Estimation et test statistiques

Ricco Rakotomalala

Estimation ponctuelle

Variable aléatoire – Loi de distribution, Stat. paramétrique

X est une v.a. représentant une grandeur quelconque

Ex. Poids d'une baguette, IMC des sportifs, distance de freinage d'un véhicule à une vitesse donnée, etc.

Mettons que X est distribuée selon une loi normale de paramètres (μ,σ)

Fonction de densité

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Statistique paramétrique

On parle de statistique paramétrique quand on fait des hypothèses sur les distributions, et qu'on s'intéresse aux paramètres pour étudier le comportement des variables.

Estimation ponctuelle. Méthode du maximum de vraisemblance (MMV) (1)

On dispose d'un échantillon de taille n obs. (X_1, X_2, ..., X_n). Comment estimer les paramètres de la loi (ex. μ) ?

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

MMV. Produire un estimateur (EMV) qui maximise la probabilité d'obtenir l'échantillon observé : la fonction de vraisemblance.

$$L = P(x_1, x_2, \ldots, x_n; \mu)$$

⇨ $$L = \prod_{i=1}^{n} f(x_i\,; \mu)$$ Parce que X_i son i.i.d.

On manipule plus volontiers la log-vraisemblance.

$$LL = \sum_{i=1}^{n} \log f(x_i\,; \mu)$$

MMV (2)

EMV(μ)

$$\frac{\partial LL}{\partial \mu} = 0 \quad \Rightarrow \quad \hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Il faut aussi que la dérivée seconde soit négative pour un max, c'est le cas ici.

Avec de « belles » propriétés

- Asymptotiquement sans biais
- Convergent
- Asymptotiquement normal

De paramètres

$$E(\bar{X}) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}(n \times \mu) = \mu$$

$$\mathrm{V}(\bar{X}) = \mathrm{V}\left[\frac{1}{\mathrm{n}}\sum_{i=1}^{n} V(X_i)\right] = \frac{1}{\mathrm{n}^2}\sum_{i=1}^{n} V(X_i) = \frac{\sigma^2}{n}$$

Distribution de la moyenne empirique

Les X_i suivent une loi normale, forcément $\bar{X}$ suit une loi normale

Puisque $\bar{X}$ est constituée à partir d'une somme de lois normales.

Mais si les X_i ne suivent pas une loi normale, qu'en est-il ?

$\bar{X}$ tend vers la loi normale en vertu du « Théorème central limite » : « une somme de v.a. i.id. tend (le plus souvent) vers une v.a. gaussienne ».

Le résultat est donc valable pour l'utilisation d'une proportion pour l'estimation d'une probabilité dans une loi de Bernouilli $\mathcal{B}(1,p)$

$$X_i = \begin{cases} 1 \text{ avec la proba « p »} \\ 0 \text{ avec la proba « 1 - p »} \end{cases}$$

$$F = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La fréquence empirique est en réalité une moyenne sur v.a. 0/1.

Estimation par intervalle

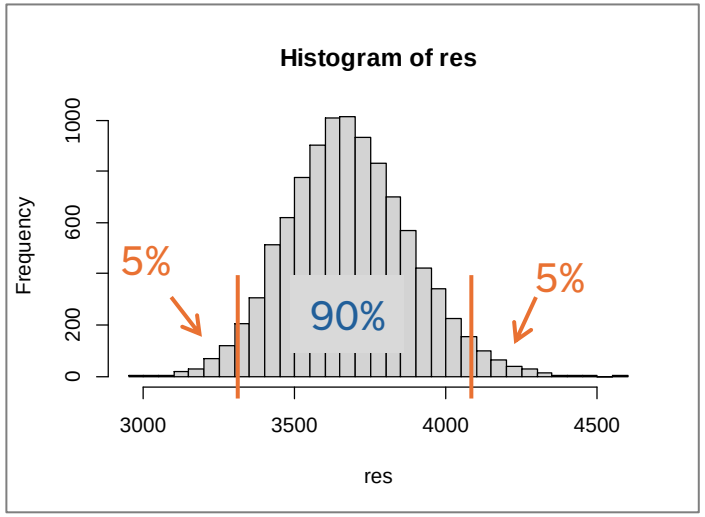
Fluctuations d'échantillonnage – Intervalle de confiance

Une estimation sur un échantillon est forcément entachée d'imprécision.

