Echantillonnage et estimation des paramètres

HDHIRI I.
GM1

2019-2020

Echantillonnage et estimation des paramètres

- Etude Statistique = Etude des caractéristiques (variables statistiques) d'une population.
- L'inférence statistique est définie comme le processus d'utilisation des données d'un échantillon pour estimer ou tester des hypothèses sur les caractéristiques numériques (« paramètres ») d'une population.
- Une population (ou « population mère ») est l'ensemble de tous les éléments d'intérêt dans une étude particulière.
- Un échantillon est un sous-ensemble de la population.

Pourquoi un échantillon? Le recensement de toute la population est coûteux, long, impossible (population infinie), mesures destructrices ..

$\Rightarrow$ On n'étudie qu'une partie de la population : un échantillon. On cherche alors à extrapoler à la population entière les propriétés mises en évidence sur l'échantillon :

Méthode d'échantillonnage aléatoire:principe Soit une population de N unités statistiques (objets, individus) sur laquelle nous désirons prélever un échantillon de taille n. Nous supposons que l'on dispose d'une liste de toutes les unités qui constituent la population, sans omission, ni répétition. Cette liste est la base de sondage. Une façon de construire un échantillon est d'attribuer à chaque unité de la population un numéro unique et prélever ensuite par tirage au sort, n numéros. Les unités correspondantes à ses numéros constituent l'échantillon requis.

Principe de la construction d'un échantillon:
Pour construire un échantillon aléatoire le tirage peut s'effectuer de deux manières:

- Tirage sans remise: les unités tirées ne sont pas remises dans la population. Chaque unité figure au plus une fois dans la population. La composition de la base d'échantillonnage varie à chaque tirage.
- Tirage avec remise: chaque unité tirée au hasard dans la base de sondage est observée puis remise à la population avant qu'une autre unité ne soit tirée. Une unité peut être désignée plusieurs fois. La composition de la base d'échantillonnage est inchangée.

Les méthodes aléatoires : Reposent sur le tirage au hasard d'échantillons et sur le calcul des probabilités.

- Echantillonnage aléatoire simple : On prélève dans la population, des individus au hasard, sans remise
- Echantillonnage aléatoire stratifié : Suppose que la population soit stratifiée, i.e. constituée de sous-populations homogènes, les strates. (ex : stratification par tranche d'age). Dans chaque strate, on fait un échantillonnage aléatoire simple, de taille proportionnelle à la taille de strate dans la population (échantillon représentatif)
- Echantillonnage par grappe : on tire au hasard des grappes ou familles d'individus, et on examine tous les individus de la grappe.

Dans toute la suite du cours, on se place dans le cadre d'un échantillonnage aléatoire simple, sauf mention contraire.

Soit une population de taille N sur laquelle est observée une caractéristique dont on connaît la moyenne μ et la variance σ^2. On supposera que la taille de la population est infinie, ou que le taux de sondage est faible.
Si on prélève n individus dans cette population, on obtient n valeurs $x_1, x_2, ..., x_n$.

- L'observation x_i peut être considérée comme une observation d'une variable aléatoire X_i de même loi que X;

Definition

- Les v.a. $(X_1, X_2, ..., X_n)$ sont indépendantes et de même loi. Elles constituent un **échantillon**
- Toute application définie sur l'échantillon est appelée statistique

Exemples de statistiques:

- Moyenne d'échantillon : $\overline{X}_n = \frac{1}{n}(X_1 + ... + X_n)$
- Variance de l'échantillon: $\Sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$
- Variance corrigée de l'échantillon: $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$

Remarque

Il ne faut pas confondre ces statistiques qui sont des v.a., donc des applications avec les valeurs prises par ces applications sur un ensemble de n individus qui sont des valeurs numériques.

Paramètres de la distribution de $\overline{X}_n$: La moyenne d'échantillon suit une loi de probabilité dont la moyenne est:

$$E(\overline{X}_n) = E(X) = \mu$$

et la variance

$$Var(\overline{X}_n) = \frac{\sigma^2}{n}$$

L'écart-type de la moyenne appelé également erreur-type de la moyenne est donné par

$$\sigma(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$$

.

