MAT 2777
Probabilités et statistique pour ingénieur.e.s

Chapitre 5
L'estimation et les intervalles de confiance

P. Boily (uOttawa)

Hiver 2023

Aperçu

5.1 – L'inférence statistique (p.2)
- Les statistiques (p.4)
- La variance d'estimation et l'erreur-type (p.5)

5.2 – L'intervalle de confiance de μ lorsque σ est connu (p.8)
- La règle du $68 - 96 - 99.7$ et les intervalles de confiance (p.9)
- L'intervalle de confiance de μ lorsque σ est connu (reprise) (p.14)

5.3 – Le choix de la taille de l'échantillon (p.26)

5.4 – L'intervalle de confiance de μ lorsque σ est inconnu (p.30)

5.5 – L'intervalle de confiance d'une proportion (p.35)

Annexe – Résumé (p.39)

5.1 – L'inférence statistique

L'un des objectifs de l'**inférence statistique** est de pouvoir tirer des conclusions sur une **population** à partir d'un échantillon aléatoire de cette population.

Exemples :

- Peut-on évaluer la fiabilité du processus de fabrication d'un produit en sélectionnant au hasard un échantillon du produit final et en déterminant combien d'entre eux sont conformes à un certain schéma d'évaluation de la qualité ?

- Peut-on déterminer qui va remporter une élection en interrogeant un petit échantillon de répondants ?

Plus précisément, nous cherchons à estimer un **paramètre** θ inconnu, disons, à l'aide d'une seule quantité, l'**estimé ponctuel** $\hat{\theta}$.

Cette estimation ponctuelle est obtenue à l'aide d'une **statistique**, une fonction d'un échantillon aléatoire. La distribution de probabilité de la statistique est sa **distribution d'échantillonnage**. Leur description est une des principales voies de recherche.

Exemple : Considérons un processus qui fabrique des roues dentées (dans un certain calibre). Soit X la v.a. qui enregistre le poids d'une roue dentée choisie au hasard. Quelle est la moyenne de la population $\mu_X = \mathrm{E}[X]$?

Solution : en l'absence d'une f.d.p. $f(x)$, nous pouvons estimer μ_X à l'aide d'un échantillon aléatoire $X_1, \ldots, X_n$ de mesures du poids, via la statistique de la moyenne de l'échantillon :

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \approx \mathcal{N}\left(\mu, \sigma^2/n\right) \text{ selon le TLC.}$$

Les statistiques

Voici quelques exemples de statistiques :

- la moyenne et la médiane de l'échantillon
- la variance et l'écart-type d'échantillon
- les quantiles de l'échantillon (médiane, quartiles, percentiles)
- les statistiques de test (t, χ^2, f, etc.)
- les statistiques d'ordre (le max./min. et l'étendue de l'échantillon, etc.)
- les moments de l'échantillon et leurs fonctions (l'asymétrie, l'aplatissement, etc.)

La variance d'estimation et l'erreur-type

L'**erreur-type** d'une statistique est l'**écart-type de sa distribution d'échantillonnage**.

Par exemple, si les observations $X_1, \ldots, X_n$ proviennent d'une population de **moyenne inconnue** μ et de **variance connue** σ^2, alors $\mathrm{Var}(\overline{X}) = \sigma^2/n$ et l'**erreur-type** de $\overline{X}$ est

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Si la variance de la population d'origine est **inconnue**, alors elle s'approche de la variance de l'échantillon S^2 et l'erreur-type approximative de $\overline{X}$ est

$$\hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{n}}, \quad \text{où} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Exemples :

1. Voici un échantillon de 20 tailles de joueurs de baseball (en pouces) :

   74,74,72,72,73,69,69,71,76,71,73,73,74,74,69,70,72,73,75,78.

   Soit $\overline{X}$ la moyenne d'échantillon. Alors ,

$$\overline{X} = \frac{X_1 + \cdots + X_{20}}{20} = 72.6$$

   et la variance d'échantillon S^2 est

$$S^2 = \frac{1}{20-1}\sum_{i=1}^{20}(X_i - 72.6)^2 \approx 5.6211.$$

L'erreur-type de $\overline{X}$ est donc

$$\hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{20}} \approx \sqrt{\frac{5.6211}{20}} \approx 0.5301.$$

2. Considérons un échantillon $\{X_1, \ldots, X_{100}\}$ d'observations indépendantes prélevées d'une population normale $\mathcal{N}(\mu, \sigma^2)$, où $\sigma = 50$ est connu mais μ ne l'est pas. Quelle est la meilleure estimation de μ ? Quelle est la distribution d'échantillonnage de cette estimation ?

   Solution : la moyenne d'échantillon $\overline{X} = \frac{X_1 + \cdots + X_{100}}{100}$ fournit la meilleure estimation de $\mu_X = \mu_{\overline{X}}$.

   L'erreur-type de $\overline{X}$ est $\sigma_{\overline{X}} = \frac{50}{\sqrt{100}} = 5$. Étant donné que les observations sont prélevées indépendamment d'une population normale avec une moyenne de μ et un écart-type de 50, $\overline{X} \sim \mathcal{N}(\mu, 5^2) = \mathcal{N}(\mu, 25)$, selon le TLC.

