Estimation Par Intervalles

I) INTRODUCTION :

1- Définition

On appelle «intervalle aléatoire» tout intervalle dont une extrémité au moins est une variable aléatoire.

Exemple 1 : Soit un $\chi^2_{(9)}$ l'intervalle $[0, \chi^2_{(9)}]$ est aléatoire. Par exemple la probabilité que le nombre 19 appartienne à cet intervalle est égale à

$$\mathbf{P}\left(19 \in [0, \chi^2_{(9)}]\right) = \mathrm{P}\left(\chi^2_{(9)} \geqslant 19\right) = 0{,}025$$

Exemple 2 Considérons la variable normale $\mathrm{X} = \mathrm{N}(\mu, 1)$ cherchons la probabilité que l'intervalle aléatoire $[\mathrm{X} - 1, \mathrm{X} + 1]$ couvre la valeur fixe μ.

$$\mu \in [\mathrm{X} - 1, \mathrm{X} + 1] \Leftrightarrow (\mu < \mathrm{X} + 1, \mu > \mathrm{X} - 1) \Leftrightarrow$$
$$(\mathrm{X} - \mu > -1, \mathrm{X} - \mu < 1)$$
$$\mathrm{P}(-1 < \mathrm{X} - \mu < 1) = \Phi(1) - \Phi(-1) = 2\,\Phi(1) - 1 = 0{,}682$$

- Définition 2:

Soit θ un paramètre à estimer. Si I est un intervalle aléatoire tel que la probabilité qu'il couvre la valeur θ est égale à $\mathrm{P}(\theta \in \mathrm{I}) = p$ et si un échantillon a conduit à l'observation $\mathrm{I} = \mathrm{I}_0$ (I_0 intervalle numérique). On conviendra de dire que I_0 est un «intervalle de confiance» pour θ ayant un «coefficient de confiance» p.

Pratiquement cela veut dire que sur un grand nombre d'intervalles proposés par le statisticien, une proportion p couvrira le paramètre et une proportion $1-p$ ne le couvrira pas.

Exemples importants :

1) Intervalle de confiance pour la moyenne μ d'une population normale dont la variance σ^2 est connue.

Considérons un échantillon $(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n)$ de la population $\mathrm{N}(\mu, \sigma_0^2)$ on sait que $\overline{\mathrm{X}}$ se distribue selon une Normale de moyenne μ et de variance $\frac{\sigma_0^2}{n}$. Donc $\frac{\overline{\mathrm{X}} - \mu}{\sigma_0/\sqrt{n}}$ est une variable normale centrée et réduite

si a et b sont deux réels tels que $b > a$. Il vient :

$$\mathrm{P}\left(a < \frac{\overline{\mathrm{X}} - \mu}{\sigma_0/\sqrt{n}} < b\right) = \Phi(b) - \Phi(a)$$

Si nous voulons que cette probabilité soit égale à 0,95 par exemple,

c'est-à-dire $\Phi(b) - \Phi(a) = 0{,}95$ nous aurons une infinité de valeurs pour a et b. Par contre si nous voulons un intervalle symétrique, c'est-à-dire $-a = b$, on aura :

$$P\left(-b < \frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} < b\right) = \Phi(b) - \Phi(-b) = 2\Phi(b) - 1 = 0{,}95$$

c'est-à-dire : $\Phi(b) = 0{,}975$ qui donne $b = 1{,}96$

Ainsi la probabilité que $(\overline{X} - \mu)$ tombe entre $-1{,}96\frac{\sigma_0}{\sqrt{n}}$ et $1{,}96\frac{\sigma_0}{\sqrt{n}}$ est 0,95 C'est-à-dire que l'intervalle aléatoire :

$$\left[\overline{X} - 1{,}96\frac{\sigma_0}{\sqrt{n}}, \overline{X} + 1{,}96\frac{\sigma_0}{\sqrt{n}}\right]$$

couvre la moyenne μ avec une probabilité égale à 0,95. Si les observations ont conduit à $\overline{X} = \bar{x}$, nous pouvons dire que l'intervalle numérique $\bar{x} \pm 1{,}96\frac{\sigma_0}{\sqrt{n}}$ est un intervalle de confiance pour μ avec un coefficient de confiance égal à 0,95 (brièvement à 95 %). D'une manière générale, si au lieu de 0,95 nous voulons un coefficient de confiance égal à p, le nombre b sera donné par l'équation

$\Phi^{-1}\left(\frac{1+p}{2}\right) = b$ ce qui fournit tout intervalle désiré.