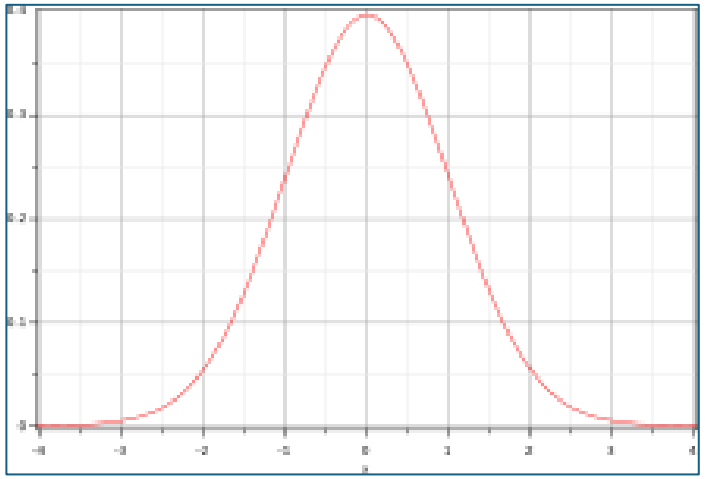
Multiplier les échantillons et calculer empiriquement les bornes des fluctuations. Hélas impraticable.

S'appuyer sur le fait que $\bar{X}$ tend vers la loi normale.

On change d'échantillon, le résultat est différent. Il faudrait en réalité une « fourchette » d'estimation avec un indicateur de fiabilité : intervalle de confiance.



En simplifiant, il y a 90% de chances que l'intervalle contienne la « vraie » valeur de μ.



Intervalle de confiance d'une moyenne – « σ » connu

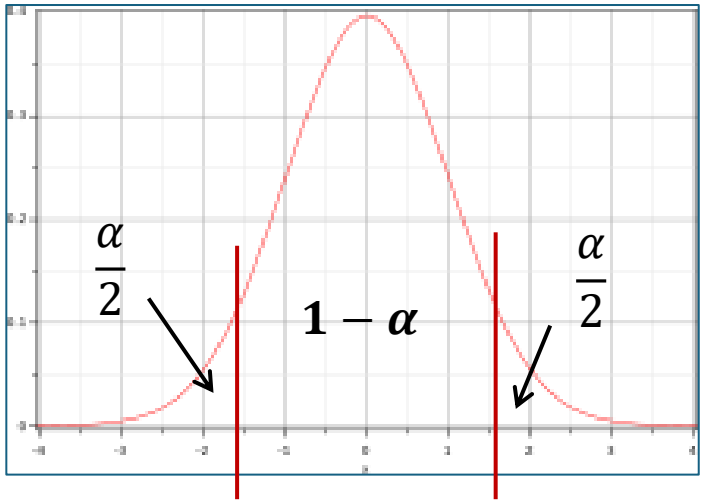
$\bar{X}$ suit asymptotiquement une loi normale de paramètres $(\mu, \sigma_{\bar{X}})$

⇨ $\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}\right) \rightsquigarrow \mathcal{N}(0,1)$ Loi normale centrée et réduite (CR)

Pour un intervalle de confiance au niveau (1 - α), les bornes sont obtenues avec

$$\bar{x} \pm u_{1-\alpha/2} \times \sigma_{\bar{X}}$$

Sachant que $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$



Où « u » sont les quantiles de la loi normale (0, 1)

$u_{\alpha/2}$ $u_{1-\alpha/2}$

⇕

$-u_{1-\alpha/2}$

Puisque la loi normale CR est symétrique autour de 0 !

Intervalle de confiance d'une moyenne – « σ » inconnu (1)

Dans la plupart des cas, « σ » n'est pas disponible, il doit être estimé sur notre échantillon.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Variance « échantillon » qui, à la différence de la variance « population », est non-biaisée.

Quelle distribution ?