5.2 – L'intervalle de confiance de μ lorsque σ est connu

Soit un échantillon $\{x_1, \ldots, x_n\}$ prélevé d'une population normale dont la variance σ^2 est **connue** et dont la moyenne μ est **inconnue**. La moyenne de l'échantillon

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

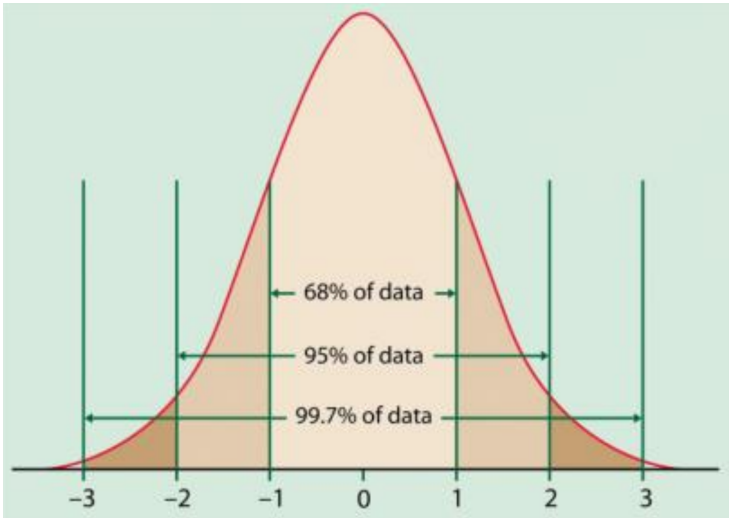
est une **estimation ponctuelle** de μ.

Bien sûr, cette estimation n'est sans doute pas exacte, car $\overline{x}$ est une valeur **observée** de $\overline{X}$; il est peu probable que la valeur observée $\overline{x}$ coïncide avec μ.

Mais nous savons que $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, et donc que

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

La règle du $68 - 96 - 99.7$ et les intervalles de confiance



$$P(-1 < Z < 1) \approx 0.683$$
$$P(-2 < Z < 2) \approx 0.955$$
$$P(-3 < Z < 3) \approx 0.997.$$

Chaque fois que nous observons une moyenne d'échantillon $\overline{X}$ en provenance d'une population normale de moyenne μ, nous nous attendons à ce que l'inégalité

$$-k < Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < k$$

se vérifie approximativement

$$g(k) = \begin{cases} 68.3\% \text{ du temps} & \text{si } k = 1 \\ 95.5\% \text{ du temps} & \text{si } k = 2 \\ 99.7\% \text{ du temps} & \text{si } k = 3 \end{cases}$$

De manière équivalente, l'**intervalle de confiance (symétrique) de μ à $g(k)$** est

$$\overline{X} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + k\frac{\sigma}{\sqrt{n}} \implies \mathsf{IC}(\mu; g(k)) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}}.$$

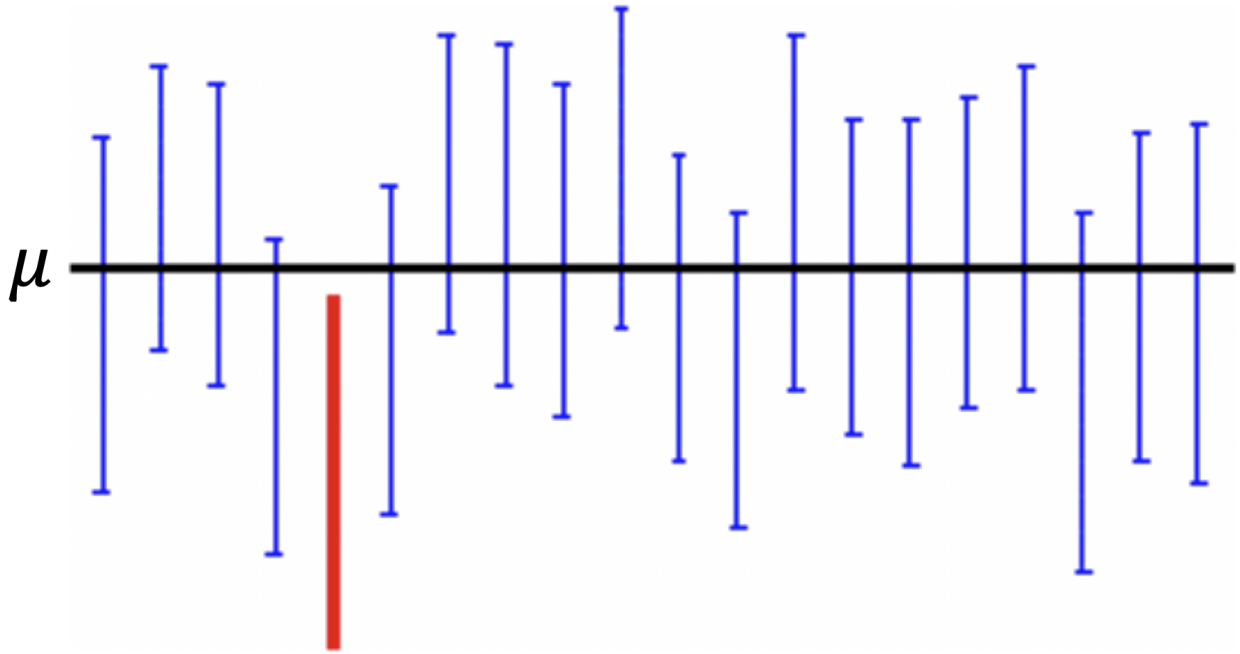
Exemples :

1. Soit $\{X_1, \ldots, X_{64}\}$ un échantillon aléatoire prélevé d'une population normale avec un écart-type de $\sigma = 72$ et une moyenne μ inconnue. La moyenne de l'échantillon est $\overline{X} = 375.2$. Construisez un intervalle de confiance de μ à 68.3%.

