Chapitre 3 : Estimation paramétrique par intervalle de confiance

I Activité introductive

• Dans une société musicale qui exécute des œuvres vocales, supposons qu'elle cherche à estimer la moyenne μ_{pop} de 40 chanteurs à partir d'un échantillon de 5 observations de cette chorale. Si on estime μ_{pop} par la moyenne d'échantillon $\mu_{\text{éch}}$, on ne peut pas raisonablement croire que $\mu_{\text{éch}} = \mu_{pop}$ exactement ; on fera une petite erreur d'estimation. D'ailleurs, comme l'échantillon est aléatoire, la valeur de $\mu_{\text{éch}}$ que vous auriez obtenue sur un autre échantillon aurait probablement été différente, quoique tout aussi pertinente. Voici quelques exemples d'échantillons ; on remarque effectivement que la valeur de $\mu_{\text{éch}}$ **fluctue** d'un échantillon à l'autre :

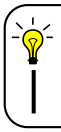
	X_1	X_2	X_3	X_4	X_5	$\mu_{\text{éch}}$
échantillon 1	1.89	1.79	1.74	1.90	1.74	1.812
échantillon 2	1.74	1.95	1.76	1.75	1.71	1.782
échantillon 3	1.84	1.84	1.88	1.85	1.89	1.86
échantillon 4	1.84	1.84	1.75	1.83	1.75	1.802
échantillon 5	1.59	1.68	1.79	1.89	1.79	1.748

• Conclusion : Pour estimer μ_{pop}, vous ne pouvez pas simplement donner la valeur de $\mu_{\text{éch}}$, mais vous devez l'accompagner d'une marge d'erreur. L'objet de ce chapitre est de comprendre comment déterminer ces **marges d'erreur**, ou en termes mathématiques, comment construire un **intervalle de confiance**.

II Principe

• L'estimation d'un paramètre inconnu $\theta \in \mathbb{R}$ par intervalle de confiance consiste à associer à un échantillon un intervalle aléatoire noté I_θ dont, pour un **risque** $\alpha \in [0,1]$ ou bien un **niveau de signification** $1-\alpha$, on a des fortes chances de croire qu'il contient la vraie valeur de θ.

• Mathématiquement, cet intervalle de confiance I_θ pour le paramètre θ, est un intervalle aléatoire de la forme $[a_\alpha, b_\alpha]$, où $(a_\alpha, b_\alpha) \in \mathbb{R}^2 \backslash \{(0,0\}$, défini comme suit :



$$\mathbb{P}(\theta \in I_\theta) = \mathbb{P}(a_\alpha \leq \theta \leq b_\alpha) = 1-\alpha$$

• Le risque α est la probabilité que le paramètre θ n'appartienne pas à l'intervalle I_θ, autrement dit c'est la probabilité que l'on se trompe en affirmant que $\theta \in I_\theta$. C'est donc une probabilité d'erreur qui doit être assez petite. Les valeurs usuelles de α sont 10%, 5%, 1%, ...

• Le niveau de signification $1-\alpha$ est la probabilité que le paramètre θ appartient à l'intervalle I_θ.

• Lorsque le paramètre θ désigne la moyenne m ou bien la proportion p, il semble alors logique de chercher un intervalle de confiance I_θ pour θ de la forme $[\widehat{\theta}-\varepsilon, \widehat{\theta}+\varepsilon]$, où $\widehat{\theta}$ est un estimateur ponctuel sans biais de θ. Selon la caractérisation ci-dessus ceci revient alors à déterminer les **marges d'erreurs** $\varepsilon > 0$ de sorte que :

$$\mathbb{P}(\widehat{\theta}-\varepsilon < \theta < \widehat{\theta}+\varepsilon) = 1-\alpha$$

.

III Construction de l'intervalle de confiance

Pour toute la suite du cours, on considère $(X_1, \ldots, X_n)$, $n > 0$, un échantillon i.i.d de taille n et on se propose de construire, pour un risque α donné, un intervalle de confiance I_θ pour le cas où l'inconnu θ est la moyenne $\mu_{pop} = \mu \in \mathbb{R}$ d'une population, ensuite pour le cas où l'inconnu θ est la variance $\sigma^2_{pop} = \sigma^2 > 0$ d'une population et enfin pour le cas où l'inconnu θ est la proportion $p \in]0,1[$ d'un caractère qualitatif relatif à une population.

