Estimation par intervalle de confiance

A. Latouche

Cours 6

Rappel : Estimation Ponctuelle

- Quelle est la fréquence, π, des pneumoconioses chez les soudeurs ?
- Quelle est la longévité, μ, des hommes à Paris ?

Estimation à partir d'un échantillon :

- Sur un échantillon de n soudeurs : il y a une proportion p de soudeurs atteints de pneumoconioses
- Sur un échantillon de n résidents parisiens : l'âge moyen du décès est de m années.

On dit que p et m sont des estimations ponctuelles de π et de μ

Variabilité de ces estimations ponctuelles ?

Intervalle de Pari ou intervalle de fluctuation

- Intervalle qui contient la valeur observée sur un échantillon d'un paramètre θ [1] dans une proportion fixée de cas.

- On souhaite calculer un intervalle de pari des réalisation d'une variable aléatoire lorsqu'on **connaît** sa loi de probabilité.

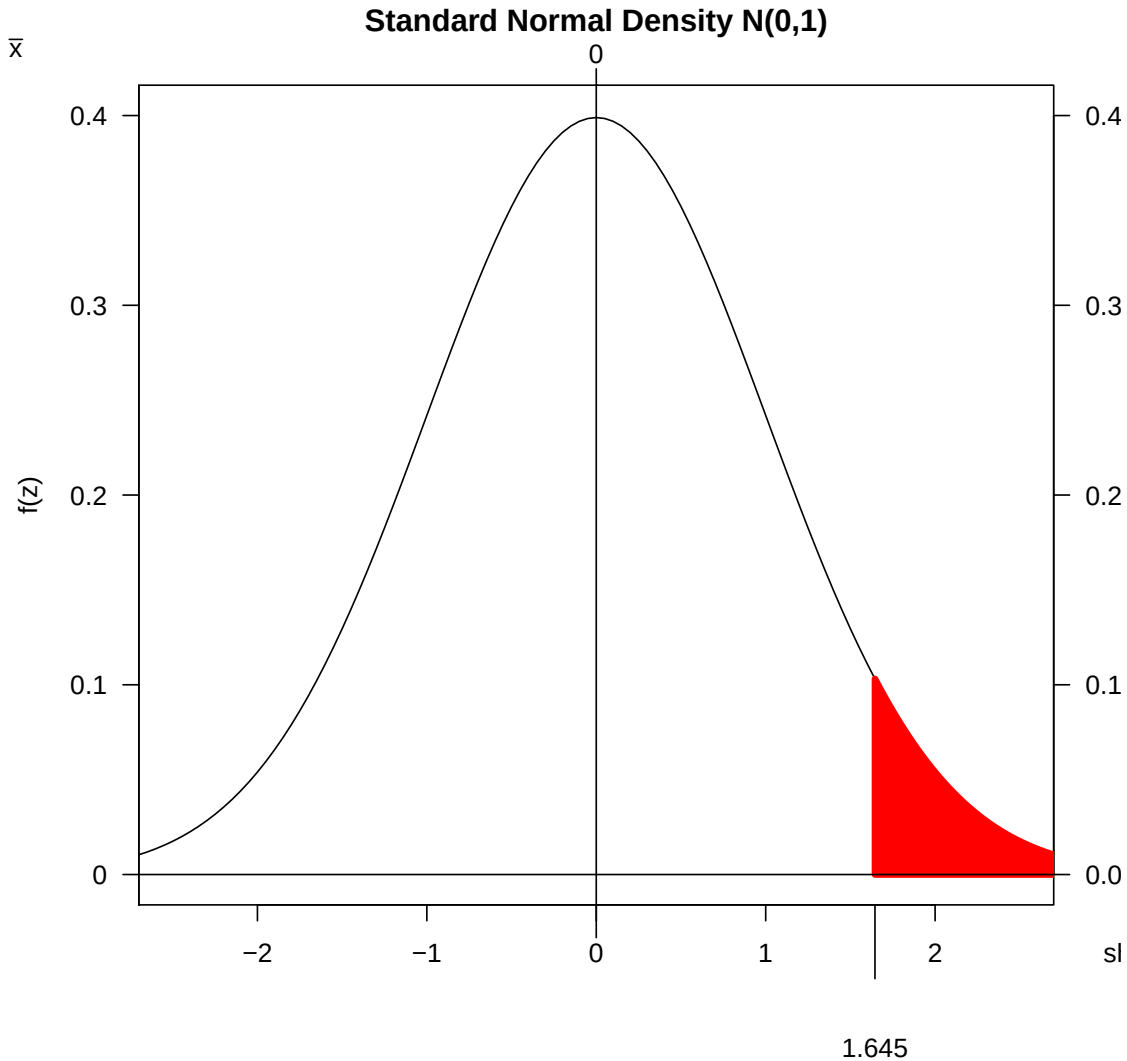
- Soit $X \sim N(m, \sigma^2)$ avec m et σ **connus**

  On cherche a et b tel que $P(a < X < b) = 0.95$

[1] pourcentage, moyenne, ...

