on a corrigé l'intervalle $[3, 7]$ par l'intervalle $[2,5 ; 7,5]$.

Pour le calcul des probabilités directement sur les points observés de la v.a. discrète X , on considère la formule $P(X = k) = P(k - 0.5 < X < k + 0.5)$, par exemple :

$$P(X = 6) = P(5,5 \leq X \leq 6,5) \approx P\left(\frac{-0,5}{\sqrt{5,64}} \leq Z \leq \frac{0,5}{\sqrt{5,64}}\right) = 2 * \phi(0,21) - 1 = 16,64\%$$

On vérifie bien la pertinence du calcul : $P(X = 6) = C_{100}^6 * 0.06^6 * 0.94^{94} = 16.57\%$

3. Méthodes d'échantillonnage

Dans un recensement, tous les sujets de la population sont examinés, on envisage un coût élevé et même dans certains cas, l'impossibilité de recenser toutes les unités du fait qu'elles ne sont pas toutes accessibles. Le sondage (on parle aussi d'enquête par sondage) cible une partie de la population, il va choisir par une méthode, un échantillon représentatif de toute la population. Ce qui va nous permettre d'étendre (ou d'extrapoler) les résultats obtenus sur l'échantillon à des conclusions concernant toute la population (inférence statistique).

On parle de méthode de sondage, ou de méthode d'échantillonnage ou tout simplement d'échantillonnage, le procédé par lequel on va choisir un échantillon de la population.

Il y a deux approches pour construire un échantillon, une première qui utilise les méthodes dites aléatoires et une deuxième qui n'est pas aléatoire.

Les méthodes aléatoires sont des méthodes qui ont recours à la théorie des probabilités, elles ont la particularité d'attribuer à chaque unité de la population une probabilité, non nulle, d'être sélectionnée (toute unité de la population a une chance d'appartenir à l'échantillon donc de représenter la population).

3.1. Méthodes aléatoires

3.1.1. L'Echantillonnage aléatoire simple

L'échantillonnage aléatoire simple consiste à choisir des individus de telle sorte que chaque membre de la population a **une chance égale** de figurer dans l'échantillon.

Ce choix peut se faire avec remise ou sans remise :

- **Avec remise**, un individu peut être choisi plusieurs fois.
- **Sans remise**, un individu déjà choisi ne peut l'être de nouveau. C'est le cas habituel.

Avantage de cette méthode : On peut espérer un échantillon «représentatif » puisque la méthode donne à chaque individu de la population une chance égale.

Difficultés : la méthode n'est applicable que lorsqu'il existe une liste exhaustive de toute la population.

Comment procéder ?

1. On numérote tous les individus de la liste correspondant aux individus de la population avec des nombres comportant un même nombre de chiffres.
2. En utilisant une table de nombres aléatoires, une calculatrice ou un programme informatique, on obtient des nombres aléatoires comportant le nombre de chiffres désiré.
3. On sélectionne les nombres qui coïncident avec la liste. On rejette les nombres qui ne coïncident pas avec la liste ou qui se répètent, on s'arrête après avoir sélectionné n individus (n représentant le nombre d'individus souhaités dans l'échantillon).

Avec Excel, on peut utiliser une des deux fonctions `alea()` ou `alea.entre.bornes()`

1. Avec le nombre aléatoire « alea »
 - Première colonne : identifie avec un nombre chaque individu de la liste de référence.
 - Deuxième colonne : `=alea()`
 - Recopier les deux colonnes en valeur à la même place.
 - Trier les deux colonnes en fonction de l'ordre croissant (ou décroissant) de la deuxième colonne.
 - Retenir les n premiers individus dans la colonne 1
2. Directement avec la fonction `alea.entre.bornes()`

Combien peut-on réaliser d'échantillons ?

Si l'on note n la taille de l'échantillon et N la taille de la population.

Pour des valeurs $n = 30$ et $N = 100$, on aura jusqu'à 10^{60} échantillons ESAR et 3×10^{25} ESSR. (On note ESAR pour Echantillon Simple Avec Remise et ESSR pour le cas Sans Remise)

Pour simplifier la construction d'un ES, on a recours à la méthode systématique qui consiste à la prédétermination des échantillons possibles à la sélection.

