Université de Lille

Estimation ponctuelle et intervalles de confiance

G. Marot-Briend
guillemette.marot@univ-lille.fr

2021-2022

Rappels

Etapes d'une étude statistique

- collecte des données issues de l'observation ou de l'expérimentation
- analyse statistique
  - analyse descriptive : résumer et présenter les données observées
  - inférence : étendre ou généraliser les conclusions obtenues

Introduction

L'inférence statistique traite principalement de deux types de problèmes :

- l'estimation des paramètres
- les tests d'hypothèse

Inférence statistique

Tirer des conclusions sur une population (grand nombre d'individus) sur la base des observations réalisées sur un échantillon, représentant une portion restreinte de la population.

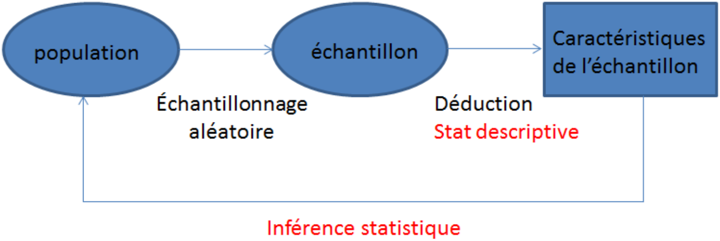
$\Rightarrow$ L'inférence statistique ne conduit jamais à une conclusion stricte, elle attache toujours une probabilité à cette conclusion.

Estimation

L'estimation a pour objectif de déterminer les valeurs inconnues des paramètres de la population (π, μ, σ^2) ou (proportion, moyenne, variance) à partir des données de l'échantillon (p ou f, $\bar{x}$, s^2_{ech}).

La précision de ces estimations est déterminée en établissant un intervalle de confiance autour de ces valeurs prédites.

Estimation



Hypothèses d'échantillonnage :

- chaque individu a la même probabilité d'appartenir à un échantillon
- les n tirages sont indépendants

Plan

1. Distribution d'échantillonnage
2. Estimateurs
3. Intervalles de confiance usuels

Estimation

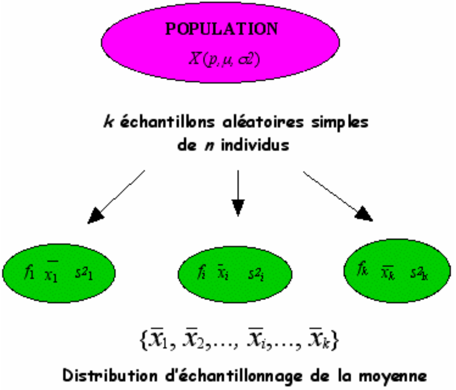
Trois concepts différents à distinguer en théorie de l'estimation :

- les paramètres de la population comme la moyenne μ dont la valeur est inconnue et certaine
  $\Rightarrow$ symbolisés par des lettres grecques
- les résultats de l'échantillonnage comme la moyenne $\bar{x}$ dont la valeur est connue et certaine
  $\Rightarrow$ symbolisés par des minuscules (cf. stat desc.)
- les variables aléatoires des paramètres, comme la moyenne aléatoire $\bar{X}$ dont la valeur est incertaine puisqu'aléatoire mais dont la loi de probabilité est souvent connue
  $\Rightarrow$ symbolisés par des majuscules

Distribution d'échantillonnage

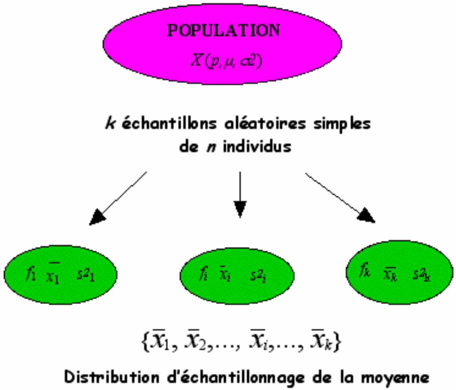
A partir d'une population (π, μ, σ^2), on tire k échantillons aléatoires de même effectif n.

$\Rightarrow$ On obtient alors k estimations du paramètre étudié (ex : k valeurs de moyennes observées).