Intervalle de fluctuation : Loi Normale $X \sim N(0,1)$

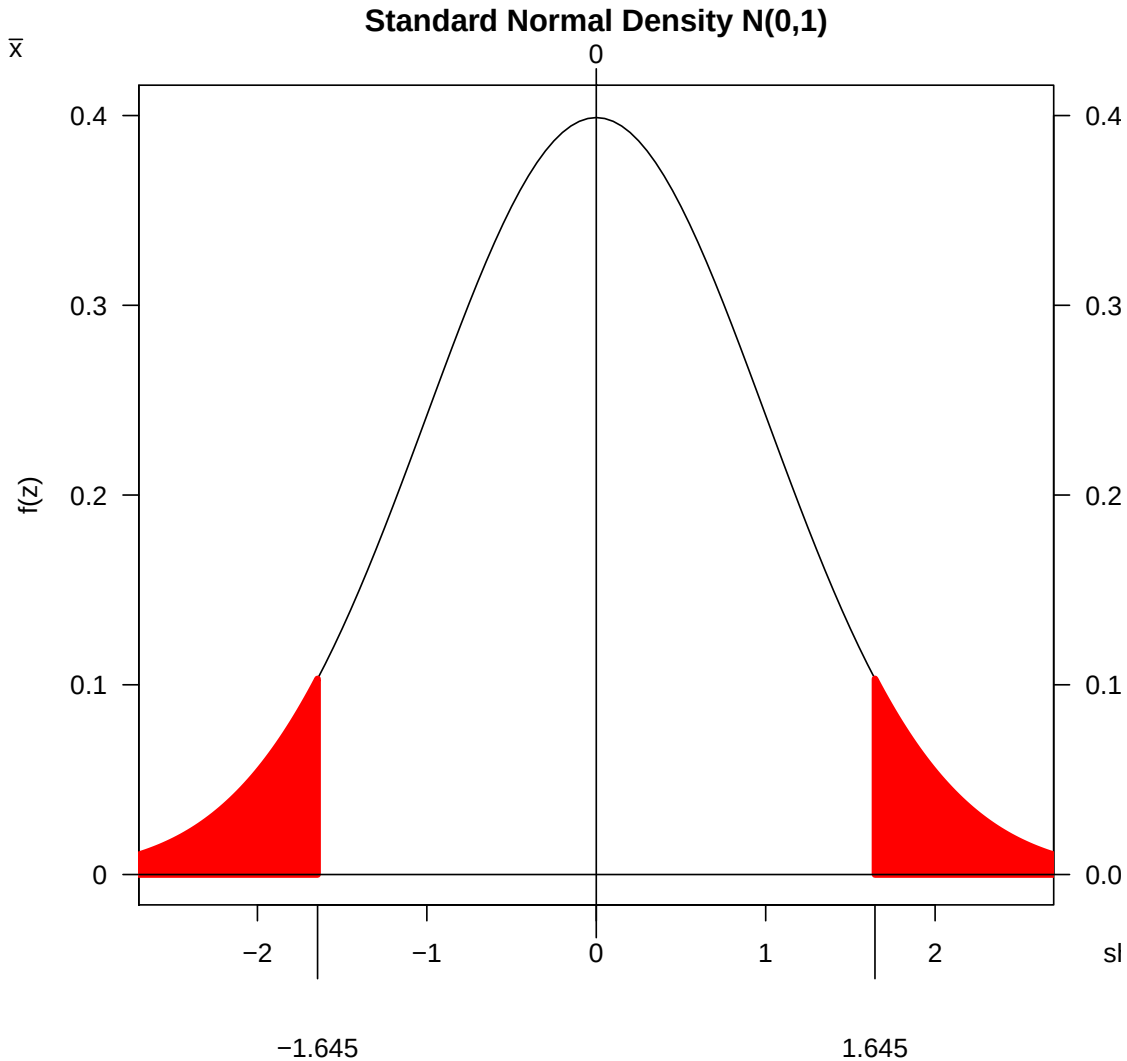
La région en rouge est appelée **Risque** (ou région critique)