Exemple : on a une population de 300 individus, on veut un échantillon de 60 individus (taux de sondage = $\frac{60}{300} = 20\%$)

- 1 On calcule ce qu'on appelle le pas de sondage $R = \frac{N}{n} = \frac{300}{60} = 5$
- 2 On détermine les 5 échantillons possibles :

- 1, 6, 11, 296
- 2, 7, 12, 297
- 3, 8, 13, 298
- 4, 9, 14, 299
- 5, 10, 15, 300

3 Pour connaître l'échantillon de l'étude, on choisit au hasard un entier entre 1 et 5.

Remarque : Si la population est distribuée au hasard dans la base de sondage, un échantillonnage systématique donnera des résultats similaires à ceux d'un échantillonnage aléatoire simple. La méthode systématique est très utilisée dans les contrôles de qualité, elle est appréciée par son implémentation facile. Un désavantage de la méthode est que les données peuvent être biaisées à cause de la périodicité (qui peut coïncider avec le pas de sondage R)

3.1.2. L'échantillonnage stratifié

Une strate est un groupe ou sous population homogène. Si on peut subdiviser la population en strates qui sont relativement homogènes et mutuellement exclusives, on doit prélever des sous-échantillons dans chaque strate.

Comment procéder ?

1. On calcule pour chaque strate, le nombre d'individus qu'il faudra sélectionner. Ce nombre est proportionnel à l'importance de la strate dans la population : plus grande est la taille de la strate, plus grand sera le sous-échantillon prélevé.
2. Dans chaque strate, on choisit au hasard les individus en nombre nécessaire.

Pour réussir cette méthode d'échantillonnage, les variables de stratification doivent être simples à utiliser, faciles à observer et **étroitement reliées** au thème de l'enquête.

L'avantage de la méthode, c'est qu'il est fort probable de choisir un échantillon raisonnable de la population puisqu'on s'assure de la présence proportionnelle de tous les divers sous-groupes composant la population.

Remarques :

- La méthode suppose l'existence d'une liste de la population.

- La variance totale des données est la somme de la variance intra-strate (à l'intérieur de la strate) et de la variance inter-strate (entre les différentes strates).

On cherche à avoir la plus petite variance intra-strate (des individus homogènes à l'intérieur de la même strate) et une grande variance inter-strate (le facteur ou la variable de stratification explique bien la classification ou le regroupement en strates).

Exemple : on veut choisir par échantillonnage stratifié 10 étudiants dans un groupe de 60, en tenant compte du fait que 50% d'entre eux sont en Gestion, 30% en Économie et Gestion et 20% en Économie. Le facteur de stratification est le caractère « Option » qui prend trois modalités : « Gestion », « Économie et Gestion » et « Économie ».

3.1.3. L'échantillonnage par grappes

Dans les méthodes précédentes, l'unité statistique était choisie individuellement. La technique de l'échantillonnage en grappes entraîne la division de la population en groupes ou grappes. On sélectionne au hasard un certain nombre de grappes (unités primaires) pour représenter la population. On sélectionne ensuite tous les individus des grappes choisies.

L'avantage de la méthode, c'est qu'elle ne nécessite pas une liste globale de la population, puisque seuls les individus inclus dans les grappes comptent. Elle permet de limiter l'échantillon à des groupes compacts, ce qui permet de réduire les coûts de déplacement, de suivi et de supervision.

La méthode peut entraîner des résultats imprécis (moins précis que les méthodes précédentes) puisque les unités voisines ont tendance à se ressembler. Et elle ne permet pas de contrôler la taille finale de l'échantillon.

Pour réussir un échantillonnage par grappes, il est recommandable :

- d'éviter de construire de grosses grappes afin d'en garantir un nombre suffisant.
- d'avoir des tailles de grappes le plus uniforme que possible (tailles se rapprochant d'une taille unique).
- que les unités composant la grappe soient aussi hétérogènes que possible du point de vue du caractère étudié.

Exemple : Dans la faculté des sciences, on choisit aléatoirement des groupes de TP pour constituer l'échantillon final, chaque groupe de TP est constitué d'un certain nombre d'étudiants, les tailles des groupes peuvent varier.