   Solution : c'est le cas $k = 1$. Selon la formule, $\mathrm{IC}(\mu; 68.3\%)$ est

   $$375.2 \pm 1 \cdot \frac{72}{\sqrt{64}} \quad \Longrightarrow \quad \mathrm{IC}(\mu; 68.3\%) \equiv (366.2, 384.2).$$

   TRÈS IMPORTANT: ceci ne dit pas que nous sommes certains à 68.3% que μ se situe entre 366.2 et 384.2 – lorsqu'un échantillon de taille 64 est prélevé dans une population normale $\mathcal{N}(\mu, 72^2)$ et qu'on construit l'intervalle $\mathrm{IC}(\mu; 68.3\%)$, μ se retrouve entre les extrémités de l'intervalle environ 68.3% du temps.



Dans un IC à 95%, nous nous attendons à ce que 19 échantillons sur 20, prélevés d'une population unique, produisent des intervalles de confiance qui contiennent le paramètre de population d'intérêt, en moyenne.

2. Construisez $\mathsf{IC}(\mu; 95.5\%)$ avec les données du problème précédent.

   Solution : c'est le cas $k = 2$, alors

$$375.2 \pm 2 \cdot \frac{72}{\sqrt{64}} \quad \Longrightarrow \quad \mathsf{IC}(\mu; 95.5\%) \equiv (357.2, 393.2).$$

3. Construisez $\mathsf{IC}(\mu; 99.7\%)$ avec les données du problème précédent.

   Solution : c'est le cas $k = 3$, alors

$$375.2 \pm 3 \cdot \frac{72}{\sqrt{64}} \quad \Longrightarrow \quad \mathsf{IC}(\mu; 99.7\%) \equiv (348.2, 402.2).$$

L'IC de μ lorsque σ est connu (reprise)

Une autre approche de construction de l'IC consiste à spécifier la proportion d'intérêt de l'aire sous la f.d.p. $\phi(z)$, puis à déterminer les valeurs critiques correspondantes (les extrémités de l'intervalle).

Soit $\{X_1, \ldots, X_n\}$ un échantillon prélevé de $\mathcal{N}(\mu, \sigma^2)$; on rappelle que $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

Pour un IC à 95%, disons, on doit trouver la **valeur critique** $z^* > 0$ telle que $P(-z^* < Z < z^*) = 0.95$.

Mais le côté gauche peut être ré-écrit comme suit :

$$\begin{aligned} P(-z^* < Z < z^*) &= \Phi(z^*) - \Phi(-z^*) \\ &= \Phi(z^*) - (1 - \Phi(z^*)) = 2\Phi(z^*) - 1. \end{aligned}$$

Nous cherchons donc un $z^* > 0$ tel que

$$0.95 = 2\Phi(z^*) - 1 \Longrightarrow \Phi(z^*) = \frac{0.95 + 1}{2} = 0.975.$$

D'après la table de la f.d.p. $\phi(z)$, nous voyons que $\Phi(1.96) \approx 0.9750$, d'où

$$P(-1.96 < Z < 1.96) = P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \approx 0.95.$$

Autrement dit, l'inégalité

$$-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96$$

est valide avec une probabilité de 0.95 (sous l'interprétation préalable).

De façon équivalente,

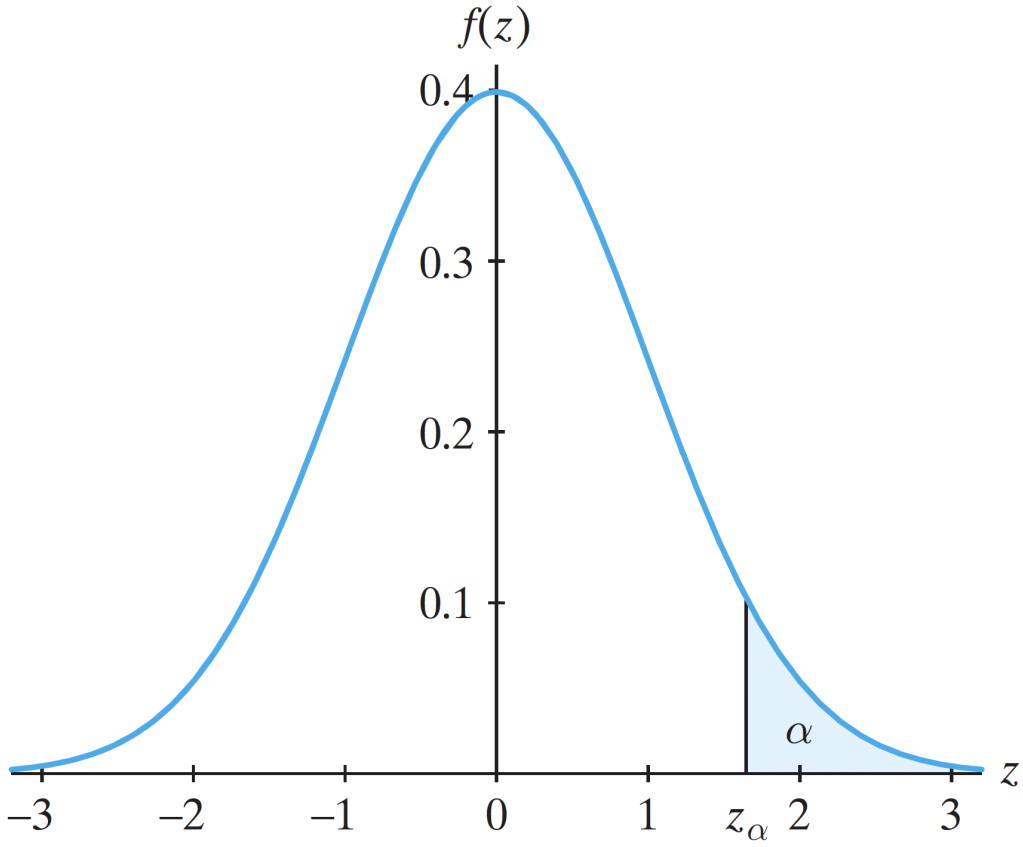
$$\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \quad \Longrightarrow \quad \mathsf{IC}(\mu; 95\%) \equiv \overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

