Cours de remise à niveau de statistique

Formation ModIA

C. Maugis-Rabusseau
GMM Bureau 116
cathy.maugis@insa-toulouse.fr

2023-2024

Bibliographie

Ces slides sont un résumé des cours suivants :

- Cours de Probabilités et Statistique de MIC2 [I2MIMT31]
- Cours de Statistique de MIC3 [I3MIMT41]

Vous pouvez consulter les polycopiés de ces cours sur la page moodle associée.

Plan

Notions de base

Définitions

- **Expérience aléatoire** $\mathcal{E}$ = expérience pour laquelle le résultat est soumis au hasard
- **Univers** Ω de $\mathcal{E}$ = ensemble de tous les résultats possibles
- **Evénement élémentaire** = tout élément ω de Ω

Exemples

1. Lancer d'un dé à 6 faces : $\Omega = \{1, 2, \ldots, 6\}$.
2. Le nombre de tirages nécessaires dans un jeu de Pile ou Face jusqu'à l'obtention de "Pile" : $\Omega = \{1, 2, 3, \ldots\} = \mathbb{N}^*$
3. Durée de vie d'un composant électronique : $\Omega = [0, +\infty[= \mathbb{R}^+$.

Evénement aléatoire

Définition : événement aléatoire

- **Evénement aléatoire** A = sous-ensemble de Ω ($A \subset \Omega$)
  On dit que l'événement A est **réalisé** si le résultat ω de l'expérience appartient à A.

Exemples

1. Lancer d'un dé à 6 faces : $\Omega = \{1, 2, \ldots, 6\}$.
   A = "on a obtenu un chiffre pair " = $\{2, 4, 6\} \subset \Omega$.
2. Le nombre de tirages nécessaires dans un jeu de Pile ou Face jusqu'à l'obtention de "Pile" : $\Omega = \mathbb{N}^*$
   A = "le nombre de tirages nécessaires est inférieur ou égal à 5"
   $A = \{1, 2, 3, 4, 5\} \subset \Omega$
3. La durée de vie d'un composant électronique : $\Omega = [0, +\infty[$
   A = "la durée de vie du composant est supérieure ou égale à 1000 heures" $A = [1000, +\infty[\subset \mathbb{R}^+$.

Tribu

Définition : Tribu

Soit $\mathcal{A}$ un ensemble de parties de l'univers Ω.
On dit que $\mathcal{A}$ est **une tribu** sur Ω si $\mathcal{A}$ satisfait les 3 propriétés suivantes :

1. Le complémentaire de tout élément de $\mathcal{A}$ est dans $\mathcal{A}$:
$$A \in \mathcal{A} \quad \Longrightarrow \quad A^c \in \mathcal{A}.$$
2. L'ensemble $\mathcal{A}$ est stable par union dénombrable :
pour toute famille $(A_n)_{n\geq 1}$ d'éléments de $\mathcal{A}$,
$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}.$$
3. L'univers est un élément de $\mathcal{A}$: $\Omega \in \mathcal{A}$.

Dans ce cas, on dit que le couple $(\Omega, \mathcal{A})$ est **un espace probabilisable.**

Définition : Probabilité

Soit $(\Omega, \mathcal{A})$ un espace probabilisable.
On appelle **probabilité** sur $(\Omega, \mathcal{A})$ une application $\mathbb{P} : \mathcal{A} \to [0, 1]$ telle que :

(i) $\mathbb{P}(\Omega) = 1$.

(ii) Si $(A_n)_{n \geq 1}$ est une famille d'événements **2 à 2 incompatibles** $(A_i \cap A_j = \emptyset, \forall i \neq j)$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Le triplet $(\Omega, \mathcal{A}, \mathbb{P})$ est appelé **espace de probabilité.**

Propriétés

Propriétés

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- Si $A \subset B$ alors $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- Si $A_1, \ldots, A_N$ sont des événements **2 à 2 incompatibles**,

$$\mathbb{P}\left(\bigcup_{n=1}^{N} A_n\right) = \sum_{n=1}^{N} \mathbb{P}(A_n).$$

Indépendance

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité, et soit A et B deux événements aléatoires. On dit que A et B sont **indépendants** (noté $A \perp\!\!\!\perp B$) si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B).$$

Variable aléatoire réelle

Définition

Soit X une application de Ω dans $\mathbb{R}$.
Si B désigne une partie de $\mathbb{R}$, on note

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} .$$

Il s'agit d'une partie de Ω, appelée **image réciproque** de B par X.