Théorème central limite: Si des échantillons aléatoires de taille n sont prélevés d'une population infinie dont les éléments possèdent un caractère mesurable X de moyenne $E(X) = \mu$ et de variance $Var(X) = \sigma^2$, alors la distribution de $\overline{X}_n$ tend à se rapprocher d'une loi Normale de moyenne μ et de variance $\frac{\sigma^2}{n}$ ou encore

$$Loi(\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}) \to \mathcal{N}(0,1),$$

et ce d'autant plus que la taille de l'échantillon est grande.

Remarque

- *On peut appliquer le théorème central limite dès que l'échantillon dépasse 30 observations.*
- *Ce théorème est très puissant car il n'impose aucune restriction sur la distribution de X dans la population*
- *Si σ^2 est inconnu, un grand échantillon ($n \geq 30$) permet d'approcher σ^2 par $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2$*

Paramètres de la distribution de S_n^2 On a:

$$E[S_n^2] = \sigma^2; \quad Var(S_n^2) \longrightarrow_{n\to\infty} 0,$$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors

$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\sigma^2} \text{ suit la loi } \chi^2_{n-1}$$

et

$$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \text{ suit la loi } \chi^2_n.$$

Distribution d'échantillonnage d'une proportion: On cherche à étuduer la proportion **p** d'individus possédant un caractère qualitatif donné. La proportion **f** obtenue dans un n-échantillon est la valeur observée d'une variable aléatoire F, appelée proportion d'échantillon.
On a:

$$F = \frac{1}{n}(X_1 + ...X_n);$$

où X_i suivent des lois de Bernouilli de paramètre p D'où

$$E[F] = p, \;\; Var(F) = \frac{p(1-p)}{n}$$

et d'après T.C.L,

$$\sqrt{n}\frac{F-p}{\sqrt{p(1-p)}} \to \mathcal{N}(0,1).$$

Estimation des paramètres

- On s'intéresse à la caractéristique X d'une population, dont la loi dépend d'un paramètre inconnu θ. On note $f_\theta(x)$ la densité de X si X est continue et $P_\theta(X = x); x \in \mathbb{R}$, la loi de X si X est discrète.
- Estimer le paramètre θ consiste à donner une valeur approchée à ce paramètre à partir d'un sondage de la population.
- On dispose d'un sondage de taille n de la population (l'observation de X sur n individus) , noté $(x_1, ...x_n)$ et on note $(X_1, ...X_n)$ l'échantillon aléatoire associé à ce sondage (il s'agit d'un vecteur aléatoire dont une réalisation particulière est $(x_1, ...x_n)$).

Lorsque un paramètre θ d'une population est estimé par un seul nombre, déduit des résultats de l'échantillon, ce nombre est appelé une estimation ponctuelle du paramètre θ.

Definition

- Un estimateur T_n de θ est une statistique de l'échantillon aléatoire $T_n = h(X_1, ..., X_n)$: telle que pour chaque réalisation $(x_1, ...x_n)$ de l'échantillon aléatoire, la valeur $h(x_1, ...x_n)$ prise par T_n approche θ.
- $\widehat{\theta}_n := h(x_1, ...x_n)$ s'appelle une estimation de θ. C'est une réalisation particulière de l'estimateur T_n.

Exemple:

- Un estimateur ponctuel de la moyenne μ d'une population est la moyenne de l'échantillon $\overline{X}_n$.
- Un estimateur ponctuel de la proportion P possèdant un caractère qualitatif, est la proportion F de l'échantillon.