forme un **intervalle de confiance de μ à 95% lorsque σ est connu**.

On peut aussi montrer que

$$\overline{X} - 2.575\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 2.575\frac{\sigma}{\sqrt{n}} \quad \Longrightarrow \quad \mathsf{IC}(\mu; 99\%) \equiv \overline{X} \pm 2.575\frac{\sigma}{\sqrt{n}}$$

forme un **intervalle de confiance de μ à 99% lorsque σ est connu**.



$$P(Z > z_\alpha) = \alpha$$

$$P(Z > z) = 1 - \Phi(z) = \Phi(-z)$$

Le **niveau de confiance** $1-\alpha$ est généralement exprimé en termes de **petits** α, par exemple, $\alpha = 0.05 \Longrightarrow 1-\alpha = 0.95$.

Pour $\alpha = 0.01, 0.02, \ldots, 0.98, 0.99$, les z_α correspondants sont appelés les **pourcentiles** (ou tout simplement les **centiles**) de la loi normale centrée résuite. En général,

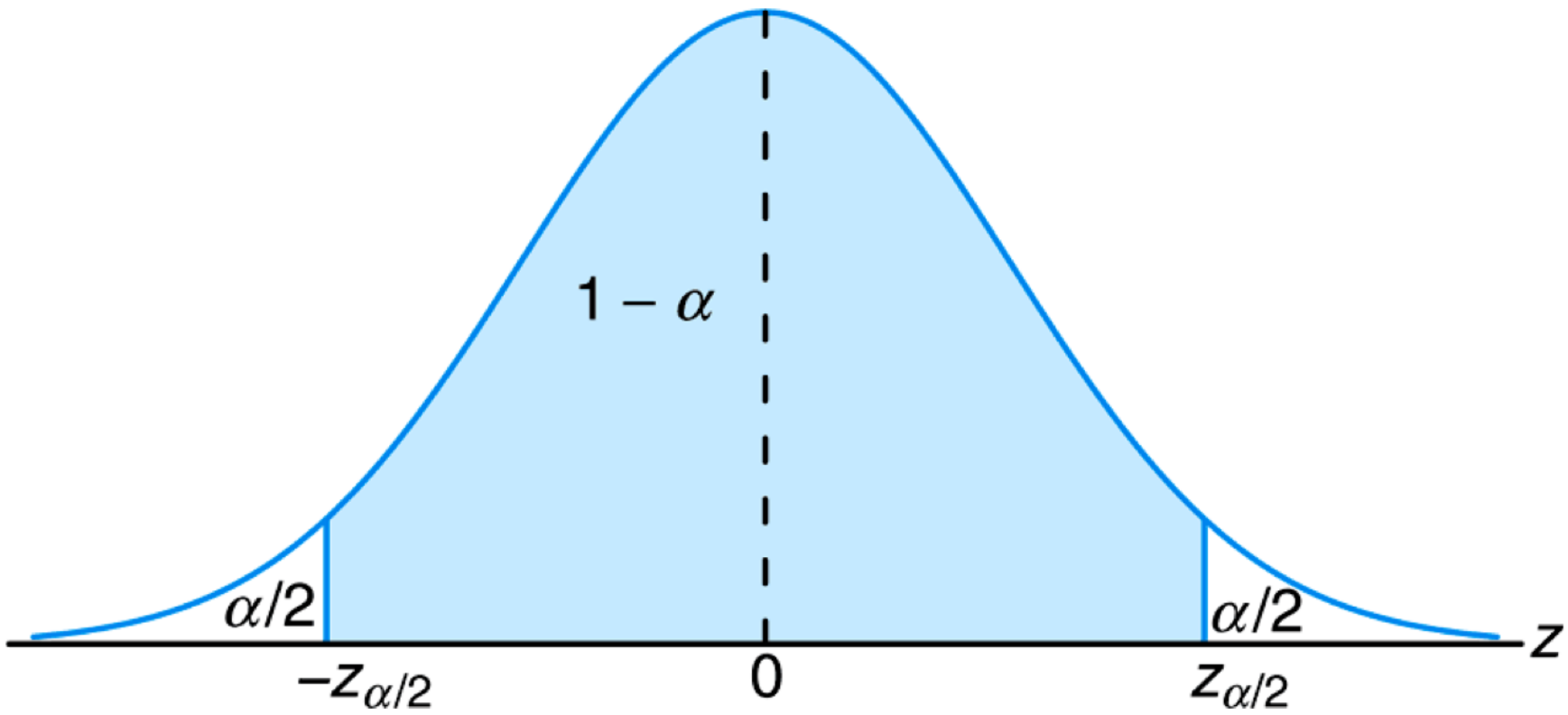
$$P(Z > z_\alpha) = \alpha \implies z_\alpha \text{ est le } 100(1-\alpha) \text{ centile.}$$