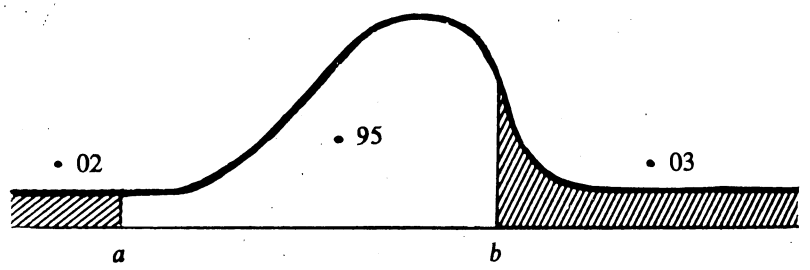
D'autre part, cherchons un intervalle de confiance non sysmètrique avec un coefficient de confiance égal à 95%.

Prenons par exemple a et b tels que $\Phi(a) = 0{,}02$ et $\Phi(b) = 0{,}97$ de façon que $\Phi(b) - \Phi(a) = 0{,}95$

C'est-à-dire $a = -2{,}05$ $b = 1{,}89$ l'intervalle

$$\left[\bar{x} - 1{,}89\frac{\sigma_0}{\sqrt{n}}, \bar{x} + 2{,}05\frac{\sigma_0}{\sqrt{n}}\right]$$ est aussi un intervalle à 95%.

Si nous avons à choisir entre cet intervalle et l'intervalle symètrique, nous choisirons bien sure l'intervalle symétrique car sa longueur $2\frac{\sigma_0}{\sqrt{n}}1{,}96$ est inférieure à celle de l'intervalle non symétrique qui est de $(1{,}89 + 2{,}05)\frac{\sigma_0}{\sqrt{n}}$



Exercice : Montrer par la méthode de LAGRANGE, que parmi tous les intervalles de confiance à 95 %, l'intervalle symétrique est le plus court.

2) Intervalle de confiance pour une moyenne d'une population normale de variance inconnue.

Considérons la quasi-variance $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$

On sait que $$\frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{\frac{\overline{X} - \mu}{\sigma/n}}{s/\sigma} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{(n-1)}}{n-1}}} = t_{(n-1)}$$

car $\frac{n-1}{\sigma^2}s^2$ se distribue selon un $\chi^2_{(n-1)}$ et le numérateur et le dénominateur sont indépendants, ce qui définit la variable de Student.

Soit F la fonction de distribution de t, on a :

$$P\left(a < \frac{\overline{X} - \mu}{s/\sqrt{n}} < b\right) = F(b) - F(a)$$

donc l'intervalle symétrique à 95% pour μ sera : $\bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$

où t est le fractile d'ordre 0,025 de la distribution de Student.

Exercice : si nous utilisons au lieu de s, un estimateur sans biais de σ, l'intervalle trouvé sera t-il plus court ou plus long que l'intervalle ci-dessus ?

3) **Intervalle de confiance pour la variance d'une population normale de moyenne connue μ_0**

On sait que $\frac{\Sigma(X_i - \mu_0)^2}{\sigma^2} = \chi^2_{(n)}$

si χ^2_α est le fractil d'ordre α, on aura :