D. Mouchiroud, 2002, MathSV, UCBL

Distribution d'échantillonnage

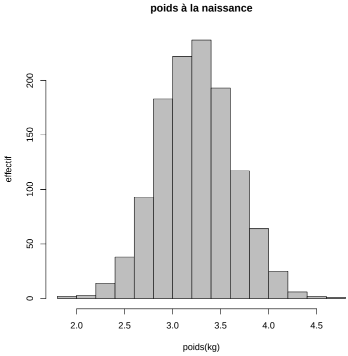


La distribution associée à ces k estimations constitue la distribution d'échantillonnage du paramètre.

<u>Ex :</u> distribution de la moyenne aléatoire $\bar{X}$ (variable aléatoire).

Distribution d'échantillonnage

Exemple : On a mesuré les poids de 1200 bébés à la naissance.

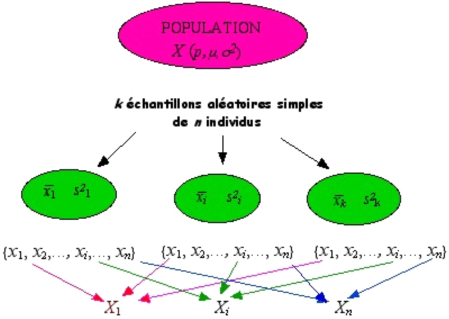


120 premiers poids :
3.1 3.0 3.0 2.9 3.7 3.2 3.6 4.1 3.8
3.1 3.7 3.0 3.9 3.9 3.4 3.6 3.2 3.9
3.8 3.4 2.4 3.5 3.0 3.3 4.0 3.5 3.4
3.9 3.5 3.6 3.0 3.7 3.3 2.6 3.3 3.5
3.3 2.8 3.1 2.9 3.4 3.2 3.3 2.9 2.7
3.0 3.3 2.6 3.3 4.1 3.6 3.1 4.0 2.9
3.4 3.6 3.2 3.4 2.9 3.1 3.9 3.5 3.3
3.3 2.8 3.0 3.7 3.3 2.9 3.6 3.0 2.7
3.0 3.4 3.4 3.9 3.7 3.3 2.9 3.4 3.5
3.8 3.2 3.3 3.0 3.5 3.5 3.0 3.0 3.1
3.8 3.3 3.9 3.4 3.3 3.2 2.7 3.0 3.5
2.8 3.1 3.5 3.4 3.4 3.7 3.7 3.2 3.6
3.5 3.6 3.6 2.8 3.2 3.4 3.6 3.8 3.0
2.7 3.0 3.3

Moyenne de la population : 3,3 kg
Ecart-type de la population : 0,4 kg

Distribution d'échantillonnage

En pratique, quand on étudie un paramètre donné d'une population, on regarde un seul échantillon.



La valeur prise par le 1$^{\text{er}}$ élément extrait de la population dépend de l'échantillon obtenu lors du tirage aléatoire. Cette valeur sera différente si on considère un autre échantillon.

Distribution d'échantillonnage

Distribution de la moyenne aléatoire

$$\bar{X} = \frac{1}{n}\sum X_i$$

Espérance :

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{1}{n}\sum X_i\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum X_i\right) \\
&= \frac{1}{n}\sum \mathrm{E}(X_i) \\
&= \frac{1}{n} n\mu \\
\mathrm{E}(\bar{X}) &= \mu
\end{aligned}$$

Distribution de la moyenne aléatoire

Variance :

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum X_i\right) \\
&= \frac{1}{n^2}\sum \mathrm{Var}(X_i) \text{ car indépendance des } X_i \\
&= \frac{1}{n^2} n\sigma^2 \\
\mathrm{Var}(\bar{X}) &= \frac{\sigma^2}{n}
\end{aligned}$$

$\Rightarrow$ la loi de probabilité de la variable aléatoire $\bar{X}$, moyenne de n va X de lois de proba $\mathcal{N}(\mu, \sigma)$ est une $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

Plan

Estimateurs

Soient X_1,X_2,...,X_n indépendantes et identiquement distribuées (iid) et θ un paramètre associé à cette loi de probabilité.