Pour les intervalles de confiance **symmétriques** (à 2 côtés), on trouve les valeurs appropriées en résolvant $P(|Z| > z^*) = \alpha$ pour z^*. Par les propriétés de $\mathcal{N}(0,1)$,

$$\alpha = P(|Z| > z^*) = 1 - P(-z^* < Z < z^*) = 1 - (2\Phi(z^*) - 1) = 2(1 - \Phi(z^*)),$$

de sorte que

$$\Phi(z^*) = 1 - \alpha/2 \implies z^* = z_{\alpha/2}.$$


1 − α
α/2
α/2
z
−z_{α/2}
0
z_{α/2}

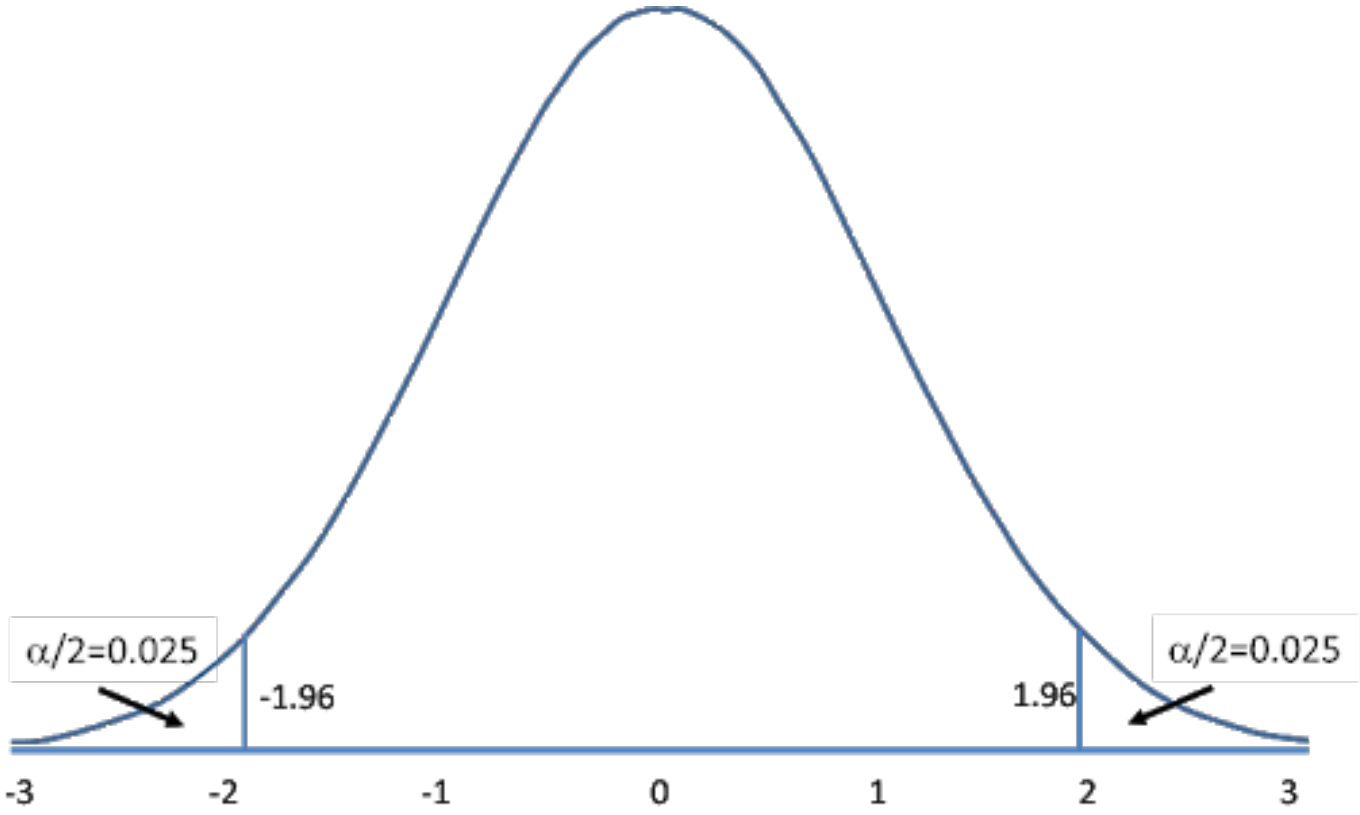
Par exemple,

$$P(|Z| > z_{0.025}) = 0.05 \implies z_{0.025} = 1.96$$
$$P(|Z| > z_{0.005}) = 0.01 \implies z_{0.005} = 2.575.$$

Dans ce même contexte ($X_i \sim \mathcal{N}(\mu, \sigma^2)$, σ connu), l'intervalle de confiance de μ pour un α donné prend généralement la forme

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies \text{IC}(\mu; 100(1-\alpha)\%) \equiv \overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Il importe alors de savoir calculer les centiles $z_{\alpha/2}$ (soit à l'aide de tables, soit à l'aide d'un logiciel).