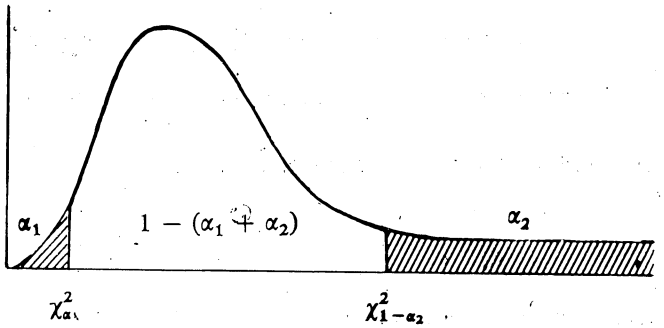
$$P\left(\chi^2_{\alpha_1} < \frac{\Sigma (X_i - \mu_0)^2}{\sigma^2} < \chi^2_{1-\alpha_2}\right) = 1 - (\alpha_1 + \alpha_2)$$

ou $$P\left(\frac{\Sigma (X_i - \mu_0)^2}{\chi^2_{1-\alpha_2}} < \sigma^2 < \frac{\Sigma (X_i - \mu_0)^2}{\chi^2_{\alpha_1}}\right) = 1 - (\alpha_1 + \alpha_2)$$

qui permet d'avoir l'intervalle cherché.
Si nous désirons avoir l'intervalle à 95% par exemple nous choisirons α_1 et α_2 tels que $\alpha_1 + \alpha_2 = 0{,}05$ et à l'aide de la table de $\chi^2_{(n)}$ nous pourrons lire les fractiles appropriés.

4) Intervalle de confiance pour la variance d'une population normale lorsque la moyenne est inconnue μ

La même procédure que dans le cas précédent est applicable sauf que $\frac{\Sigma (X_i - \mu)^2}{\sigma^2}$ se distribue selon un $\chi^2_{(n-1)}$ au lieu d'un χ^2 de degré n.

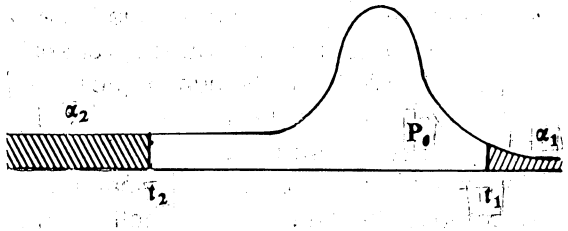


Exercice : Est-il vrai que l'intervalle symétrique $\alpha_1 = \alpha_2$ est le plus court ?

3- Méthode générale de construction des intervalles de confiance.

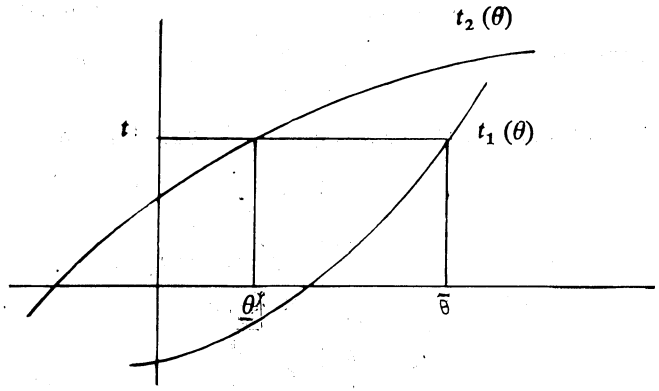
Soit T estimateur ponctuel du paramètre θ pour lequel nous voulons construire un intervalle de confiance ayant pour coefficient de confiance le nombre P. L'intervalle sera basé sur T. Soient α_1 et α_2 deux nombres compris entre 0 et 1 tels que $\alpha_1 + \alpha_2 = (1 - P)$ Il existe deux nombres t_1 et t_2 tels que :

$$\begin{cases} P_\theta (T < t_1) = \alpha_1 \\ P_\theta (T > t_2) = \alpha_2 \end{cases}$$



où P_θ est la loi de probabilité lorsque le paramètre vaut θ, remarquer que $t_1 < t_2$ car $\alpha_1 + \alpha_2 < 1$ et que $P(t_1 \leqslant T \leqslant t_2) = P$.
Il est clair que t_1 et t_2 dépendent de θ. Donc nous avons défini deux fonctions $t_1(\theta)$ et $t_2(\theta)$.

Si ces fonctions sont strictement monotones, la méthode suivante permet d'obtenir un intervalle de confiance pour θ.