3.1.4. L'échantillonnage à plusieurs degrés

Ressemble à l'échantillonnage en grappes, sauf que dans ce cas on prélève un échantillon à l'intérieur de chaque grappe. On a au moins deux degrés :

On identifie au premier les grandes grappes (unités primaires). Ces grappes renferment plus d'unités qu'il n'en faut dans l'échantillon

Au second degré, à l'intérieur de chaque grappe, on sélectionne les unités (unités secondaires) qui vont faire partie de l'échantillon

On peut utiliser plus de 2 degrés, par **exemple**, si on veut un échantillon des enseignants des universités marocaines, on sélectionne quelques universités du royaume. Dans chaque université sélectionnée, on choisit quelques établissements et dans chaque établissement choisit, on fait un tirage au sort de quelques départements. Tous les enseignants du département sélectionné seront dans l'échantillon final.

L'avantage de cette méthode c'est qu'on aura un échantillon plus concentré, ce qui réduit les coûts, de plus on n'a pas besoin de disposer de la liste de toutes les unités.

3.1.5. L'échantillonnage à plusieurs phases

Les données de base sont collectées auprès d'un échantillon d'unité de grande taille, ensuite pour un sous-échantillon de ces unités, la collecte des données est plus détaillée.

Le plus couramment on utilise deux phases, par exemple, on a besoin d'information sur les éleveurs de bétail. Dans la base de sondage, sont énumérés les types d'exploitation agricoles: bétail, grains, volaille, fruit, légumes... mais on n'a aucune donnée auxiliaire. C'est pour cette raison qu'on pourrait mener une enquête sur un premier échantillon de grande taille dont la seule question serait : *Vous consacrez vous en totalité au bétail ?* Puis sur la sélection des éleveurs uniquement de bétail, on extrait un second échantillon plus petit que le premier auquel on pose des questions détaillées.

3.2. Méthodes empiriques ou non aléatoires

Les méthodes non aléatoires sont des méthodes où le concept de «chance égale» est absent. Ce sont des méthodes généralement peu fiables. Elles ne nécessitent pas de base de sondage, elles sont souvent utilisées

- pour des études exploratoires;
- pour réduire les coûts;
- quand il est impossible ou non envisageable d'utiliser la méthode aléatoire.

L'avantage essentiel de ces méthodes, c'est qu'elles sont moins coûteuses et plus faciles à réaliser.

Le calcul de probabilité n'est pas possible puisqu'on ne peut pas connaître à priori la probabilité qu'une unité sélectionnée appartienne à l'échantillon.

Parmi les méthodes non aléatoires, on distingue :

3.2.1. L'échantillonnage par quotas

Il est largement utilisé dans les enquêtes d'opinion et les études de marché notamment parce qu'il ne suppose pas de liste des individus de la population. On parle aussi d'échantillonnage dirigé ou par choix raisonné. On demande aux enquêteurs de faire un nombre d'entrevues dans divers groupes établis en fonction du secteur géographique, de l'âge, du sexe ou d'autres caractéristiques...

Le principe de cette méthode est de reproduire un échantillon qui va présenter les distributions statistiques de certains caractères appelés variables de contrôle, de telles distributions doivent être identiques aux distributions connues de la population.

L'idée d'une telle démarche est qu'on a plus de chances pour que les distributions de l'échantillon pour les autres caractères étudiés soit les mêmes que celles de la population.

Le choix des unités est laissé à la discrétion de l'enquêteur, qui doit respecter son quota.

Les quotas sont déterminés en multipliant le taux de sondage par les effectifs des différentes modalités des variables de contrôle.

Pour choisir une variable de contrôle, on doit vérifier si sa distribution dans la population est connue. En plus, comme dans le cas du choix de la variable de stratification, la variable de contrôle doit être en corrélation avec la (ou les) variable(s) étudiée(s).

Exemple : on veut sélectionner un échantillon par quotas dans un groupe d'étudiants de $N = 600$, le taux de sondage est $n/N = 20\%$. La taille de l'échantillon sera $n = 120$.