α/2=0.025
-1.96
1.96
α/2=0.025
-3
-2
-1
0
1
2
3

Pour un niveau de confiance α donné, les intervalles de confiance **plus étroits** sont primés par rapport à l'estimation de la moyenne :

- les estimations deviennent "**meilleures**" lorsque la taille n de l'échantillon **augmente** ;
- les estimations deviennent "**meilleures**" lorsque σ diminue.

Si $\alpha_1 > \alpha_2$, l'intervalle de confiance de μ à $100(1-\alpha_1)\%$ est plus étroit que l'intervalle de confiance de μ à $100(1-\alpha_2)\%$.

En particulier,

$$\mathrm{IC}(\mu; 95\%) \subseteq \mathrm{IC}(\mu; 99\%).$$

Si l'échantillon provient d'une population normale, alors l'intervalle de confiance est **exact**. Autrement, nous pouvons utiliser le TLC afin d'obtenir un intervalle de confiance **approximatif**, lorsque n est suffisamment élevé.

Exemples :

1. Un échantillon de $n = 9$ observations provenant d'une population normale ayant un écart-type $\sigma = 5$ connu a une moyenne d'échantillon $\overline{X} = 19.93$. Donnez IC$(\mu; 0.95)$ et IC$(\mu; 0.99)$ sur la base de cet échantillon.

   Solution : l'estimation ponctuelle de μ est $\overline{X} = 19.93$, et

$$\text{IC}(\mu; 100(1-\alpha)\%) \equiv \overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

   Ainsi,

$$95\% : \overline{X} \pm z_{0.025}\frac{\sigma}{\sqrt{n}} \Longrightarrow 19.93 \pm 1.96 \cdot \frac{5}{\sqrt{9}} \Longrightarrow (16.66, 23.20)$$

$$99\% : \overline{X} \pm z_{0.005}\frac{\sigma}{\sqrt{n}} \Longrightarrow 19.93 \pm 2.575 \cdot \frac{5}{\sqrt{9}} \Longrightarrow (15.64, 24.22)$$

2. Un échantillon de 25 observations provenant d'une population normale ayant un écart-type $\sigma = 5$ connu a une moyenne d'échantillon $\overline{X} = 19.93$. Donnez IC$(\mu; 0.95)$ et IC$(\mu; 0.99)$ sur la base de cet échantillon.

   Solution : l'estimation ponctuelle de μ est toujours $\overline{X} = 19.93$; les IC$(\mu; 100(1-\alpha)\%)$ recherchés sont

$$95\% : \overline{X} \pm z_{0.025}\frac{\sigma}{\sqrt{n}} \Longrightarrow 19.93 \pm 1.96 \cdot \frac{5}{\sqrt{25}} \Longrightarrow (17.97, 21.89)$$

$$99\% : \overline{X} \pm z_{0.005}\frac{\sigma}{\sqrt{n}} \Longrightarrow 19.93 \pm 2.575 \cdot \frac{5}{\sqrt{25}} \Longrightarrow (17.35, 22.51)$$

3. Un échantillon de 25 observations provenant d'une population normale ayant un écart-type $\sigma = 10$ connu a une moyenne d'échantillon $\overline{X} = 19.93$. Donnez IC$(\mu; 0.95)$ et IC$(\mu; 0.99)$ sur la base de cet échantillon.

   Solution : l'estimation ponctuelle de μ est toujours $\overline{X} = 19.93$; les IC$(\mu; 100(1-\alpha)\%)$ recherchés sont

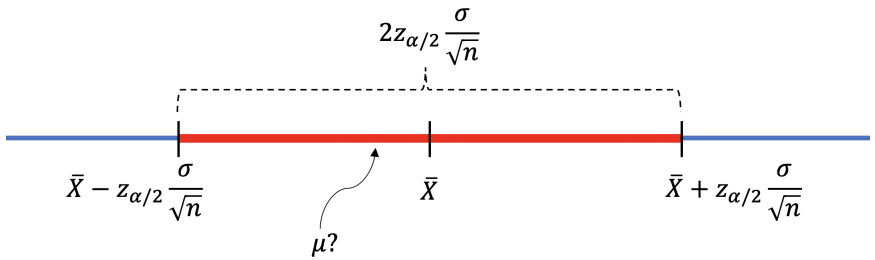
$$95\% : \overline{X} \pm z_{0.025}\frac{\sigma}{\sqrt{n}} \Longrightarrow 19.93 \pm 1.96 \cdot \frac{10}{\sqrt{25}} \Longrightarrow (16.01, 23.85)$$

$$99\% : \overline{X} \pm z_{0.005}\frac{\sigma}{\sqrt{n}} \Longrightarrow 19.93 \pm 2.575 \cdot \frac{10}{\sqrt{25}} \Longrightarrow (14.78, 25.08)$$

   Notez comment les intervalles de confiance sont affectés par α, n, et σ.