Soit T_n un estimateur de θ

Estimateur convergent: T_n est dit convergent si $\forall \epsilon > 0$,

$$P[|T_n - \theta| > \epsilon] \rightarrow_{n \rightarrow \infty} 0.$$

D'après l'inégalité de Markov, on a si $E[T_n] \rightarrow_{n \rightarrow \infty} \theta$ alors T_n est convergent

Ecart quadratique: $E[|T_n - \theta|^2] = Var(T_n) + \underbrace{[E(T_n) - \theta]^2}_{\text{Biais}}$

Un estimateur ponctuel doit posséder certaines qualités pour fournir des bonnes estimation, nous le résumons comme suit.

- **Estimateur non biaisé:** un estimateur T_n de θ est dit non biaisé ou sans biais si $E(\widehat{T_n}) = \theta$.
  Exemple:
  - $\overline{X}_n$ est un estimateur sans biais de μ: $E[\overline{X}_n] = \mu$
  - Σ^2 est un estimateur biaisé de σ^2: $E[\Sigma^2] = \frac{n-1}{n}\sigma^2$
- **Estimateur efficace:** Le choix parmi plusieurs estimateurs sans biais s'effectue en comparant les variances des estimateurs. Un estimateur sans biais mais de variance élevée peut fournir des estimations très éloignées de la vraie valeur. Un estimateur sans biais est plus efficace si sa variance est la plus faible parmi celles des autres estimateurs sans biais.

Estimation par intervalle de confiance:

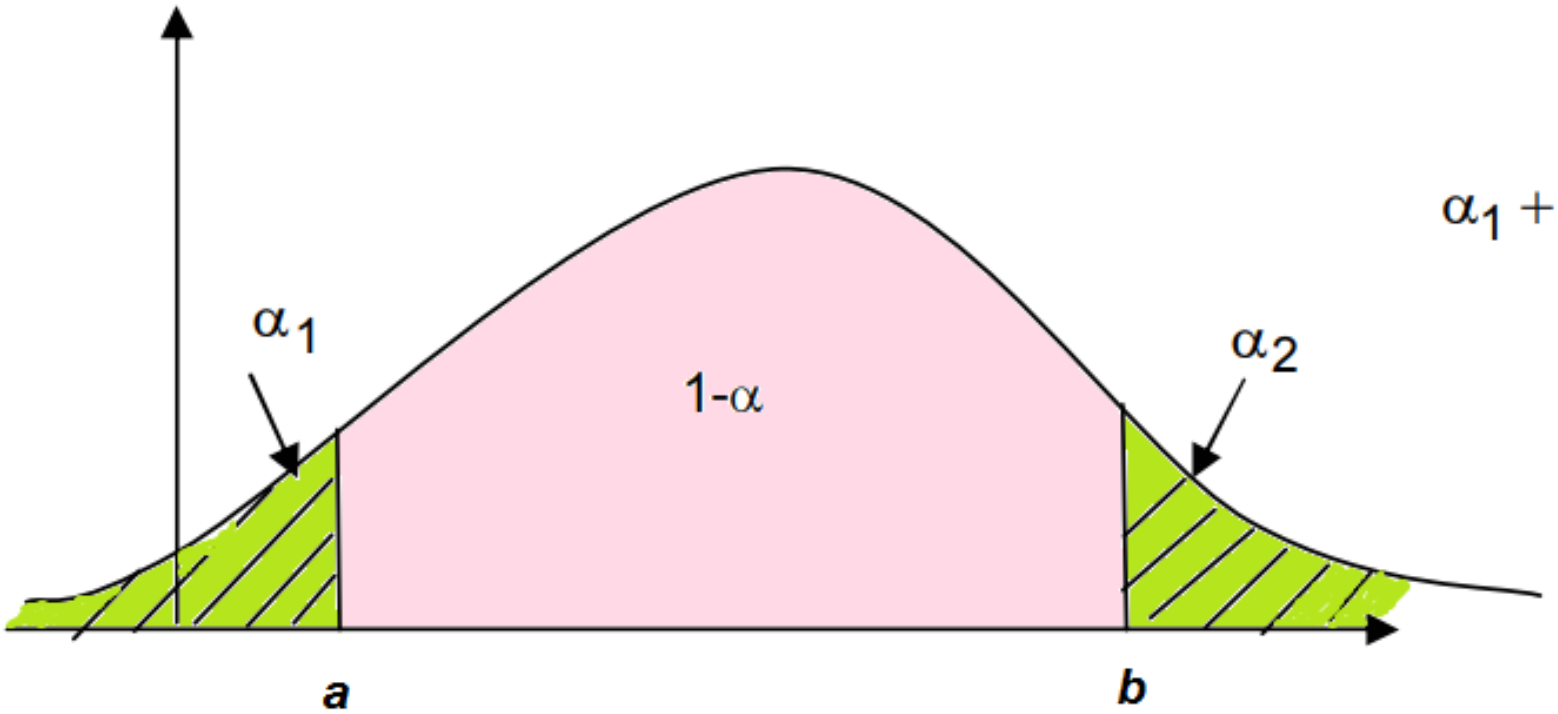
L'estimation par intervalle de confiance d'un paramètre inconnu consiste à calculer à partir d'un estimateur choisi, un intervalle dans lequel il est vraisemblable que la valeur correspondante du paramètre s'y trouve. L'intervalle de confiance est défini par deux limites auxquelles est associée une certaine probabilité, fixée à l'avance.