III.1 Intervalle de confiance pour la moyenne

• Cas des petits échantillons $n < 30$:

- Soit $(X_1, \ldots, X_n)$, $n > 0$, un n-échantillon de loi **normale** $\mathcal{N}(\mu, \sigma^2)$, où μ est la moyenne et σ^2 est la variance. On considère les estimateurs ponctuels classiques (sans biais et convergents) de μ et de σ^2 respectivement la moyenne empirique et la variance empirique corrigée données par :

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} \quad \text{et} \quad S'^2_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

- L'idée principale de la construction de l'intervalle de confiance I_μ pour μ, avec un risque α fixé, est la suivante :

Soit $Z \sim \mathcal{N}(0,1)$ alors il existe un réel a tel que $\mathbb{P}(Z \geq a) = \frac{\alpha}{2} \Rightarrow a$ est dit le **quantile** de $\mathcal{N}(0,1)$ d'ordre $1-\frac{\alpha}{2}$ qu'on le note dorénavant par $z_{\frac{\alpha}{2}}$ et on le détermine à partir de la table de $\mathcal{N}(0,1)$ (lecture inverse de la table).

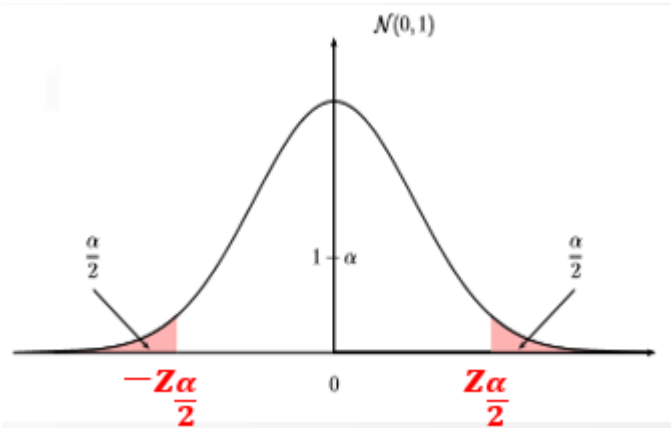
D'où on a :

$$\mathbb{P}(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

- Par la symétrie de $\mathcal{N}(0,1)$, on a aussi :

$$\mathbb{P}(Z \leq -z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

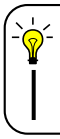
Ceci est illustré graphiquement par :



Ce qui implique que :

$$\begin{aligned}\mathbb{P}(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) &= \mathbb{P}(Z \geq -z_{\frac{\alpha}{2}}) - \mathbb{P}(Z \geq z_{\frac{\alpha}{2}}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\ &= 1 - \alpha\end{aligned}$$

Alors pour construire un intervalle de confiance de μ avec un niveau de confiance $1-\alpha$ fixé, il suffit alors d'utiliser ce résultat qu'on vient d'établir :



$$\mathbb{P}(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

De plus, étant donné que l'échantillon est de loi normale $\mathcal{N}(\mu, \sigma^2)$ alors :

$$\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

et par conséquent l'idée est d'essayer de construire à partir de $\overline{X}_n$ une **variable aléatoire** $\sim \mathcal{N}(0,1)$. C'est pour cela il fallait distinguer le deux situations suivantes : si σ^2 est connue et si σ^2 est inconnue.

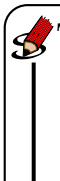
Cas variance σ^2 connue

On obtient que si :

$$\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Leftrightarrow \quad Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

D'après le résultat ci-dessus , on montre que :

