L'échantillonnage et L'estimation

Dr. Meriem Bouhadjar

2020/2021

1. L'échantillonnage et L'estimation

L'échantillonnage

On veut, à partir d'un échantillon de la population, déduire des informations sur cette population. Le problème qui se pose alors est le suivant : comment choisir une partie de la population qui reproduit le plus fidélement possible ses caractéristiques. C'est le problème de l'échantillonnage.

L'échantillonnage

Echantillon : ensemble des unités de base sélectionnées et réellement observées au cours d'un sondage.
Echantillonnage : ensemble des opérations qui permettent de sélectionner de façon organisée les éléments de l'échantillon.
Sondage : Enquête incomplète, enquête partielle ou enquête par échantillonnage, c'est une enquête au cours de laquelle seulement une partie des unités de base de la population sont observée.
Base de sondage : énumération ou présentation ordonnée de toutes les unités de base constituant la population.

Prélèvement d'un échantillon (échantillonnage)

1. Echantillonnages sur la base des méthodes empiriques

La Méthode des quotas (respect de la composition de la population pour certains critères) est la plus utilisée

Remarque. La méthode des quotas est la méthode la plus employée par les instituts de sondage. Cette méthode ne contient pas d'élément aléatoire et par conséquent sa fiabilité ne peut être mathématiquement calculée puisqu'on ne peut pas utiliser le calcul des probabilités.

C'est une méthode d'échantillonnage qui consiste à s'assurer de la représentativité d'un échantillon, en lui affectant une structure similaire à celle de la population mère

Prélèvement d'un échantillon (échantillonnage)

2. Echantillonnages aléatoires

Quand la probabilité de sélection de chaque élément de la population est déterminée avant même que l'échantillon soit choisi.
Il permet de juger objectivement la valeur des estimations.
On a quelque méthodes d'échantillonnage probabilistes
Echantillonnage aléatoire simple on tire au hasard et avec remise les unités dans la population concernée.
Echantillonnage stratifié
– Subdiviser d'abord la population échantillonner en sous-ensembles (strates) relativement homogènes.
– Extraire de chaque strate un échantillon aléatoire simple.
– Regrouper tous ces échantillons.

Prélèvement d'un échantillon (échantillonnage)

Echantillonnage par grappes
– Choisir un échantillon aléatoire d'unités qui sont elles-mêmes des sous-ensembles (ex : diviser la ville en quartiers ; un certain nombre de quartiers sont choisis pour faire partie de l'échantillon ; on fait l'enquête auprès de toutes les familles résidant dans ces quartiers).

Quelques statistiques classiques

Rappels

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

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

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

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

si X, Y indépendantes,

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

La moyenne empirique et la variance empirique

Posons $E(X) = \mu, V(X) = \sigma^2$ (inconnues)

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

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

Sa réalisation est $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$(qui est la moyenne de l'échantillon) aussi appelée *moyenne observée.*

•**Propriétés :**

$$\begin{cases} E(\overline{X}) = \mu \\ V(\overline{X}) = \frac{1}{n}\sigma^2 \end{cases}$$

La moyenne empirique et la variance empirique

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

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

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

•**Propriétés :**

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

Lois de probabilité des statistiques

• Théorème limite centrale (pour l'échantillon) (rappel) :
soit X une v.a. t.q. $E(X) = \mu, V(X) = \sigma^2 \neq 0$
Soit $(X_1, ..., X_n)$ un n- échantillon de X
$\bar{X} = \frac{1}{n}(X_1, ..., X_n)$
$Alors\ \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \backsim \mathcal{N}(0,1)$pour $n \to \infty$
(loi approximative)
(ou bien $\overline{X} \backsim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ pour $n \to \infty$

• **2 cas à étudier :**
– a) la taille n de l'échantillon est grande
– b) X suit une loi gaussienne

a) Taille n grande
(d'après le thm. limite centrale)
1/ $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$suit approximativement $\mathcal{N}(0,1)$

Lois de probabilité des statistiques

Quelques statistiques classique :

$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \backsim \mathcal{N}(0,1)$pour $n \to \infty$

Ou bien $\overline{X}$ suit approximativement $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ en pratique $n > 30$

