

Estimation

I Estimation ponctuelle

Soit Λ une population suppose très grande

on veut estimer une grandeur X de cette population

on prélève une petite échantillon de n individus
 $w_1, \dots, w_n$ de manière aléatoire

Soit $X_1 = X(w_1)$, $\dots$, $X_n = X(w_n)$

la moyenne empirique $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$

$E(\bar{X}_n) =$ moyenne de X

on dit que $\bar{X}_n$ est un estimateur sans biais de la
 moyenne de X

$\text{Var}(\bar{X}_n) = \frac{1}{n} \text{Var} X$

la variance empirique $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

$\lim_{n \rightarrow +\infty} E(S_n^2) = \text{Variance de } X$ mais $E(S_n^2) \neq \text{Variance de } X$

on dit que $E(S_n^2)$ est un estimateur biaisé
 de la variance de X

En effet $E(S_n^2) = \frac{n-1}{n} \text{Variance de } X \quad \sigma^2$ P43

Revenons à notre

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

$$= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X}_n - \mu)^2$$

donc

$$E(S_n^2) = \frac{1}{n} \sum_{i=1}^n E(X_i - \mu)^2 - E(\bar{X}_n - \mu)^2$$

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

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

$$= \frac{1}{n} n \sigma^2 - \frac{1}{n^2} \text{Var}(X_1 + \dots + X_n)$$

indépendant

$$= \sigma^2 - \frac{1}{n^2} n \sigma^2 = \frac{n-1}{n} \sigma^2$$

la variance débiaisée $\tilde{S}_n^2 = \frac{n}{n-1} S_n^2$

$$E(\tilde{S}_n^2) = \frac{n}{n-1} E(S_n^2) = \sigma^2$$

donc $\tilde{S}_n^2$ est un estimateur sans biais de la variance σ^2

Exemple 2.4.3 Luck.

Soit $\mathcal{A}$ = la population des tous les étudiants de l'université. Soit X leur taille.

La mesure des tailles de $n = 10$ étudiants pris au hasard donne

168, 172, 173, 175, 176, 176, 177, 180, 183, 188

$$\bar{X}_{10} = \frac{168 + \dots + 188}{10} = 176,8$$

$$s_n = \sqrt{\frac{(168-176,8)^2 + (172-176,8)^2 + \dots + (188-176,8)^2}{10}} \approx 5,42$$

La taille moyenne des étudiants de l'université est donc estimée à 176,8 tandis que son écart-type est

$$\text{estimé à } \tilde{s}_{10} = \sqrt{\frac{10}{10-1}} \times 5,42 \approx 5,71$$

II Intervalles de confiance de la moyenne

on a vu que la moyenne empirique $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$ est un estimateur sans biais de la moyenne μ de X

Le problème est que sur un échantillon de n individus, on ne connaît pas la différence, l'erreur entre le $\bar{X}_n$ mesuré et μ théorique. Soit $\alpha \in]0, 1[$ donné. Pour estimer mieux, on cherche à connaître un intervalle $]\bar{X}_n - \varepsilon_\alpha, \bar{X}_n + \varepsilon_\alpha[$ tq

$$P(\mu \in]\bar{X}_n - \varepsilon_\alpha, \bar{X}_n + \varepsilon_\alpha[) = 1 - \alpha$$

on dit que $]\bar{X}_n - \varepsilon_\alpha, \bar{X}_n + \varepsilon_\alpha[$ est un intervalle de confiance

1°) cas d'un grand échantillon

on suppose que le nombre d'individus n dans notre échantillon est grand $n \gg 30$.

Quelque soit la loi de la variable aléatoire obtenue en regardant la grandeur X d'un individu pris au hasard dans notre population. D'après le théorème de la limite centrale

$$Y = \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

Y ont une loi normale centrée réduite $N(0,1)$ P46

on peut donc lire sur la table

$$P(|Y| > u_\alpha) = \alpha \quad \text{souvent } \alpha = 0,05$$

c à d $P(|Y| < u_\alpha) = 1 - \alpha \quad u_{0,05} = 1,960$

$$|Y| < u_\alpha \Leftrightarrow |\bar{X}_n - \mu| < \frac{\sigma}{\sqrt{n}} u_\alpha$$

$$\Leftrightarrow -\frac{\sigma}{\sqrt{n}} u_\alpha < \mu - \bar{X}_n < \frac{\sigma}{\sqrt{n}} u_\alpha$$

$$\Leftrightarrow \mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} u_\alpha, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_\alpha \right]$$

donc $P\left(\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} u_\alpha, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_\alpha \right]\right) = 1 - \alpha$

$1,960 \qquad \qquad \qquad 1,960 \qquad \qquad \qquad 0,95$

L'intervalle de confiance est donc

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

Le problème est qu'en pratique, on ne connaît pas σ !