Supposons que l'on ait observé la valeur t de la statistique T et supposons pour fixer les idées, que les fonctions $t_1(\theta)$ et $t_2(\theta)$ sont croissantes, on sait laors qu'elles admettent deux fonctions réciproques ce qui permet de déterminer les nombres :
$\underline{\theta} = t_2^{-1}(t)$ et $\overline{\theta} = t_1^{-1}(t)$ (voir figure)
Si le paramètre vaut θ et si t_1 et t_2 désignent les nombres $t_1(\theta)$ et $t_2(\theta)$ on peut alors affirmer que :

$$\theta \in [\underline{\theta}, \overline{\theta}] \Leftrightarrow t \in [t_1, t_2]$$

Par conséquent la probabilité que l'intervalle aléatoire $[\underline{\theta}, \overline{\theta}]$ couvre le paramètre θ est la même que la probabilité que la variable aléatoire T appartienne à l'intervalle numérique $[t_1, t_2]$

C'est-à-dire : $P(\underline{\theta} \leqslant \theta \leqslant \overline{\theta}) = P$

Une fois les valeurs $\underline{\theta}$ et $\overline{\theta}$ sont observées on obtient un intervalle de confiance à 100p % pour le paramètre θ. En pratique il n'est pas nécessaire de connaître les fonctions $t_1(\theta)$ et $t_2(\theta)$, car $\underline{\theta}$ et $\overline{\theta}$ peuvent être calculés à partir des équations :

$P_{\underline{\theta}}(T > t) = \alpha_2$ et $P_{\overline{\theta}}(T < t) = \alpha_1$

Exemple 1 : Application à la construction d'un intervalle de confiance pour la moyenne d'une population normale de variance connue σ_0^2

Prenons par exemple $\alpha_1 = \alpha_2 = 0{,}025$ et comme statistique $T = \overline{X}$

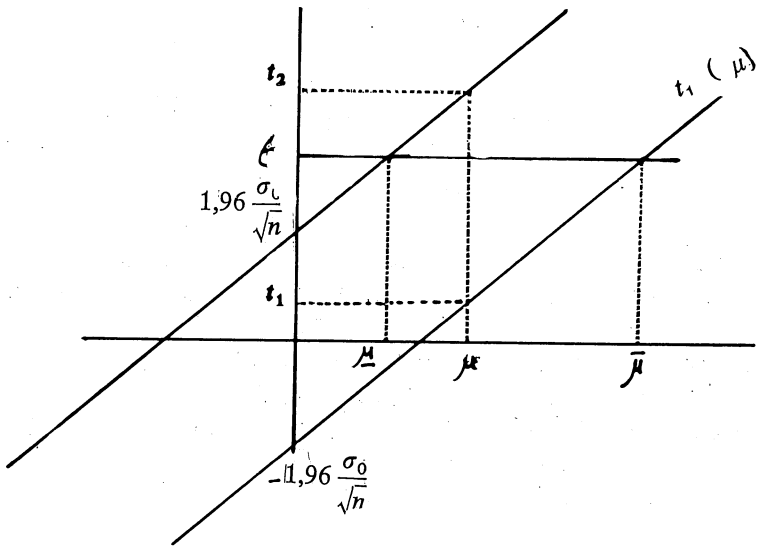
Donc $P_\mu(\overline{X} < t_1) = 0{,}025$, $P_\mu(\overline{X} > t_2) = 0{,}025$

c'est-à-dire

$$\Phi\left(\frac{t_1 - \mu}{\sigma_0/\sqrt{n}}\right) = 0{,}025 \quad , \quad \Phi\left(\frac{t_2 - \mu}{\sigma_0/\sqrt{n}}\right) = 0{,}975$$

ou bien : $t_1 = \mu - 1{,}96\frac{\sigma_0}{\sqrt{n}}$, $t_2 = \mu + 1{,}96\frac{\sigma_0}{\sqrt{n}}$

Dans ce cas les fonctions $t_1(\mu)$ et $t_2(\mu)$ sont linéaires :



On aurait pu écrire directement : $P_{\overline{\mu}}(\overline{X} < \bar{x}) = 0{,}025$ $P_{\underline{\mu}}(\overline{X} > \bar{x}) = 0{,}025$

c'est-à-dire $\frac{\bar{x} - \overline{\mu}}{\sigma_0/\sqrt{n}} = -1{,}96$, $\frac{\bar{x} - \underline{\mu}}{\sigma_0/\sqrt{n}} = 1{,}96$