$P(X > 1.645) = 0.05$: Unilatéral

Intervalle de fluctuation : Loi Normale $N(0,1)$

La région en rouge est appelée **Risque** (ou région critique)



$P(|X| > 1.645) = 0.1$ et $P(|X| < 1.645) = 0.9$: Bilatéral

Intervalle de fluctuation : Loi Normale $X \sim N(0,1)$

Soit $X \sim N(0,1)$

- $P(X > 1.96) = 0.025$
- $P(X > -1,96) = 0.975$

$\Rightarrow P(-1,96 < X < 1,96) = P(X > -1,96) - P(X > 1,96) = 0,95$

$[-1,96; 1,96]$= intervalle de fluctuation à 95%

Intervalle de fluctuation

De façon plus générale :

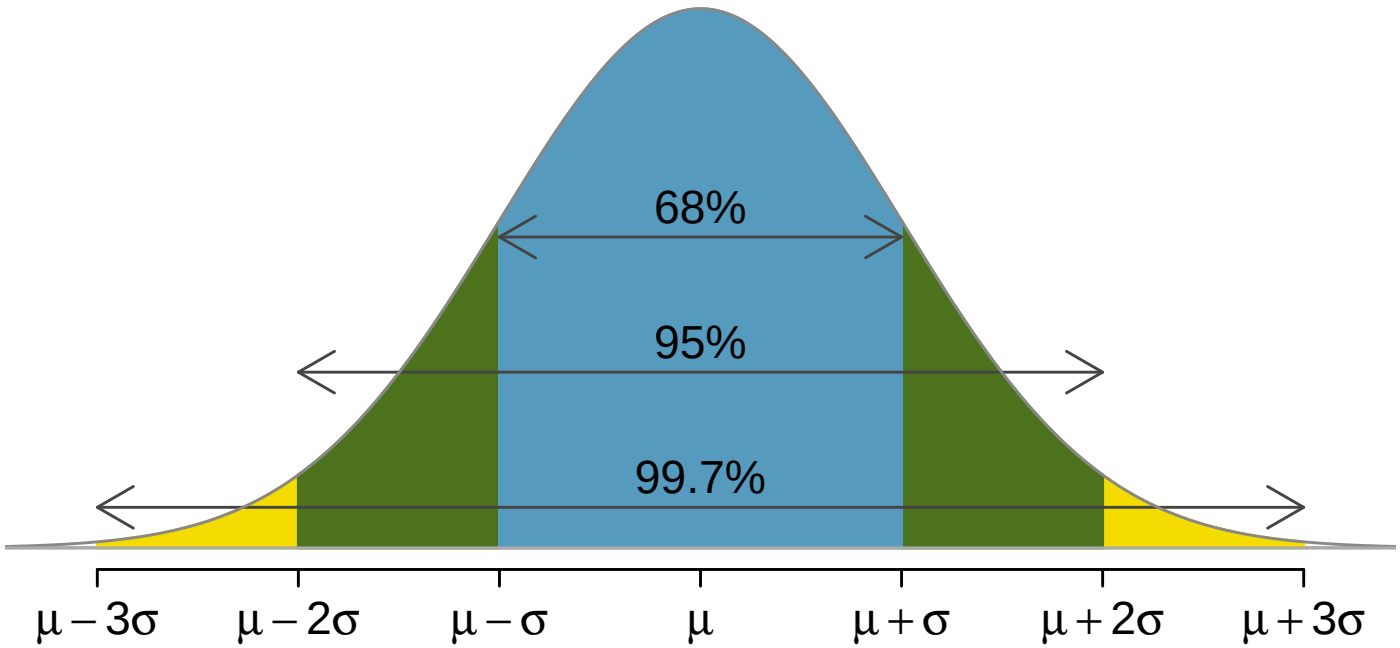
- $P(Z > z_{\alpha/2}) = \alpha/2$
- $P(Z < -z_{\alpha/2}) = \alpha/2$

Donc $\left[-z_{\alpha/2}; +z_{\alpha/2}\right]$ est l'intervalle de fluctuation de Z à $1 - \alpha$ % ou au risque α

Intervalles *classiques*

- $[-1.645; 1.645]$ = intervalle de fluctuation à 90%
- $[-2.576; 2.576]$ = intervalle de fluctuation à 99%

Loi normale (μ, σ^2) : Quelques valeurs



Source `http://www.openintro.org/`

Intervalle de confiance (IC)

En plus de l'estimation ponctuelle d'un paramètre θ, la quantification de la précision est primordiale.

- On considère un intervalle qui a une *grande* probabilité de contenir la vraie valeur du paramètre θ .