Remplaçons σ par $\tilde{S}_n$ la racine carrée de la variance débiaisée et ça marche ! car

th (Admis)

$$\tilde{Y} = \frac{\bar{X}_n - \mu}{\tilde{S}_n / \sqrt{n}} \quad \text{sont encore une loi normale centrée réduite } N(0,1)$$

Blm Si $n \geq 30$, Pour $\alpha \in]0, 1[$ donné

on lit en la table de la loi $N(0, 1)$

$$P\left(\left|\tilde{Y}\right| \geq u_{\alpha}\right) = \alpha$$

et on en déduit que

$$P\left(\mu \in \left[\bar{X}_n - \frac{\tilde{S}_n}{\sqrt{n}} \mu_{\alpha}, \bar{X}_n + \frac{\tilde{S}_n}{\sqrt{n}} \mu_{\alpha}\right]\right) = 1 - \alpha$$

2°) cas où X suit une loi normale

on suppose que la variable aléatoire obtenue en regardant la grandeur X d'un individu pris au hasard

dans notre population suit une loi normale $N(\mu, \sigma)$

Blm (admis) Saporta p. 45

La somme de deux variables aléatoires indépendantes suivant des lois normales $N(\mu_1, \sigma_1)$ et $N(\mu_2, \sigma_2)$ est une variable aléatoire qui suit une loi normale $N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$

Remarque

Il est clair que si Y suit une loi normale $N(\mu, \sigma)$ alors $a > 0$, $aY + b$ suit une loi normale $N(a\mu + b, a\sigma)$

Preuve

$$\frac{aY + b - E(aY + b)}{\sqrt{\text{Var}(aY + b)}} = \frac{aY + b - (a\mu + b)}{a\sigma} = \frac{Y - \mu}{\sigma}$$

suit une loi normale $N(0, 1)$

P48

Soit n le nombre d'individus n dans notre échantillon
Ici n peut être très petit
thm

$$Y = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \text{ suit encore une loi}$$

normale centrée réduite $N(0, 1)$

Remarque $X_i =$ valeur de X pour le i ème individu
suit la loi normale $N(\mu, \sigma)$

donc $X_1 + \dots + X_n$ suit la loi $N(n\mu, \sqrt{n}\sigma)$
d'après le théorème additif

Donc d'après Remarque

$$\bar{X}_n = \frac{1}{n} (X_1 + \dots + X_n) \text{ suit } N\left(\frac{1}{n} n\mu, \frac{\sqrt{n}\sigma}{n}\right)$$

$$\text{donc } Y = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \text{ suit une loi } N(0, 1) \quad \square$$

En lisant u_α on la table $P(|Y| \geq u_\alpha) = \alpha$
comme avant on trouve que

$$P\left(\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} u_\alpha, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_\alpha\right]\right) = 1 - \alpha$$

Exemple Jerome Pages Statistiques générales pour utilisateurs
1 - Méthodologie
p 28

Sur un échantillon de $n = 4$ parcelles de blé,
on a observé les rendements suivant en quintaux
par hectares $\left\{ \begin{matrix} 79, & 79, & 80, & 82 \\ \underset{\underset{x_1}{\parallel}}{\parallel} & \underset{\underset{x_2}{\parallel}}{\parallel} & \underset{\underset{x_3}{\parallel}}{\parallel} & \underset{\underset{x_4}{\parallel}}{\parallel} \end{matrix} \right\}$ $\bar{x}_4 = \frac{79+79+80+82}{4} = 80$

L'institut technique des Cereales et Fourrages par sa
longue experience sait que le rendement X suit une
loi normale d'écart-type connu $\sigma = 1$

Prendons $\alpha = 5\% = 0,05$ $u_\alpha = u_{0,05} = 1,960$

donc $P\left(\mu \in \left[\bar{x}_4 - \frac{6}{\sqrt{4}} u_{0,05}, \bar{x}_4 + \frac{6}{\sqrt{4}} u_{0,05}\right]\right) = 0,95$

cà d $P\left(\mu \in \left[80 - \frac{1,96}{2}, 80 + \frac{1,96}{2}\right]\right) = 0,95$

La moyenne μ des rendements $\in [79,02; 80,98]$

avec 95 % de chances

Souvent en pratique, on ne connaît pas σ !