D'où l'intervalle de confiance :

$$\left[\bar{x} - 1{,}96\frac{\sigma_0}{\sqrt{n}}, \bar{x} + 1{,}96\frac{\sigma_0}{\sqrt{n}}\right]$$

Exemple 2 : Intervalle de confiance pour le paramètre de Bernoulli. Considérons un échantillon $(X_1, X_2, \ldots X_n)$ et prenons la statistique $\Sigma X_i = X$ Les limites supérieures et inférieures $\underline{p}$ et $\overline{p}$ seront tels que $P_{\overline{p}}(X < x) = \alpha_1$ et $P_{\underline{p}}(X > x) = \alpha_2$

c'est-à-dire $\sum_{i=0}^{i=x-1} \binom{n}{i} \overline{p}^i (1 - \overline{p})^{n-i} = \alpha_1$ et $\sum_{i=x+1}^{i=n} \binom{n}{i} \underline{p}^i (1 - \underline{p})^{n-i} = \alpha_2$

En principe, il suffit de lire sur des tables binomiales détaillées la valeur $\overline{p}$ qui correspond à α_1. De même pour $\underline{p}$. Mais il y a parfois une difficulté qui vient du fait que les Σ n'atteignent pas exactement α à cause de la nature discrète des distributions. Dans ce cas nous nous contenterons d'une approximation

En fait, les tables de la fonction Beta (β) incomplète sont plus appropriées et il est préférable de les utiliser surtout lorsque n est assez grand. (Voir chap. 2 de la 4e partie). Signalons qu'il existe des abaques des fonctions $t_1(p)$ et $t_2(p)$ sur lesquelles nous pouvons lire les valeurs $\underline{p}$ et $\overline{p}$ immédiatement.

- QUELQUES REMARQUES IMPORTANTES CONCERNANT L'INTERVALLE DE CONFIANCE.

1) Intervalles approximatifs :

Soit T une statistique dont on connait la loi de probabilité approximative, il est possible de construire un intervalle tel que la probabilité que celui-ci couvre le paramètre se rapproche du coefficient de confiance voulu.

Par exemple s'il s'agit de construire un intervalle de confiance à 95 % pour la moyenne μ d'une population de variance connue σ^2 on peut écrire d'après le théorème limite centrale. que ;

(Voir Chap. VI 2e Partie) :

$$P\left(\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} \leqslant 1{,}96\right) \approx 0{,}95$$

où $\overline{X}$ est la moyenne de l'échantillon.

Donc $\bar{x} \pm 1{,}96\frac{\sigma}{\sqrt{n}}$ est un intervalle de confiance pour μ avec un coefficient de confiance qui se rapproche de 95 % d'autant plus que n est grand.

Comme autre exemple : Construisons un intervalle de confiance pour le paramètre p de Bernoulli sur la base de la fréquence empirique f

On sait par le théorème de la limite centrée que $\frac{f-p}{\sqrt{pq/n}}$ se distribue selon $N(0, 1)$ Donc l'intervalle $f \pm 1{,}96\sqrt{pq/n}$ couvre le paramètre P avec une probabilité approximativement égale à 0,95

Nous n'avons pas encore construit l'intervalle de confiance cherché. Car l'amplitude de l'intervalle ci-dessus dépend elle'même de p, mais comme f est un estimateur convergent de p, l'intervalle de confiance approximatif à 95 % est alors ;

$$f \pm 1{,}96\sqrt{f(1-f)/n}$$

Exercice : Démontrer avec précision ce que nous avons dit ci-dessus à propos du coefficient de confiance approximatif (comment intervient la notion d'estimateur convergent).

2) Intervalles basés sur l'inégalité de TCHEBYCHEV

Soit T une statistique quelconque. On sait d'après l'inégalité de Tchebychev que

$$P(|T - E(T)| \leqslant k \cdot \sigma_T) \geqslant 1 - \frac{1}{k^2}$$

où $E(T)$ et σ_T sont respectivement l'espérance mathématique et l'écart-type de T. k étant un nombre positif.