Lois de probabilité des statistiques

•**Exercice :**Soit un lot de 500 chocolats. Le poids d'un chocolat est une v.a. telle que $\mu = 5g$ et $\sigma = 0,5g$. Queelle st la probabilité qu'une boîte de 50 chocolats issus de ce lot ait un poids total supérieur à $260g$?

solution

L'échantillon étant grand$(n = 50 > 30)$ et on peut appliquer la première formule :

$\overline{X} \backsim \mathcal{N}\left(5; \frac{0,5}{\sqrt{50}}\right)$

approximativement

on pose $T = 50\overline{X}$; cette nouvelle v.a. suit approximativement :

$T \backsim \mathcal{N}\left(5.50; \frac{50.0,5}{\sqrt{50}}\right) = \mathcal{N}\left(250; 0,5\sqrt{50}\right)$

calculons

$P(T > 260) = P\left(Z > \frac{260-250}{0,5\sqrt{50}}\right) = P(Z > 2,83) = 1 - P(Z < 2,83) =$

$1 - 0,9977 = 0,0023$

Lois de probabilité des statistiques

b) Echantillon gaussien

Soit $X \backsim \mathcal{N}(\mu, \sigma)$

(d'après l'additivité pour des v.a. suivant des lois normales)

1/ $\overline{X} \backsim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

ou bien

$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \backsim \mathcal{N}(0,1)$

c'est une loi exacte et non une approximation comme dans le cas d'un échantillon de grande taille où la loi n'est pas connue.

2/ $\frac{n}{\sigma^2} S^2 \backsim \mathcal{X}^2_{n-1}$

3/ $\frac{\overline{X}-\mu}{\sqrt{S^2}/\sqrt{n-1}} \backsim t_{n-1}$

Fréquence empirique F

Soit une population comportant deux modalités A et B. Soit la proportion d'individus de la population possédant la modalité A. $1-\pi$ est donc la proportion des individus de la population possédant la modalité B. On extrait de la population un échantillon de taille n. Soit K_n la v.a qui représente le nombre d'individus dans l'échantillon ayant la modalité A.

•**Définition :** La v.a. $F = \frac{K_n}{N}$ s'appelle *fréquence empirique.*
Sa réalisation f est la proportion d'individus dans l'échantillon ayant la modalité A.

•**Propriétés :**

$$\begin{cases} K \backsim \beta(n,\pi) \text{ donc} \\ E(F) = \pi \\ V(F) = \frac{\pi(1-\pi)}{n} \end{cases}$$

Fréquence empirique F

- **Loi de probabilité pour F :**

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

dès que $n > 30$, $\pi \in [0.1, 0.9]$. On trouve aussi $n > 5$, $n(1-\pi) > 5$
ou les seules conditions $n > 5, n(1-\pi) > 5)$
(loi approximative).
$\frac{F-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \backsim \mathcal{N}(0,1)$

L'estimation

Les premiers problèmes d'inférence statistique auxquels s'applique la théorie des distributions d'échantillonnage sont les problèmes d'estimations. Le but poursuivi est d'estimer, à partir d'un échantillon, la ou les valeurs numériques d'un ou de plusieurs paramètres de la population considérée et de déterminer la précision de cette ou de ces estimations.
On distingue deux formes d'estimations : l'estimation ponctuelle et l'estimation par intervalle de confiance.

L'estimation

Notations.

• les paramètres à estimer seront notés par des lettres grecques minuscules

μ : *moyenne*

σ : *écart − type*

σ^2 : *variance*

π : *proportion*

• les réalisations d'échantillon seront notées par des lettres latines minuscules

$x_1, ..., x_n$: *valeur de l'échantillon*

$\bar{x}$: *moyenne de l'échantillon*

s : *écart − type de l'échantillon*

s^2 : *variance de l'échantillon*

p : *proportion dans l'échantillon*

• les estimateurs (v.a. ou statistiques) seront notés par des majuscules

$\bar{X}$

S^2

F

Intervalle de confiance pour une moyenne

a) cas où n, la taille de l'échantillon, est petite $n < 30$
On suppose que $X \backsim \mathcal{N}(\mu, \sigma)$.
On distingue deux cas σ connu et σ inconnu.

a-1) σ **connu**

- $\bar{X} \backsim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ ou bien $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \backsim \mathcal{N}(0,1)$