5.3 – Le choix de la taille de l'échantillon

Lorsque l'on prélève un échantillon de taille n d'une population normale dont l'écart-type σ est connu, l'erreur que l'on commet en estimant la moyenne μ à l'aide de la moyenne empirique $\overline{X}$ est en général bornée par $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, à un niveau de confiance de $100(1-\alpha)\%$.



Afin de contrôler l'erreur, il faut contrôler la **taille de l'échantillon** :

$$E > \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \implies n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \frac{z_{\alpha/2}^2\sigma^2}{E^2}.$$

Exemples :

1. Un échantillon $\{X_1, \ldots, X_n\}$ est prélevé d'une population normale dont l'écart-type $\sigma = 100$ est connu. Quelle taille d'échantillon n est requise afin de s'assurer que l'erreur sur l'estimation est au plus $E = 10$, à un niveau de confiance $\alpha = 0.05$?

   Solution : tant que

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025} \cdot 100}{10}\right)^2 = (19.6)^2 = 384.16,$$

   alors l'erreur d'estimation commise en utilisant $\overline{X}$ pour approcher μ sera d'au plus $E = 10$, avec une probabilité de 95%.

2. On répète le premier exemple, mais avec $\sigma = 10$.

   Solution : on doit avoir

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025} \cdot 10}{10}\right)^2 = (1.96)^2 = 3.8416.$$

3. On répète le premier exemple, mais avec $E = 1$.

   Solution : on doit avoir

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025} \cdot 100}{1}\right)^2 = (196)^2 = 38416.$$

4. On répète le premier exemple, mais avec $\alpha = 0.01$.

   Solution : on doit avoir

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.005} \cdot 100}{10}\right)^2 = (25.75)^2 = 663.0625.$$

5. On répète le premier exemple, mais avec $\sigma = 10$, $E = 1$, et $\alpha = 0.01$.

   Solution : on doit avoir

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.005} \cdot 10}{1}\right)^2 = (25.75)^2 = 663.0625.$$

La relation entre α, σ, E, et n est simple, mais pas toujours intuitive !

5.4 – L'intervalle de confiance de μ lorsque σ est inconnu

Jusqu'à présent, nous nous sommes trouvés dans la situation heureuse d'échantillonner à partir d'une population dont la variance σ^2 est connue.

Que faire lorsque cette dernière est **inconnue** ?

Nous donnons un estimé de σ par l'entremise de la **variance d'échantillon**

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

(rappelez-vous que la moyenne de la population μ est également inconnue... c'est ce que nous cherchons !) et l'**écart-type empirique** $S = \sqrt{S^2}$.

Si l'on connaît σ, nous savons grâce au TLC que $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ est approximativement $\mathcal{N}(0,1)$.

Si σ est inconnu, on peut montrer que $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ suit approximativement une **loi t de Student** avec $n-1$ **degrés de liberté**, $t(n-1)$.

Par conséquent, pour un niveau de confiance α,

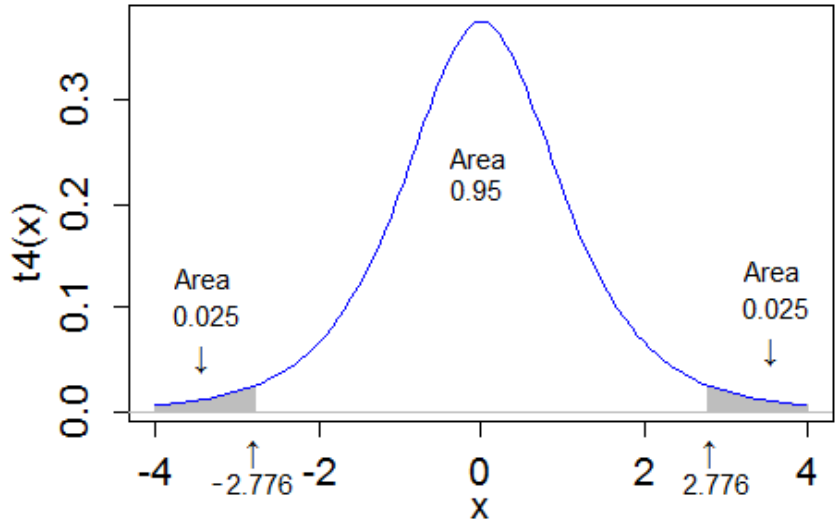
$$P\left(-t_{\alpha/2}(n-1) < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) \approx 1-\alpha,$$

où $t_{\alpha/2}(n-1)$ est le $100(1-\alpha/2)^{\text{e}}$ centile de $t(n-1)$ (on peut les lire dans les tables). L'égalité est atteinte si la population sous-jacente est **normale**:

$$\text{IC}_S(\mu; 100(1-\alpha)\%) \equiv \overline{X} \pm t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}.$$

Par exemple, si $\alpha = 0.05$ et $\{X_1, X_2, X_3, X_4, X_5\}$ est un échantillon prélevé d'une normale dont la variance σ^2 est inconnue, alors

$$t_{0.025}(5-1) = 2.776 \quad \text{et} \quad P\left(-2.776 < \frac{\overline{X} - \mu}{S/\sqrt{5}} < 2.776\right) = 0.95.$$



Exemples :

1. Pour une année donnée, on obtient $n = 9$ mesures de la concentration d'ozone :

   `3.5 5.1 6.6 6.0 4.2 4.4 5.3 5.6 4.4`

   En supposant que les concentrations d'ozone mesurées suivent une loi normale avec $\sigma^2 = 1.21$, construisez IC$(\mu; 95\%)$. Notez que $\overline{X} = 5.01$ et que $S = 0.97$.