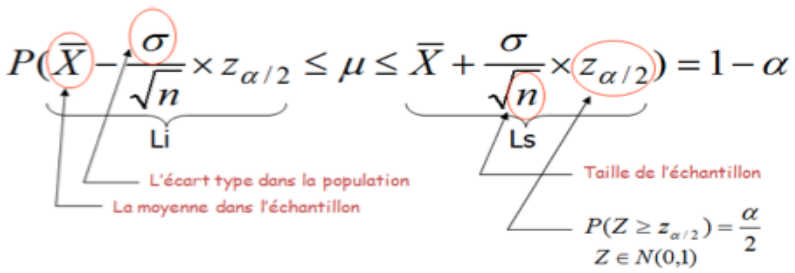
$$\begin{aligned}\mathbb{P}(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) &= 1-\alpha \\ \mathbb{P}(-z_{\frac{\alpha}{2}} < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}) &= 1-\alpha \\ \mathbb{P}(-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \overline{X}_n - \mu < z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) &= 1-\alpha \\ \mathbb{P}(-\overline{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) &= 1-\alpha \\ \mathbb{P}(\overline{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) &= 1-\alpha\end{aligned}$$



Théorème 1

Un intervalle de confiance de seuil α pour le paramètre μ de la loi $\mathcal{N}(\mu, \sigma)$ lorsque σ^2 est connue est donné par :

$$IC(\mu) = [\overline{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$



Exercice 1 *On suppose que le poids d'un nouveau né est une variable aléatoire normale d'écart-type égal à* $0,5$ *kg. Au mois de janvier* 2004 *dans l'hôpital de Charleville-Mézières, on observe* 25 *enfants nés dont le poids moyen* $\overline{x_n} = 3,6$ *kg.*

1. *Déterminer un intervalle de confiance de niveau de confiance* 95% *pour la moyenne* m *du poids d'un nouveau né ?*
2. *Quel serait le nombre d'enfants observés pour que l'intervalle de confiance soit de longueur* $0,1$ *?*

Cas variance σ^2 inconnue

- On se propose de donner un intervalle de confiance de risque α pour μ avec σ^2 est inconnue. Une idée naturelle est alors de remplacer σ^2 par son estimateur ponctuel classique (sans biais et convergent) :

$$S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

et par la suite on définit la variable aléatoire suivante :

$$T = \frac{\overline{X}_n - \mu}{\sqrt{\frac{S_n'^2}{n}}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{S_n'^2}} \sim t(n-1) \quad \text{(attention elle } \textbf{ne suit pas } \mathcal{N}(0,1))$$

où $t(n-1)$ est la loi de Student à **n-1** ddl.

- Grâce à la symétrie de la loi de Student, on obtient le résultat suivant :

$$\mathbb{P}(-t_{\frac{\alpha}{2},n-1} < T < t_{\frac{\alpha}{2},n-1}) = 1-\alpha$$

avec $t_{\frac{\alpha}{2},n-1}$ désigne le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-1$ ddl, et il est déterminé à partir de la table de $t(n-1)$ (lecture inverse de la table).

- Par conséquent :

$$\begin{aligned}
\mathbb{P}\Big(-t_{\frac{\alpha}{2},n-1} < \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{S_n'^2}} < t_{\frac{\alpha}{2},n-1}\Big) &= 1-\alpha \\
\mathbb{P}\Big(\overline{X}_n - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{S_n'^2}{n}} < \mu < \overline{X}_n + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{S_n'^2}{n}}\Big) &= 1-\alpha
\end{aligned}$$

Théorème 2

Un intervalle de confiance de seuil α pour le paramètre μ de la loi $\mathcal{N}(\mu,\sigma^2)$ lorsque σ^2 est inconnue est de la forme :

$$IC(\mu) = \Big[\overline{X}_n - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{S_n'^2}{n}}, \overline{X}_n + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{S_n'^2}{n}}\Big]$$

- **Cas des grands échantillons $n \geq 30$:**

Lorsque l'échantillon **n'est pas de loi normale** mais sa taille $n \geq 30$, alors on obtient les intervalles de confiances asymptotiques suivants :

Théorème 3

Lorsque $n \geq 30$ et σ^2 **connue**, alors le T.C.L nous permet de construire un intervalle de confiance de seuil α pour le paramètre μ sous la forme suivante :

$$IC(\mu) = [\overline{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

Théorème 4

Lorsque $n \geq 30$ et σ^2 **inconnue**, on a $t(n) \approx \mathcal{N}(0,1)$, et par la suite un intervalle de confiance de seuil α pour le paramètre μ est de la forme :

$$IC(\mu) = \left[\overline{X}_n - z_{\frac{\alpha}{2}} \sqrt{\frac{S'^2_n}{n}}, \overline{X}_n + z_{\frac{\alpha}{2}} \sqrt{\frac{S'^2_n}{n}}\right]$$