Définition

Intervalle de confiance de θ au seuil (1-α)% :

- *θ a une probabilité de (1- α) de se trouver dans cet intervalle*
- *θ a un* **risque** *α de ne pas se trouver dans l'intervalle*

Exemple

On observe une réduction de mortalité de 20% avec un IC à 95% de $[5\%; 35\%]$.

- une baisse de 20% ait été observée ponctuellement dans l'essai,
- il n'est pas possible d'exclure que l'efficacité du traitement soit en réalité
  - plus petite (au pire elle peut être de 5%)
  - plus grande (au mieux de 35%).

En d'autre terme, dans cet essai une réduction de 5% n'est pas statistiquement différente de 20%

Intervalle de Confiance de la moyenne

on observe les réalisations d'une suite (X_n) de variables aléatoires i.i.d $N(\mu, \sigma^2)$ où μ inconnue et σ^2 connue.
Alors

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

et l'intervalle de confiance de μ au risque α est

$$\mu \in [\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Rq : Sans hypothèse de normalité il faut $n \geq 30$

Exemple

Tension artérielle de 41 hommes de plus de 65 ans

- Echantillon 1 : $\bar{X} = 14,97$ et $\sigma^2 = 85,91$

I.C. de m = 14,97 +/- 1,96 * $\sqrt{85,91/41}$= [12,13 ; 17,81]

- Echantillon 2 : $\bar{X} = 15,24$ et $\sigma^2 = 78,12$

I.C. de m = 15,24+/- 1,96$\sqrt{78,12/41}$=[12,53 ; 17,95]

IC moyenne et variance inconnue

Classiquement, on observe les réalisations d'une suite (X_n) de variables aléatoires i.i.d où $X_i \sim N(\mu, \sigma^2)$ **inconnues**. Or

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \to T_{n-1}$$

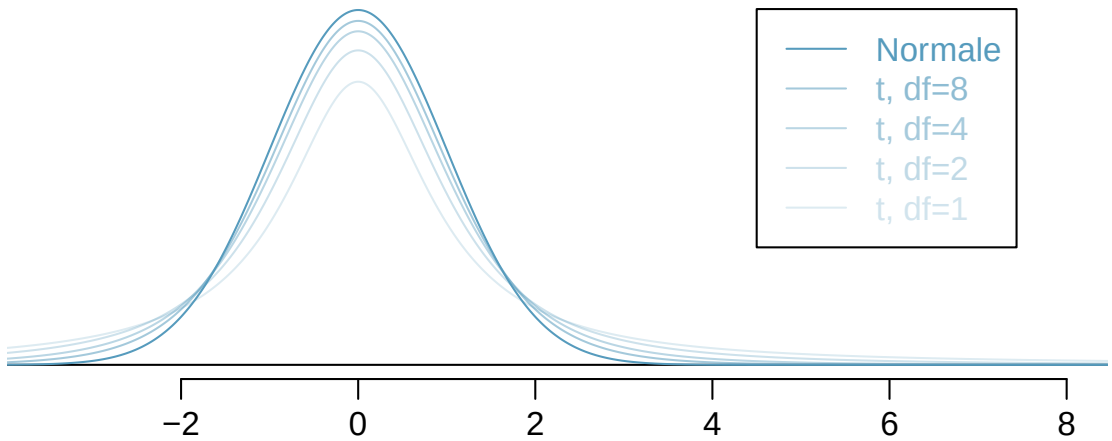
soit $t_{\alpha/2}$ tel que $P(|Z| \leq t_{\alpha/2}) = 1 - \alpha$ et Z suit une loi de Student(n-1)
On en déduit l'intervalle suivant au risque α

$$\mu \in [\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}]$$

Condition d'applications : Données Normales ou n grand ($n \geq 30$)

Loi de Student

Si les degrés de libertés augmentent, la loi de Student converge vers la loi Normale



Loi du χ^2

Définition

Soit $X_1, \ldots, X_n$ *i.i.d* $N(0,1)$, *alors*

$$X_1^2 + \ldots + X_n^2 \sim \chi^2(n)$$

Loi de Student

Définition

Soit

- *Z : variable suivant une loi normale centrée réduite*
- *Y : variable suivant une loi de χ^2 à k ddl indépendante de Z*

alors

$$T = \frac{Z}{\sqrt{Y/k}}$$

suit une loi de Student à k ddl

IC moyenne : Quelques Remarques

- Intervalle est construit autour de la moyenne observée
- La largeur de l'intervalle diminue quand n augmente.
- L'intervalle de confiance est d'autant plus étroit que l'effectif de l'échantillon est grand.
- En augmentant le risque (α), la largeur de l'intervalle de confiance diminue.
- $E(\bar{X}) = \mu$