Il s'agit de construire une « fourchette de valeurs numériques permettant de situer » θ avec une probabilité $1-\alpha$.

$$P[a < \theta < b] = 1 - \alpha.$$

$1-\alpha$ est dit degré de confiance.
La démarche comprend deux étapes :

- avant le tirage d'un échantillon de taille n, un estimateur T a été choisi et la loi de probabilité de T permet de construire un intervalle aléatoire susceptible de contenir la valeur du paramètre θ avec une probabilité $1-\alpha$ fixée a priori.
- après le tirage, la valeur particulière t T calculée à partir des données de l'échantillon permet de déterminer les bornes de l'intervalle de confiance recherché



$$\alpha_1 + \alpha_2 = \alpha$$

On parlera :

- d'intervalle bilatéral symétrique si : $\alpha_1 = \alpha_2 = \alpha/2$
- d'intervalle bilatéral si $\alpha_1 > 0$ $\alpha_2 > 0$
- d'intervalle unilatéral à gauche si $\alpha_2 = 0$
- d'intervalle unilatéral à droite si $\alpha_1 = 0$

Lorsque l'estimateur est sans biais, il est naturel de construire un intervalle centré sur l'estimation ponctuelle obtenue pour θ.

Estimation de la moyenne d'une population normale de variance connue: A partir d'un échantillon aléatoire de taille n d'une population normale de variance connue σ^2. On a $\frac{\overline{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}} \backsim \mathcal{N}(0,1)$