Exercice 2 *Des tests sur un échantillon de taille* 10 *sur la conductivité thermique d'un métal ont permis d'obtenir les données suivantes :*

41.60	*41.48*	*42.34*	*41.95*	*41.86*	*42.18*	*41.71*	*42.26*	*41.81*	*42.04*

Soit X la conductivité thermique du métal. On suppose que X suit une loi Normale des paramètres inconnus.

1. *Donner un intervalle de confiance de μ de niveau 95%.*
2. *Supposons que $\sigma^2 = 0.3$, déterminer la taille nécessaire de l'échantillon pour construire un intervalle de confiance pour μ de niveau de confiance* 95% *et d'amplitude égale à* 0.06.

III.2 Intervalle de confiance d'une proportion

Le problème connu sous le nom d'intervalle de confiance pour une proportion est en faite le problème de la détermination d'un intervalle de confiance pour le paramètre $p \in]0,1[$ de la loi de Bernoulli au vu d'un échantillon $(X_1, \ldots, X_n) \sim \mathcal{B}(p)$.

De ce fait, une proportion n'est que la fréquence de la valeur 1 dans l'échantillon. On rappelle qu'on a déjà montré qu'un estimateur ponctuel de p est $\widehat{P}_n = \overline{X}_n$, or pour un échantillon qui n'est pas normal, la loi de la statistique $\widehat{P}_n$ n'est pas évident de la trouver et par la suite la détermination de l'intervalle de confiance n'est plus possible, mais en faveur du Théorème Central Limite (T.C.L) lorsque n suffisamment grand, on admet le résultat suivant :

Théorème 4

Si $np > 5$ et $n(1-p) > 5$ (ou n assez grand), alors l'intervalle de confiance de niveau de signification $1-\alpha$ pour une proportion p se présente comme suit :

$$IC(p) = \left[\widehat{P}_n - z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{P}_n(1-\widehat{P}_n)}{n}}, \widehat{P}_n + z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{P}_n(1-\widehat{P}_n)}{n}}\right]$$

Exercice 3 *Sur* 500 *personnes interrogées,* 274 *ont déclaré qu'elles voteraient pour le candidat A.*

1. *Donner une estimation de p la proportion de personnes favorables au candidat A dans la population par intervalle de confiance au niveau de signification* $0,95$.

2. *Pour quel degré de confiance a-t-on la borne inférieure exactement égale à* 50% *?*

III.3 Intervalle de confiance d'une variance

- Le problème est le suivant : il faut encadrer la variance σ^2 de la population qui est inconnue. On recherche donc deux valeurs a_α et b_α encadrant σ^2 qui vérifient :

$$\mathbb{P}(a_\alpha < \sigma^2 < b_\alpha) = 1 - \alpha$$

Cas moyenne μ connue

- Soit une population normale de variance σ^2 inconnue, dans ce cas l'estimateur ponctuel proposé pour σ^2 est :

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

- La variable aléatoire $Y = \frac{n}{\sigma^2} \times \widehat{\sigma}_n^2 = \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}$ suit une loi du $\chi^2(n)$ à $n-$degrés de liberté, qui n'est pas une loi symétrique.

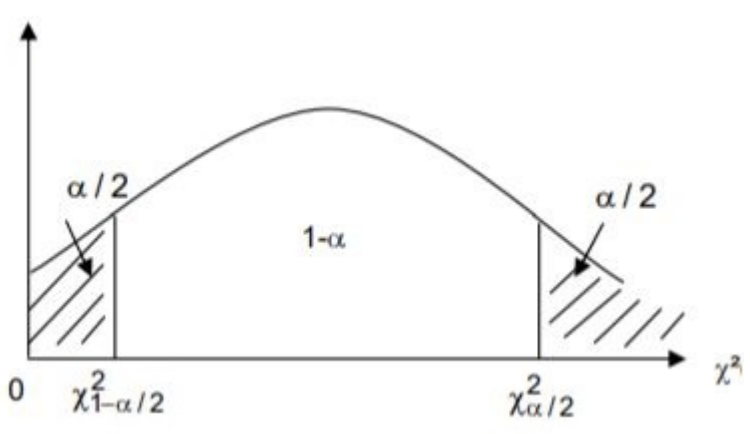
- L'idée principale de la construction de l'intervalle de confiance I_{σ^2} pour σ^2, avec un risque α fixé, est la suivante :
On cherche $\chi_{1-\frac{\alpha}{2},n}$ qui vérifie :