IC d'une proportion

- On observe une proportion p et on souhaite estimer π
- La loi Binomiale peut être approximer par une loi normale $N(np, np(1-p))$
- Condition d'approximation $n \geq 30$, $np > 5$ et $n(1-p) > 5$

En centrant et réduisant la v.a. on obtient l'IC au risque α pour la proportion π

$$\pi \in \left[p - u_{\alpha/2} \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_{\alpha/2} \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

soit

$$\pi \in [\pi_1; \pi_2]$$

Verifier à posteriori que $n\pi_1 > 5$ et $n\pi_2 > 5$

Exemple IC

Avec un nouveau traitement que vous êtes le seul à utiliser, vous avez guéri 20 patients sur un total de 200.

- Quelle proportion des patients en France seront guéris par ce traitement?[2]

- Condition d'application :

[2] On doit supposer que vos patients sont représentatifs des patients de France et que vous êtes aussi un médecin représentatif des autres médecins.

Exemple IC

Avec un nouveau traitement que vous êtes le seul à utiliser, vous avez guéri 20 patients sur un total de 200.

- Quelle proportion des patients en France seront guéris par ce traitement?[2]

- Condition d'application : On a bien np et $n(1-p) > 5$, donc, en choisissant arbitrairement une confiance de 95%

[2] On doit supposer que vos patients sont représentatifs des patients de France et que vous êtes aussi un médecin représentatif des autres médecins.

Exemple IC

Avec un nouveau traitement que vous êtes le seul à utiliser, vous avez guéri 20 patients sur un total de 200.

- Quelle proportion des patients en France seront guéris par ce traitement?[2]

- Condition d'application : On a bien np et $n(1-p) > 5$, donc, en choisissant arbitrairement une confiance de 95%

p=20/200

- IC(π)= [0.06; 0.14]
- $n\pi_1 = 12 > 5$ et $n\pi_2 = 28 > 5$

[2] On doit supposer que vos patients sont représentatifs des patients de France et que vous êtes aussi un médecin représentatif des autres médecins.

IC d'une proportion : exemple

Lors d'une étude sur une hémopathie maligne, on constitue un échantillon représentatif de 480 malades. Cet échantillon est constitué de 40% de femmes.

- Calculer l'intervalle de confiance à 95% de la proportion d'hommes atteints par cette maladie

- Calculer l'intervalle de confiance à 95 % de la proportion de femmes atteintes par cette maladie

- Quelle devrait être la taille de l'échantillon pour avoir une *précision* de 0.2 ?

Définition : La précision est la demi-longueur de l'IC

IC d'une proportion : exemple

Correction : Proportion observée de femmes = 1 - proportion observée d'homme

- IC(p_H)=p_H +/- 1.96 $\sqrt{p_H(1-p_H)/480}$=[0.56;0.64]

- IC(p_F)=[0.36; 0.44]
- Précision= 1.96 $\sqrt{p_H(1-p_H)/n}$
- Donc $n = \frac{p_H(1-p_H)}{(0.2/1.96)^2}$ soit $n = 23$

Exercice

Vous savez qu'au sein d'une population A, la variable *indice de masse corporelle* (IMC, en kg/m^2) suit une loi normale de moyenne égale à 25 kg/m^2 et de variance égale à 9.
Vous tirez au sort un échantillon de 100 sujets au sein de A et vous vous intéressez à la moyenne observée dans cet échantillon.
Quelles sont les affirmations vraies?

1. Vous allez calculer un intervalle de confiance
2. Vous allez calculer un intervalle de pari
3. L'échantillon est représentatif de la population A
4. On peut affirmer avec une confiance de 95% que la moyenne de l'IMC y sera comprise dans l'intervalle [25 - 1.96*3/10 ; 25 + 1.96*3/10]
5. On peut affirmer avec une confiance de 95% que la moyenne de l'IMC dans cet échantillon sera comprise dans l'intervalle [25 - 1.96*3 ; 25 + 1.96*3]

.

1. Faux : on est dans une situation d'échantillonnage. On connaît la distribution de la variable dans la population et on cherche à prédire le résultat que l'on obtiendrait dans un échantillon. On calcule donc un intervalle de pari.
2. Vrai
3. Vrai : car il y a eu un tirage au sort de la population A
4. Vrai : $m = 25$, $1 - \alpha = 0.95$ donc $u = 1.96$, $s^2 = 9$ d'où intervalle de Pari $[25 - 1.96 * 3/10; 25 + 1.96 * 3/10]$
5. Faux