Remplaçons σ par $\tilde{s}_n$ la racine carrée de la variance
échantillonnée et tout change! car

Eh (admis)

$\tilde{Y} = \frac{\bar{x}_n - \mu}{\tilde{s}_n / \sqrt{n}}$ suit une loi de Student

à $n-1$ degrés de liberté

on lit sur une table le réel t_α

$$t_q \quad P(|\tilde{Y}| \geq t_\alpha) = \alpha$$

et on en déduit que

$$P\left(\mu \in \left[\bar{X}_n - \frac{\tilde{S}_n}{\sqrt{n}} t_\alpha, \bar{X}_n + \frac{\tilde{S}_n}{\sqrt{n}} t_\alpha\right]\right)$$

Exemple Pages p 30

on reprend les rendements des $n=4$ parcelles de blé
sans supposer que σ est connu

on lit sur la table $P(|\tilde{Y}| \geq t_\alpha) = 0,05$
pour $t_\alpha = 3,182$ pour $4-1$ degrés de
liberté

$$\bar{X}_4 = 80 \quad S_4^2 = \frac{1^2 + 1^2 + 2^2}{4} = \frac{6}{4}$$

$$S_4^2 = \frac{n-1}{n} \text{Var } X \quad S_4^2 = \frac{n}{n-1} S_4^2 = \frac{4}{4-1} \frac{6}{4} = 2$$

donc $S_4^2 = 2$ variance débiaisée

$$P\left(\mu \in \left[\underbrace{80 - \frac{\sqrt{2}}{\sqrt{4}}}_{77,75} 3,182, \underbrace{80 + \frac{\sqrt{2}}{\sqrt{4}}}_{82,25} 3,182\right]\right) = 0,95$$

Il y a 95% de chances que la moyenne μ des
rendements $\in]77,75, 82,25[$

Remarque $t_\alpha \gg u_\alpha$ toujours quand n augmente t_α tend vers u_α
la loi de Student converge vers $\mathcal{N}(0,1)$

III ^{probabilité} fréquence d'un caractère ^{proportion}

P.51

On s'intéresse à un caractère X dans une population pouvant prendre deux valeurs

Ex 1 fille ou garçon

Ex 2 a l'intention de voter pour Sarkozy
ou n'a pas l'intention de voter pour Sarkozy

on va dire que $X = 1$ si l'individu est une fille
et $X = 0$ si l'individu est un garçon
on s'est donc ramené à une grandeur X
possédant deux valeurs

La variable aléatoire obtenue en regardant le caractère X
d'un individu pris au hasard suit une loi
de Bernoulli de paramètre p
où p et $1-p$ sont les proportions du caractère.

$E(\bar{X}_n) =$ moyenne μ de X ^{ici} $= p$
est un estimateur sans biais de p

$$\text{Var}(\bar{X}_n) = \frac{1}{n} \text{Var} X \stackrel{\text{ici}}{=} \frac{1}{n} p(1-p)$$

$X_1 + X_2 + \dots + X_n$ suit une loi binomiale de paramètre n et p

Supposons que $n \gg 30$, $np \gg 5$ et $n(1-p) \gg 5$

alors on a n qui approximativement grâce au lemme de limite centrale

$X_1 + \dots + X_n$ suit une loi normale $N(np, \sqrt{np(1-p)})$

$$\text{Autrement dit } Y = \frac{X_1 + \dots + X_n - np}{\sqrt{np(1-p)}} = \frac{\bar{X}_n - \mu}{\sqrt{\frac{p(1-p)}{n}}}$$

sont encore une loi normale centrée réduite $N(0,1)$

En lisant u_α sur une table de la loi $N(0,1)$
comme avant

$$P \left(\mu \in \left[\bar{X}_n - \frac{\sqrt{\mu(1-\mu)}}{\sqrt{n}} u_\alpha, \bar{X}_n + \frac{\sqrt{\mu(1-\mu)}}{\sqrt{n}} u_\alpha \right] \right) = 1 - \alpha$$

Le problème pour calculer cet intervalle de confiance
est qu'on ne connaît pas μ qu'il s'agit
justement d'estimer !

thém (admis) Saporta p. 306

La formule pour l'intervalle de confiance reste valable
lorsqu'on remplace μ par son estimateur sans biais :

$$P \left(\mu \in \left[\bar{X}_n - \frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}} u_\alpha, \bar{X}_n + \frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}} u_\alpha \right] \right)$$

Preuve Saporta

$$|\bar{X}_n - \mu| \leq \frac{\sqrt{\mu(1-\mu)}}{\sqrt{n}} u_\alpha$$

$$\Leftrightarrow (\bar{X}_n - \mu)^2 \leq \frac{\mu(1-\mu)}{n} u_\alpha^2$$

on prend une inégalité du 2nd degré d'inconnues μ
puis on approxime.

question pourquoi ne pas appliquer
l'écrit de la même façon ?

super résumé de Rivet