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 \rightsquigarrow \sum \mathcal{N}(0,1)^2 \equiv \chi^2(ddl)$$

ddl = (n – nombre de paramètres estimés dans la formule), 1 seul ici avec $\bar{x}$, par conséquent ddl = n-1

Loi de Student

La loi de Student à « k » degrés de liberté est défini par

$$\mathcal{T}(k) = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(k)}{k}}}$$

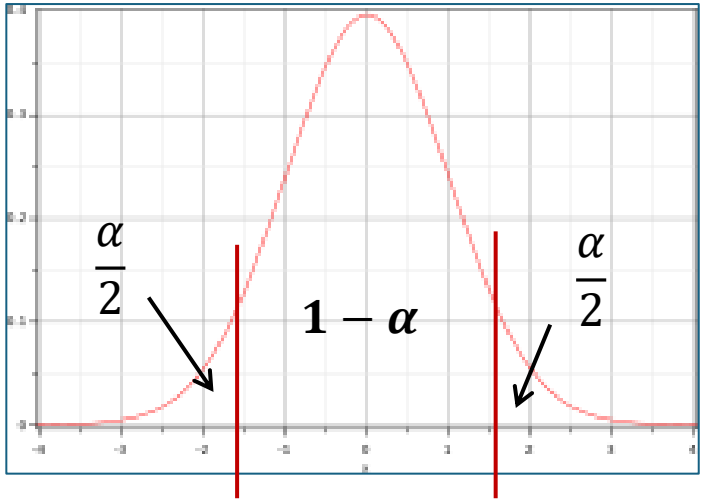
Intervalle de confiance d'une moyenne – « σ » inconnu (2)

Statistique utilisée avec l'estimation de l'écart-type

$$\frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\frac{s^2}{\sigma^2}}{n-1}}} = \frac{\bar{x}-\mu}{s/\sqrt{n}} \rightsquigarrow \mathcal{T}(n-1)$$

Intervalle de confiance au niveau (1 - α)

$$\bar{x} \pm t_{1-\alpha/2} \times \frac{s}{\sqrt{n}}$$



$t_{\alpha/2}$ ⇕ $-t_{1-\alpha/2}$ $\qquad$ $t_{1-\alpha/2}$

Où « t » sont les quantiles de la loi de Student (n-1)

Puisque la loi de Student est aussi symétrique autour de 0 !

Test d’hypothèses

Test de conformité à un standard (à une référence)

On émet des hypothèses les caractéristiques de la population (sur les paramètres des lois en stat. paramétrique)

- H0 : Hypothèse nulle (qui sert de référence)
- H1 : Hypothèse alternative

Mécanisme du test : on cherche à savoir si les données observées s'écartent « significativement » de l'hypothèse nulle

Calculer une statistique de test, vérifier si elle est située dans la « région critique » (région de rejet) c.-à-d. prend une valeur qui amène à rejeter l'hypothèse nulle.

Types de risques

Décision à partir des données

Réalité	Décider H0	Rejeter H0 (décider H1)
H0 est vraie	Bonne décision	α (risque de 1ère espèce)
H1 est vraie	β (risque de 2è espèce)	Bonne décision

Quelques exemples pour la moyenne

$$\begin{cases} H0 : \mu = \mu_0 \\ H1 : \mu \ (< \text{ ou } \neq \text{ ou } >) \ \mu_0 \end{cases}$$

$$\begin{cases} H0 : \mu = 250 \\ H1 : \mu < 250 \end{cases}$$

Le poids moyen d'une baguette est censé être égal à 250g, est-ce que le boulanger du village essaie de gruger ses clients ?

Test unilatéral à gauche

$$\begin{cases} H0 : \mu = 23.15 \\ H1 : \mu > 23.15 \end{cases}$$

L'IMC moyen des joueurs de foot est de 23.15 (Mondial 2018). Est-ce que les joueurs de Montpellier (2024/2025) ont besoin de faire un régime ?

Test unilatéral à droite

$$\begin{cases} H0 : \mu = 32 \\ H1 : \mu \neq 32 \end{cases}$$

Le diamètre usuel d'une pizza normale est de 32 cm. Est-ce que notre pizzaiolo respecte cette norme ?

Test bilatéral

Mécanisme du test d'hypothèses

Principe

Privilégier H0 en maîtrisant le risque de première espèce

α = P(rejeter H0 sur la base des observations / H0 est vraie dans la population)

Mécanisme

On cherche à savoir si on s'écarte significativement de H0 (en direction de H1)

- Fixer un risque d'erreur α (ex. α = 5%, α = 10%, α = 1%...). Parler du rôle de α.
- Calculer la statistique de test. En déterminer la loi de distribution.
- Comparer la statistique calculée avec le quantile théorique de la loi, vérifier :
  - Que nous sommes dans la région d'acceptation c.-à-d. les données ne permettent pas de réfuter H0
  - Ou dans la région de rejet (région critique) c.-à-d. les données contredisent significativement H0