On connaît la répartition du groupe selon les deux caractères Age et Sexe (deux variables de contrôle) :

Sexe	Effectifs	Fréquence
Féminin	360	60%
Masculin	240	40%
Total	600	100%

Age	Effectifs	Fréquence
17- 19	280	46.67%
19-21	220	36.67%
21-25	60	10%
25-33	30	5%
33-51	10	1.67%
Total	600	100%

Alors on multiplie le taux de sondage aux différents effectifs pour obtenir les quotas :

Sexe	Effectifs	Fréquence
Féminin	72	60%
Masculin	48	40%
Total	120	100%

Age	Effectifs	Fréquence
17- 19	56	46.67%
19-21	44	36.67%
21-25	12	10%
25-33	6	5%
33-51	2	1.67%
Total	120	100%

Si on demande à un enquêteur de réaliser 24 interviews, on lui fournit un tableau de contrôle qu'il doit respecter :

Sexe	Effectifs
Féminin	14
Masculin	10

Age	Effectifs
17- 19	11
19-21	9
21-25	2
25-33	1
33-51	1

3.2.2. L'échantillonnage à l'aveuglette ou de commodité

L'échantillonnage de commodité est une technique d'échantillonnage non probabiliste où les individus sont choisis en raison de leur accessibilité et de la proximité de l'enquêteur. Simple à réaliser, elle sera utilisée par l'enquêteur quand il estime non utile de représenter toute la population, cela va entraîner un biais entre les résultats obtenus et les hypothèses théoriques provenant de toute la population.

Exemples :

- Déguster des yaourts d'une marque, goûter à des plats principaux, afin de juger de la qualité.
- Un journaliste d'une chaîne télévisée interviewe des gens dans la rue pour comprendre comment la population perçoit un enjeu ou une question.
- On veut interroger 30 clients d'une grande surface, on décide de sélectionner les trente premiers clients qui entrent.

3.2.3. L'échantillonnage de volontaires

On sélectionne un échantillon parmi des volontaires.

Par exemple, il est impossible de choisir au hasard des personnes du grand public, pour des expériences médicales ou psychologiques. Pour cette raison qu'on prélève l'échantillon à partir d'un groupe de volontaires, qui acceptent la possibilité de subir tout un processus long ou exigeant. Ce type d'échantillonnage entraîne des biais puisqu'on ne peut pas accéder à toute la population.

3.2.4. L'échantillonnage au jugé

Cette méthode implique la sélection d'individus en fonction de l'idée qu'on se fait de la composition de la population (hypothèse sur laquelle l'enquêteur sélectionnera des unités qui représentent la population). On le fait pour des essais auprès des groupes cibles.

L'inconvénient principal, c'est qu'on doit se fier à l'hypothèse de l'enquêteur qui a ses propres idées préconçues (qui peuvent se révéler inexactes), au risque d'avoir des biais dans les résultats.

3.3. Les erreurs

Les méthodes d'échantillonnage peuvent être sources d'erreurs. Un certain nombre d'erreurs pourront être éliminées, certaines pourront être réduites, mais d'autres persisteront :

- **Les erreurs dues aux instruments de mesure :**

Un instrument est fidèle s'il répond exactement de la même façon quand il est placé dans deux situations identiques. Exemple le thermomètre. Une question claire est dite fidèle quand tout le monde la comprend de la même façon.

Un instrument est valide lorsqu'il mesure vraiment ce qu'il est censé mesurer.

- **Les erreurs dues à l'organisation :**

Ce sont les erreurs qui se glissent lors de la collecte des données.

Est-ce que les consignes ont été respectées ?

Les enquêteurs ont-ils agi de la même façon ?

Pour éviter ces erreurs, il faut utiliser les mêmes instruments et les mêmes conditions.

- **Les erreurs dues à la méthode d'échantillonnage**

Il faut toujours vérifier, à la lumière des objectifs de l'étude statistique, que la méthode d'échantillonnage est adaptée. En particulier éviter la surreprésentation de certaines parties de la population.

- **Les erreurs dues au phénomène de non-réponse**

Même avec la meilleure méthode d'échantillonnage, il se présente toujours un certain nombre de non-répondants, ce qui peut entacher la représentativité de l'échantillon et amener des conclusions erronées.

- **L'erreur d'échantillonnage**

Le fait d'étudier un échantillon plutôt qu'un autre engendre forcément une erreur. Cette erreur appelée erreur d'échantillonnage est inévitable.

Pour en savoir plus sur les sondages et leurs applications, voici quelques références :

Bibliographie

1. *Méthodes statistiques*, techniques statistiques, B. Grais.
2. *L'enquête par questionnaire*, manuel à l'usage du praticien. C.Javeau.
3. *Au royaume des sondages*, J.-J. Droesbeke et G. Thoveron.
4. *Les techniques de sondage* P. Ardilly, édition TECHNIP.
5. *Documents pédagogiques* sur le site de *Statistique Canada* www.statcan.ca