$$\mathbb{P}\Big(\chi^2(n) \leq \chi_{1-\frac{\alpha}{2},n}\Big) = \frac{\alpha}{2}$$

et $\chi_{\frac{\alpha}{2},n}$ qui vérifie :

$$\mathbb{P}\Big(\chi^2(n) \geq \chi_{\frac{\alpha}{2},n}\Big) = \frac{\alpha}{2}$$

Ce qui implique que :



$$\mathbb{P}\left(\chi_{1-\frac{\alpha}{2},n} < \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} < \chi_{\frac{\alpha}{2},n}\right) \quad = \quad 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

alors on montre que :

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{\frac{\alpha}{2},n}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{1-\frac{\alpha}{2},n}}\right) \quad = \quad 1 - \alpha$$

Et on admet le résultat suivant :

Théorème 5

Un intervalle de confiance de seuil α pour le paramètre σ^2 de la loi $\mathcal{N}(\mu, \sigma)$ lorsque μ est connue est de la forme :

$$IC(\sigma^2) = \left[\frac{n\widehat{\sigma}_n^2}{\chi_{\frac{\alpha}{2},n}} , \frac{n\widehat{\sigma}_n^2}{\chi_{1-\frac{\alpha}{2},n}} \right]$$

Cas moyenne μ inconnue

On se propose de donner un intervalle de confiance de niveau de confiance $1-\alpha$ pour σ^2 avec μ inconnue. Dans ce cas l'estimateur ponctuel proposé pour σ^2 est :

$$S_n'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Par conséquent $Y = \frac{(n-1)}{\sigma^2} \times S_n'^2$ suit loi du $\chi^2(n-1)$ à $(n-1)-$degrés de liberté. En suivant le même démarche que celui de la situatuion précédente on aboutit au résultat suivant :

Théorème 6

Un intervalle de confiance de seuil α pour le paramètre σ^2 de la loi $\mathcal{N}(\mu, \sigma)$ lorsque μ est inconnue est :

$$IC(\sigma^2) = \left[\frac{(n-1)S_n'^2}{\chi_{\frac{\alpha}{2},n-1}} , \frac{(n-1)S_n'^2}{\chi_{1-\frac{\alpha}{2},n-1}} \right]$$

Exercice 4 *On a mesuré la quantité totale d'alcool (exprimée en g/l) contenue dans un échantillon de 10 bouteilles de cidre doux du marché. On a obtenu des valeurs $x_1, x_2, x_3, \ldots, x_{10}$ tels que :*

$$\sum_{i=1}^{10} x_i = 62 \quad et \quad \sum_{i=1}^{10} x_i^2 = 388.4124$$

On modélise la quantité d'alcool contenue dans une bouteille par une variable aléatoire X suivant une loi normale d'espérance μ et de variance σ^2, où les paramètres μ et σ étant inconnus.

1. *À partir de l'échantillon observé, proposer des estimations ponctuelles de μ et σ^2.*
2. *Construire un intervalle de confiance pour la moyenne μ au niveau de confiance de $1-\alpha = 95\%$.*
3. *Déterminer un intervalle de confiance à 80% de la variance σ^2.*
4. (a) *Si n désigne la taille d'un grand échantillon ($n > 50$), exprimer en fonction de n l'amplitude de l'intervalle de confiance de μ au niveau de confiance de 95%.*
   (b) *On souhaite construire un intervalle de confiance de μ au niveau de confiance 95% ayant une amplitude de $0,2$. Quelle est la taille de l'échantillon pour une variance échantionnelle $s_n'^2 = 0,6$ g/l ?*