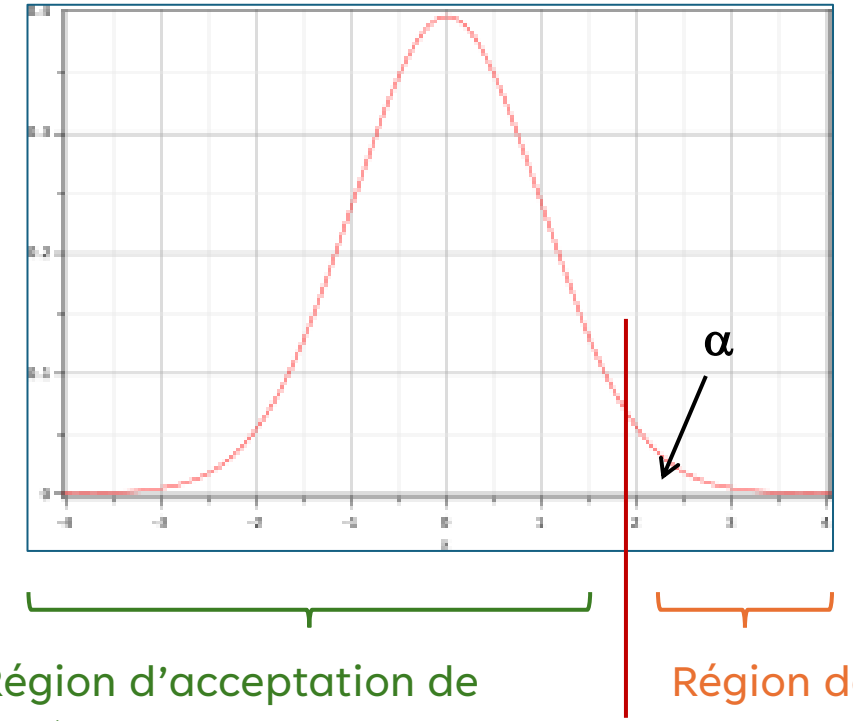
Exemple pour la moyenne

$$\begin{cases} H0 : \mu = \mu_0 \\ H1 : \mu > \mu_0 \end{cases}$$

Statistique de test $t_{calculé} = \left(\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \right) \rightsquigarrow \mathcal{T}(n-1)$

Région critique : est-ce que $\bar{x}$ excède significativement μ_0 au risque α ?

$t_{calculé} > t_{théorique}$?



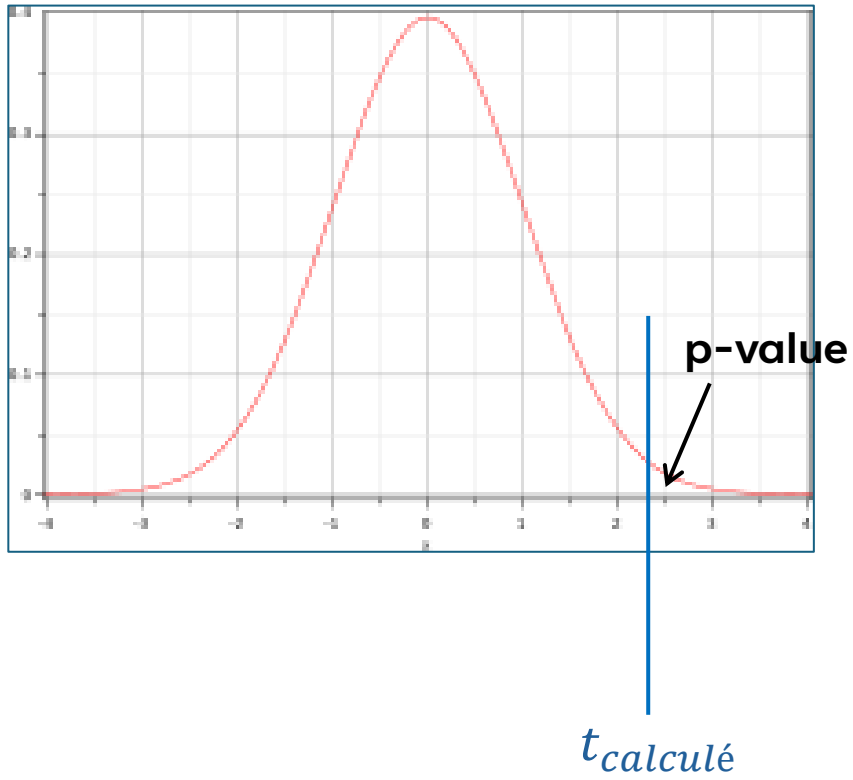
Test unilatéral à droite ici

Exemple pour la moyenne – La p-value

Plutôt que d'opposer la statistique calculée à une valeur seuil, les logiciels de statistique produisent souvent directement la p-value (probabilité critique)