- **Conclusion** : si $\bar{x}$ est une réalisation de $\bar{X}$, l'intervalle de confiance de μ de seuil α est

$$IC = \left[\bar{x} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Intervalle de confiance pour une moyenne

- **exemple :** $n = 15$,$\sigma = 3.75$,$\alpha = 5\%$, $\sum_{i=1}^{15} x_i = 2400$ alors $\bar{x} = \frac{2400}{15} = 160$
$Z_{1-\frac{\alpha}{2}} = 1.96$ car $P(Z < -1.96) = 0.025$
on suppose X gaussienne et on obtient l'intervalle de confiance :
$IC = \left[160 - 1,96\frac{3,75}{\sqrt{15}}; 160 + 1,96\frac{3,75}{\sqrt{15}}\right] = [158,10; 161,90]$

Intervalle de confiance pour une moyenne

a-2) σ inconnu

$\frac{\overline{X}-\mu}{S\sqrt{n-1}} \backsim t_{n-1}$

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

$$IC = \left[\bar{x} - t_{n-1\left(1-\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n-1}}; \bar{x} + t_{n-1\left(1-\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n-1}}\right]$$

Intervalle de confiance pour une moyenne

b) cas où n, la taille de l'échantillon, est grande $n > 30$
Il n'est plus nécessaire de supposer que X est Gaussienne.

b-1) σ **connu**

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$

Si $\bar{x}$ est une réalisation de $\bar{X}$ et si s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$IC = \left[\bar{x} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Intervalle de confiance pour une moyenne

b-2) σ inconnu

$\frac{\bar{X}-\mu}{S/\sqrt{n}} \backsim \mathcal{N}(0,1)$

si $\bar{x}$ est une réalisation de $\bar{X}$ et s une réalisation de S, l'intervalle de confiance de μ de seuil α est

$$IC = \left[\bar{x} - Z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$$

Intervalle de confiance pour une proportion

• on sait que $F = \frac{K}{n}$ est un estimateur de π où π est la proportion de la population possédant le caractère considéré.

$$F \backsim \mathcal{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right) \qquad \text{pour } n\pi, n(1-\pi) > 5$$

ou bien

$$\frac{F-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \backsim \mathcal{N}(0,1) \qquad \text{pour } n\pi, n(1-\pi) > 5$$

• problème : $\pi(1-\pi)$ est inconnu !!!

• **solution 1 :** méthode par estimation de l'écart-type on remplace $\sqrt{\frac{\pi(1-\pi)}{n}}$ par $\sqrt{\frac{f(1-f)}{n}}$ f étant la valeur observée de F (estimation de π) et on a

$$IC = \left[f - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}; f + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}\right]$$

Intervalle de confiance pour une proportion

- **solution 2 :** méthode de l'ellipse (moins classique, mais plus rigoureuse)

$I = [\pi_1, \pi_2]$

La loi de Student à n degrés de liberté (I_n)

La loi de Student à n degrés de liberté $(\mathcal{I}_n)$

• Elle joue un rôle important dans l'estimation par intervalle de confiance. Elle est symétrique, de moyenne nulle et dépend d'un paramètre n appelé nombre de degrés de liberté.

• L'aspect de la courbe variera selon le nombre de degrés de liberté n (de façon générale, elle est plus aplatie que $\mathcal{N}(0,1)$ et quand n augmente $(n > 30)$ les 2 courbes se confondent)

La loi de Student à n degrés de liberté (I_n)

- **définition :** Soient $X \backsim \mathcal{N}(0,1)$ $Y \backsim \mathcal{X}_n^2$ v.a. indépendantes. Alors

$$Z = \frac{X}{\sqrt{Y/n}} \backsim t_n$$

- **remarque :** la fonction densité de probabilité de t_n est

$$f_{t_n}(t) = c_n \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

où c_n sont t.q $\int_{\mathbb{R}} f_{t_n}(t) = 1$

La loi de Student à n degrés de liberté (I_n)

- **Propriétés :**

$$\begin{cases} \text{Si } X \backsim t_n \text{ alors} \\ E(X) = 0 \quad n > 1 \\ V(X) = \frac{n}{n-2}, n > 2 \end{cases}$$

- Convergence de la loi Student vers la loi normale (approximation)

Soit $X \backsim t_n$ alors

$X \to \mathcal{N}(0,1)$en loi quand $n \to \infty$

(en pratique $n > 30$)