On appelle estimateur T de θ toute v.a. fonction des X_i

$$T = f(X_1, X_2, \ldots, X_n)$$

Si on considère n observations x_1,x_2,...,x_n, l'estimateur T fournit une estimation

$$\widehat{\theta} = f(x_1, x_2, \ldots, x_n)$$

Propriété des estimateurs

- Convergence : T est convergent si

$$\lim_{n \to \infty} T = \theta$$

- Biais d'un estimateur : $B(T) = \mathrm{E}(T - \theta)$
  Pour avoir un bon estimateur, la différence moyenne entre sa valeur et celle du paramètre qu'il estime doit être nulle
  $\Rightarrow B(T) = 0$
  Un estimateur est sans biais (ESB) si son espérance est égale à la valeur du paramètre de la population $\mathrm{E}(T) = \theta$

Propriétés des estimateurs

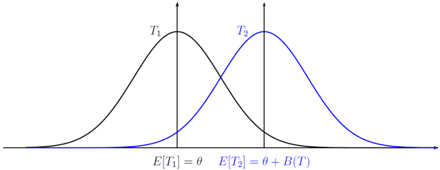
Exemples :

- $\frac{1}{n-1}\sum X_i$ et $\frac{1}{n}\sum X_i$ sont des estimateurs de μ convergents
- $\bar{X}$ est un ESB de μ

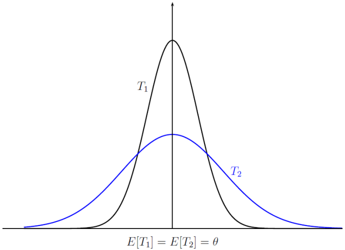
Remarques :

- Si deux estimateurs sont convergents et sans biais, le plus efficace est celui qui a la variance la plus faible.
- On peut préférer un estimateur biaisé qui a une faible variance plutôt qu'un estimateur non biaisé.

Propriétés des estimateurs



T1 est sans biais, T2 est biaisé



T1 est plus efficace que T2

Estimation ponctuelle

L'estimation d'un paramètre quelconque θ est ponctuelle si l'on associe une seule valeur à l'estimateur à partir des données observées sur un échantillon aléatoire.

Théorème :

- $\bar{X}$ est le meilleur estimateur de μ
- $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ (variance observée) n'est le meilleur estimateur de σ^2 que si μ est connue.

Estimation ponctuelle

L'estimation d'un paramètre quelconque θ est ponctuelle si l'on associe une seule valeur à l'estimateur à partir des données observées sur un échantillon aléatoire.

Théorème :

- $\bar{X}$ est le meilleur estimateur de μ
- $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ (variance observée) n'est le meilleur estimateur de σ^2 que si μ est connue.

Exercice : Montrer que $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ est biaisé.

Estimation ponctuelle

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

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 + \bar{X}^2 - 2X_i\bar{X}) \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 + \frac{1}{n}\sum_{i=1}^{n}\bar{X}^2 - \frac{1}{n}\sum_{i=1}^{n}2X_i\bar{X} \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 + \frac{1}{n}\bar{X}^2\sum_{i=1}^{n}1 - 2\bar{X}\frac{1}{n}\sum_{i=1}^{n}X_i \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 + \frac{1}{n}n\bar{X}^2 - 2\bar{X}\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2
\end{aligned}$$

Estimation ponctuelle

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

Par définition,

$$\begin{cases} \sigma^2 = E(X^2) - (E(X))^2 = E(X^2) - \mu^2 \\ \mathrm{Var}(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 = E(\bar{X}^2) - \mu^2 \end{cases}$$

donc

$$\begin{cases} E(X^2) = \sigma^2 + \mu^2 \\ E(\bar{X}^2) = \dfrac{\sigma^2}{n} + \mu^2 \end{cases}$$

Estimation ponctuelle

Finalement,

$$
\begin{aligned}
\mathrm{E}(\widehat{\sigma}^2) &= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - (\frac{\sigma^2}{n}+\mu^2) \\
&= \sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

Donc $\widehat{\sigma}^2$ est biaisé.

Estimation ponctuelle

Finalement,

$$\begin{aligned} E(\widehat{\sigma}^2) &= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - (\frac{\sigma^2}{n}+\mu^2) \\ &= \sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2 \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

Donc $\widehat{\sigma}^2$ est biaisé.

En pratique, μ est souvent inconnue, le meilleur estimateur de σ^2 (ESB) est alors :

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2$$

Estimation ponctuelle

Vocabulaire :

- écart-type de l'échantillon

$$s_{\text{ech}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- déviation standard (anglicisme)

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- erreur standard de la moyenne

$$\text{esm} = \frac{\sigma}{\sqrt{n}}$$

Estimation par intervalle

L'estimation par intervalle associe à un échantillon aléatoire un intervalle $[\widehat{\theta_1}, \widehat{\theta_2}]$ qui recouvre θ avec une certaine probabilité.
Cet intervalle est appelé intervalle de confiance.