La règle de décision devient :

- p-value < α, rejet de H0
- p-value $\geq \alpha$, pas de rejet de H0

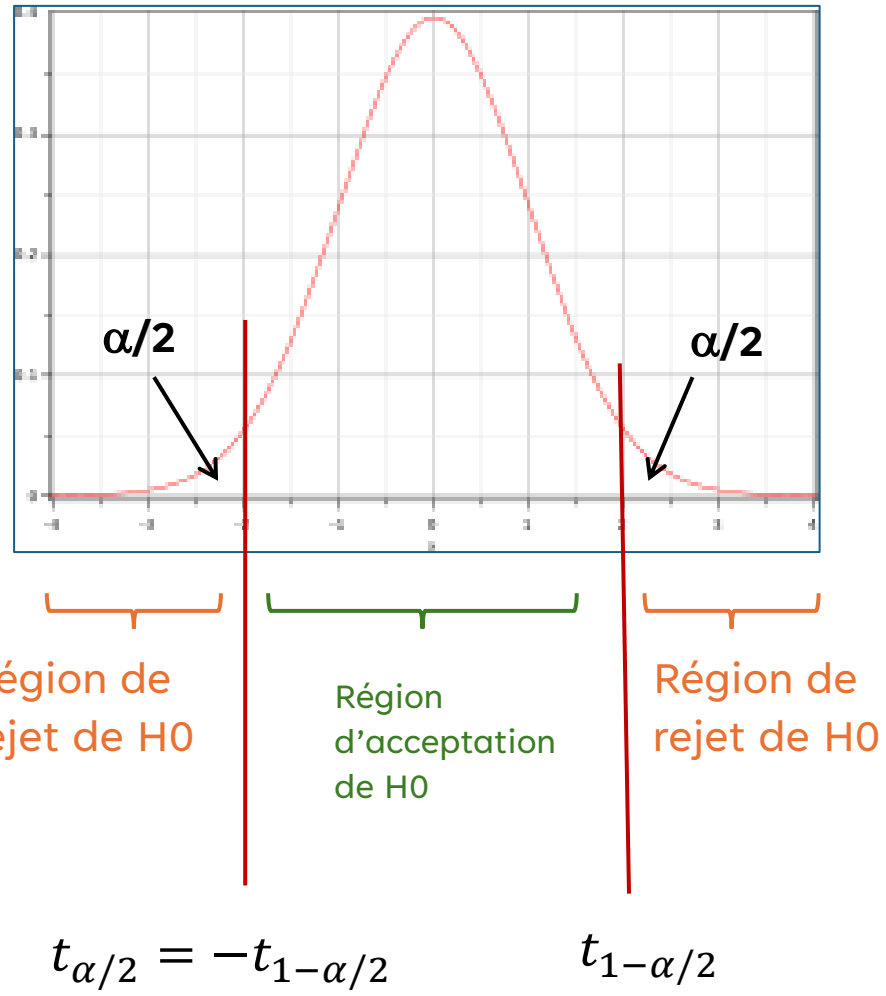


Exemple de la moyenne – Test bilatéral

$$\begin{cases} H0 : \mu = \mu_0 \\ H1 : \mu \neq \mu_0 \end{cases}$$

<u>Région critique :</u>

$|t_{calculé}| > t_{théorique}$?



Bibliographie

Quelques références

- Photis Nobelis – « Statistique » (https://nobelis.eu/photis/index.html)
- PennState – « STAT 415 - Introduction to Mathematical Statistics » (https://online.stat.psu.edu/stat415/)
- Christophe Chesneau – Supports de cours – Voir en particulier « Sur l'estimateur du maximum de vraisemblance » (https://chesneau.users.lmno.cnrs.fr/)
- Groupe des Ecoles des Mines – « Décision et Prévision Statistiques » (http://ressources.unit.eu/cours/Decision_et_prevision_statistiques/)