Si T est un estimateur sans biais de θ pour lequel nous désirons avoir un intervalle de confiance avec un coefficient de confiance égale à P et si on prend $k = \sqrt{\frac{1}{1-P}}$

alors $P(T - k\sigma_T \leqslant \theta \leqslant T + k\sigma_T) \geqslant P$

Donc l'intervalle aléatoire $T \pm k \cdot \sigma_T$ couvrira le paramètre θ avec une probabilité au moins égale à P

D'où un intervalle de confiance ayant un coefficient de confiance au moins égal à P.

Du fait que le coefficient de confiance est supèrieur à celui que l'on désire. il est naturel d'obtenir un intervalle assez long.

Par exemple si nous prenons comme paramètre la moyenne μ d'une population de Variance égale à 1, un intervalle de confiance pour μ sur la moyenne $\overline{X}$ d'un échantillon de taille 4 s'obtient en faisant :

$$P(|\overline{X} - \mu| \leqslant 1) \geqslant 1 - \frac{1}{4} = 0{,}75 \qquad k = 2$$

Donc $\bar{x} \pm 1$ est un intervalle ayant au moins 75 % pour coefficient de confiance.

Si au lieu de l'inégalité de TCHEBYCHEV, nous utilisons la distribution approximative de $\overline{X}$, on aura :

$$P\left(-1{,}15 \times \frac{1}{\sqrt{4}} \leqslant \overline{X} - \mu \leqslant 1{,}15 \cdot \frac{1}{\sqrt{4}}\right) = 0{,}75$$

Car $\Phi(1{,}15) = 0{,}875$ Donc $\Phi(1{,}15) - \Phi(-1{,}15) = 0{,}75$

l'intervalle de confiance pour μ est $\bar{x} \pm 0{,}575$ avec un coefficient de confiance égal à 0,75. Il est naturel que cet intervalle soit plus court que celui de Tchebychev car il ne peut garantir un coeffient de confiance supérieur à 75 %.

3) Tests d'hypothèses et Intervalles de confiance.

La théorie des intervalles de confiance est fortement liée à celle des Tests d'hypothèses Statistiques. Il sera plus à propos après l'étude des tests d'hypothèses de donner quelques élèments qui comptlètent ce que nous avons dit au sujet des intervalles de confiance. Par exemple tout ce qui concerne les intervalles pour les différences entre deux paramètres trouvera une meilleure place dans le prochain chapitre sur la thérie des tests d'hypothèses statistiques.

EXERCICES

1- Soit un $\varepsilon > 0$. Quelle est la taille de l'échantillon qui permet d'avoir un intervalle de confiance à 95 % pour la moyenne d'une population avec une longueur qui ne dépasse pas ε

2- Trouver les fonctions $t_1(p)$ et $t_2(p)$ relatives au paramètre de Bernoulli sachant que : $\alpha_1 = \alpha_2 = 0{,}025$ et $n = 10$

3- Soit I un intervalle de confiance pour le paramètre θ avec un coefficient de confiance égal à 95 %. Pour une suite de 100 estimations de θ , nous proposons l'intervalle I Que peut-on dire de la probabilité de ne pas se tromper plus de trois fois ?

4- Considérons un échantillon prélevé d'une population de Poisson de paramètre θ

Donner un intervalle de confiance pour θ avec un coefficient de confiance égal à 90 %

5- Considérons un échantillon prélevé à partir de deux populations normales $N(\mu_1, 1)$ et $N(\mu_2, 1)$

Comment pouvez-vous construire un intervalle de confiance pour la différence entre les moyennes μ_1 et μ_2

6- Supposons un échantillon tiré d'une population Normale inconnue $N(\mu, \sigma^2)$ Nous savons que l'expression $(\bar{x} - \mu)/s/\sqrt{n}$ se distribue selon la loi t de student. C'est-à-dire que $\mu = \bar{x} - t \cdot s/\sqrt{n}$. . Après avoir observé $\bar{x}$ et s^2 , nous obtenons une transformation linéaire de t La distribution de la nouvelle variable s'appelle «distribution fiduciaire» du paramètre μ. . Quelle est cette distribution et quel est son fractile d'ordre 80 % par exemple ?