On appelle risque d'erreur la probabilité α que l'intervalle de confiance ne contienne pas la vraie valeur du paramètre.
On appelle niveau de confiance la probabilité $1 - \alpha$ que l'intervalle de confiance contienne la vraie valeur du paramètre.

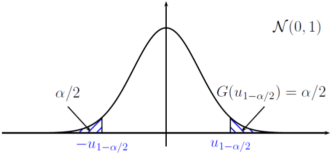
$$P(\widehat{\theta_1} < \theta < \widehat{\theta_2}) = 1 - \alpha$$

Estimation par intervalle

Exemple :

On cherche l'intervalle de confiance à 99% du poids des bébés.
Rappel : $X \sim \mathcal{N}(3,3; 0,6)$

On veut $P(a < X < b) = 0.99$.
En centrant réduisant, on cherche a* et b* tels que
$P(a^* < X^* < b^*) = 0.99$
$\Rightarrow a^* = -u_{1-0.01/2}; b^* = u_{1-0.01/2}$



$$-u_{1-\alpha/2} < \frac{X-\mu}{\sigma} < u_{1-\alpha/2}$$
$$\mu - u_{1-\alpha/2}\sigma < X < \mu + u_{1-\alpha/2}\sigma$$
$$3,3 - 2,5758 * 0,6 < X < 3,3 + 2,5758 * 0,6$$
$$IC_{99\%} = [1,8; 4,8]$$

Plan

Intervalles de confiance

Intervalles de confiance pour la moyenne μ

- Petits échantillons

Théorème :

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

1. $$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

2. $$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \mathcal{T}(n - 1 \text{ ddl})$$

Intervalles de confiance

Application : $n \leq 30$ petits échantillons

Intervalles de confiance d'une moyenne

- si σ est connu (rare)

$$IC(\mu) = \left[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- si σ est inconnu

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

Intervalles de confiance

Démonstration :

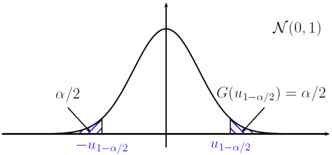
On cherche $(\hat{\mu_1},\hat{\mu_2})$ tel que $P(\hat{\mu_1} < \mu < \hat{\mu_2}) = 1 - \alpha$ (définition IC).
On suppose σ connu donc $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ (théorème).

Définissons $\bar{X}^* = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

$\rightarrow \bar{X}^* \sim \mathcal{N}(0,1)$

$u_{1-\alpha/2}$ est tel que

$P(\bar{X}^* < u_{1-\alpha/2}) = 1 - \dfrac{\alpha}{2}$



Par symétrie de la courbe, $P(-u_{1-\alpha/2} < \bar{X}^* < u_{1-\alpha/2}) = 1 - \alpha$

$$-u_{1-\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < u_{1-\alpha/2}$$

$$-\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Intervalles de confiance

- Grands échantillons

Théorème :

Soit X une v.a. qui suit une loi quelconque

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

$$n \to \infty$$

C'est le Théorème Central Limite qui garantit que la formule est valable quelle que soit la loi !

Intervalles de confiance

Application : $n > 30$ grands échantillons

Intervalles de confiance d'une moyenne

- si σ est connu (rare)

$$IC(\mu) = \left[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} ; \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- si σ est inconnu

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

Intervalles de confiance

Intervalles de confiance d'une proportion π

Soit K une v.a. discrète suivant une loi binomiale $\mathcal{B}(n, \pi)$ pour laquelle on souhaite estimer π.

Rappel loi des grands nombres : la fréquence observée du nombre de succès dans un échantillon de taille n constitue le meilleur estimateur de π :

$$F = \frac{K}{n}$$

Théorème :

Si $n\pi \geq 10$ et $n(1-\pi) \geq 10$ alors

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

Intervalles de confiance

Application

$$\widehat{\pi} = \frac{k}{n}$$

Intervalle de confiance d'une proportion

si $n\widehat{\pi} \geq 10$ et $n(1-\widehat{\pi}) \geq 10$, alors

$$\mathrm{IC}(\pi) = \left[\widehat{\pi} - u_{1-\alpha/2}\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}; \widehat{\pi} + u_{1-\alpha/2}\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}\right]$$