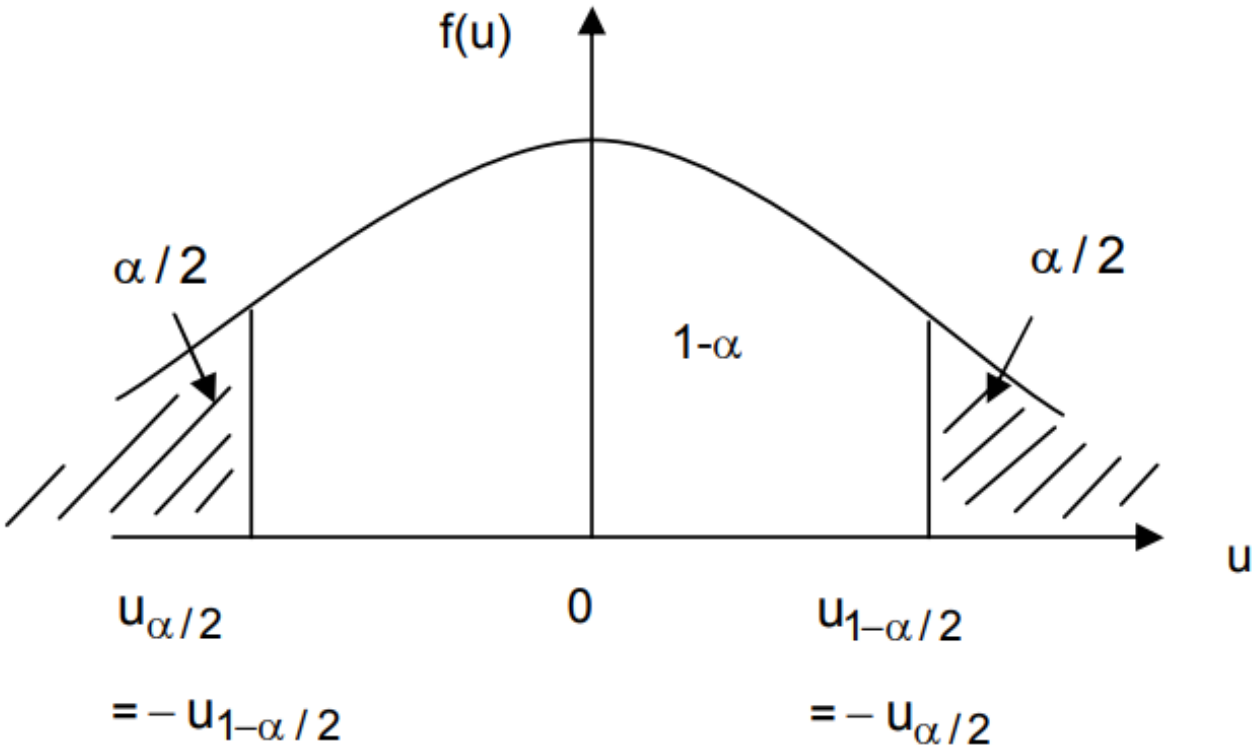
$$P\Big[\mu \in]\overline{X}_n + u_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}}), \overline{X}_n - u_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})[\Big] = 1-\alpha.$$

Ainsi, on définit un intervalle de confiance ayant un degré de confiance $(1-\alpha)$ de contenir la vraie valeur de μ comme suit:

$$\overline{X}_n + u_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}}) < \mu < \overline{X}_n - u_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}}).$$

La marge d'erreur est $|\overline{X}_n - \mu| < |u_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})|$

$\mathcal{N}(0,1)$



$$u_{\alpha/2} = -u_{1-\alpha/2}$$

Application: On cherche un intervalle de confiance bilatéral symétrique de degré de confiance 95% de la moyenne μ d'une population de loi normale dont l'écart type est $\sigma = 2$. On prélève dans cette population un échantillon de taille $n = 100$ et on calcule sa moyenne $\overline{x}_n = 10$.
On a $\alpha = 0.05$, d'où $u_{\frac{\alpha}{2}} = -1.96$. D'où

$$P\Big[\mu \in]\overline{X}_n - 1.96(\frac{\sigma}{\sqrt{n}}), \overline{X}_n + 1.96(\frac{\sigma}{\sqrt{n}})[\Big] = 1 - \alpha.$$

Ainsi l'intervalle de confiance de μ de degré de confiance 95% est:
$]10 - 1.96(\sqrt{\frac{2}{100}}), 10 + 1.96(\sqrt{\frac{2}{100}})[=]9.61, 10.39[$.

Estimation par intervalle de confiance de la moyenne d'une population lorsque la variance de la population est inconnue:

On a dans ce cas, $\frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n-1}}}$ suit la loi de Student $\mathcal{T}(n-1)$. D'où

$$P\left[\mu \in [\overline{X}_n + t_{\frac{\alpha}{2}}(\frac{S}{\sqrt{n-1}}), \overline{X}_n - t_{\frac{\alpha}{2}}(\frac{S}{\sqrt{n-1}})]\right] = 1 - \alpha.$$

où $t_{\frac{\alpha}{2}}$ est lue sur la table de Student.