   Solution: puisque nous connaissons la variance, nous devons utiliser la valeur critique $z_{\alpha/2} = z_{0.025} = 1.96$:

$$\text{IC}(\mu; 95\%) \equiv \overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 5.01 \pm 1.96 \cdot \frac{\sqrt{1.21}}{\sqrt{9}} \equiv (4.29, 5.73).$$

2. On aborde le même problème, mais en supposant cette fois que la variance de la population sous-jacente est inconnue.

   Solution: Comme nous ne connaissons pas la variance, nous devons utiliser valeur critique $t_{\alpha/2}(n-1) = t_{0.025}(8) = 2.306$ (assurez-vous de comprendre comment obtenir cette valeur à partir du tableau) :

$$\text{IC}(\mu; 95\%) \equiv \overline{X} \pm t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}} = 5.01 \pm 2.306 \cdot \frac{0.97}{\sqrt{9}} \equiv (4.26, 5.76).$$

   Lorsque nous connaissons la variance, l'IC est **plus étroit** (plus petit), ce qui est naturel puisque nous sommes plus confiants lorsque nous avons plus d'informations.

5.5 – L'intervalle de confiance d'une proportion

Si $X \sim \mathcal{B}(n, p)$ (le nombre de réussites dans n épreuves de Bernouilli), alors l'estimateur ponctuel de p est $\hat{P} = \frac{X}{n}$.

Rappelons que $\mathrm{E}[X] = np$ et $\mathrm{Var}[X] = np(1-p)$.

Nous pouvons normaliser toute variable aléatoire, normale ou non :

$$Z = \frac{X - \mu}{\sigma} = \frac{n\hat{P} - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

est approximativement $\mathcal{N}(0, 1)$ si n est suffisamment élevé.

Ainsi,

$$P\left(-z_{\alpha/2} < \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \approx 1-\alpha.$$

En utilisant l'approche précédente, on construit un IC$(p; 100(1-\alpha)\%)$ **approximatif** :

$$\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}},$$

mais ce n'est pas bien utile p est inconnu ! Au lieu, nous utilisons :

$$\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

Exemples :

1. Deux candidats (A et B) se présentent aux élections. Dans un sondage, 1000 électeurs sont choisis au hasard : 52% soutiennent A, tandis que 48% soutiennent B. Donnez un IC$(p; 0.95)$ pour le soutien de chaque candidat.

   Solution : on utilise $\alpha = 0.05$ et $\hat{P} = 0.52$. L'intervalle de confiance recherché du candidat A est ainsi

$$\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = 0.52 \pm 1.96\sqrt{\frac{0.52 \cdot 0.48}{1000}} \approx 0.52 \pm 0.031.$$

   L'intervalle de confiance pour B a la même largeur: 0.48 ± 0.031.

2. Sur la base de ce résultat de sondage, un quotidien publie la une suivante : "La candidate A devance le candidat B !". Ce titre est-il justifié ?

   Solution : bien qu'il y ait un écart de 4 point de pourcentage dans les sondages, le soutien réel pour la candidate A se retrouve dans l'intervalle de confiance

   $$(48.9\%, 55.1\%),$$

   avec une probabilité de 95%. De même, le soutien réel pour le candidat B se situe dans l'intervalle

   $$(44.9\%, 51.1\%),$$

   avec la même probabilité.

   Puisque les intervalles de confiance se **chevauchent**, il est probable que la une **n'est pas justifiée**.

Annexe – Résumé

Échantillon : $\{X_1, \ldots, X_n\}$. **Objectif :** prédire μ avec confiance α.

- Si la population est **normale**, de variance σ^2 **connue**, $\mathsf{IC}(\mu; 100(1-\alpha)\%)$ est **exact** et

$$\mathsf{IC}(\mu; 100(1-\alpha)\%) \equiv \overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

- Si la population est **non-normale**, de variance σ^2 **connue**, et n est suffisament élevé, alors $\mathsf{IC}(\mu; 100(1-\alpha)\%)$ est **approximatif** et

$$\mathsf{IC}(\mu; 100(1-\alpha)\%) \equiv \overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

- Si la population est **normale**, de variance **inconnue**, $\text{IC}_S(\mu; 100(1-\alpha)\%)$ est **exact** et

$$\text{IC}_S(\mu; 100(1-\alpha)\%) \equiv \overline{X} \pm t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}.$$

- Si la population est **non-normale**, de variance **inconnue**, et n est suffisament élevé, alors $\text{IC}_S(\mu; 100(1-\alpha)\%)$ est **approximatif** et

$$\text{IC}_S(\mu; 100(1-\alpha)\%) \equiv \overline{X} \pm z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}.$$

- Si la variance de la population est **inconnue** et si n est '**trop petit**', il n'y a rien à faire...