Définition

Soit $(\Omega, \mathcal{A})$ un espace probabilisable. On appelle **variable aléatoire (réelle)**, abrégé **v.a.r**, une application X de Ω dans $\mathbb{R}$, telle que :

$$\forall\ B \text{ intervalle de } \mathbb{R}, \quad X^{-1}(B) \in \mathcal{A}.$$

L'ensemble des valeurs prises par X est noté

$$X(\Omega) = \{X(\omega) \in \mathbb{R} : \omega \in \Omega\}.$$

Loi de probabilité

Définition

Soit X une variable aléatoire définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. On appelle **loi de probabilité** de X, notée $\mathbb{P}_X$, l'application qui à toute partie B de $\mathbb{R}$ associe le nombre

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

On notera cette quantité $\mathbb{P}(X \in B)$ par la suite.

Remarque : Lorsque $B = \{x\}$ (respectivement $B =]-\infty, x]$), on note aussi $\mathbb{P}(X = x)$ (respectivement $\mathbb{P}(X \leq x)$)

Remarque : L'application $\mathbb{P}_X$ définit une probabilité sur $\mathbb{R}$.

Fonction de répartition F_X

Définition

La **fonction de répartition** de la v.a. X est définie par

$$\forall x \in \mathbb{R},\ F_X(x) = \mathbb{P}(X \leq x).$$

Propriétés

- $\forall x \in \mathbb{R},\ 0 \leq F_X(x) \leq 1.$
- La fonction F_X converge vers 0 en $-\infty$ et vers 1 en $+\infty$.
- La fonction F_X est croissante.
- La fonction F_X est continue à droite.
- $\forall (a,b) \in \mathbb{R}^2,\ a < b,\ \mathbb{P}(a < X \leq b) = F_X(b) - F_X(a).$
- La fonction de répartition caractérise la loi.

v.a. discrète / v.a. continue

Définition

- On dit que la v.a.r. X est **discrète** si X prend un nombre fini ou dénombrable de valeurs
- Soit X une v. a.r. prenant un nombre infini non dénombrable de valeurs. Si sa fonction de répartition F_X est une fonction continue, on dit que X est une **v.a.r. continue**.

Loi d'une v.a.r

X v.a. discrète

La loi de probabilité de X est entièrement déterminée par les probabilités

$$\mathbb{P}_X(x) = \mathbb{P}(X = x), \ \forall x \in X(\Omega).$$

X v.a. continue

La fonction f_X est **la densité** de X si pour tous réels $a < b$,

$$\mathbb{P}(a < X \leq b) = \int_a^b f_X(x)dx.$$

- $\displaystyle\sum_{x \in X(\Omega)} \mathbb{P}(X = x) = 1$

- $F'_X(x) = f_X(x)$ en tout point où F_X est dérivable
- $\int_{\mathbb{R}} f(x)dx = 1$
- La densité caractérise la loi

Exemples

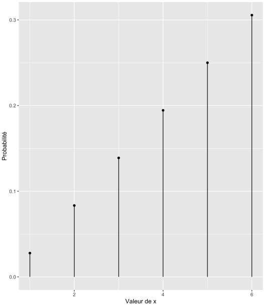
X v.a.r discrète

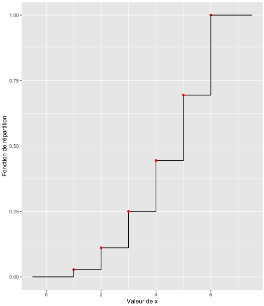
- On lance deux dés.
  X représente la valeur maximale des 2 faces.
- $X(\Omega) = \{1, 2, 3, 4, 5, 6\}$.
- Loi de probabilité de X :

$$\mathbb{P}(X=1) = \frac{1}{36}, \mathbb{P}(X=2) = \frac{3}{36},$$

$$\mathbb{P}(X=3) = \frac{5}{36}, \mathbb{P}(X=4) = \frac{7}{36},$$

$$\mathbb{P}(X=5) = \frac{9}{36}, \mathbb{P}(X=6) = \frac{11}{36}$$



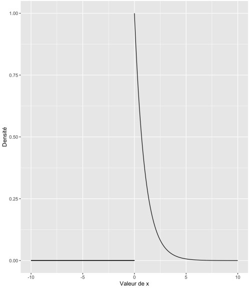


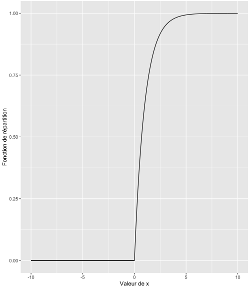
Exemples

X v.a.r continue

- X v.a.r de densité $f_X(x) = e^{-x}\,\mathbb{1}_{x\geq 0}$
- Fonction de répartition

$$
\begin{aligned}
F_X(x) &= \mathbb{P}(X \leq x) \\
&= (1 - e^{-x})\,\mathbb{1}_{x\geq 0}
\end{aligned}
$$





Espérance

Soit X une v.a.r et $h : \mathbb{R} \to \mathbb{R}$

X v.a. discrète

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x \, \mathbb{P}(X = x)$$

$$\mathbb{E}[h(X)] = \sum_{x \in X(\Omega)} h(x) \, \mathbb{P}(X = x)$$

à condition que la série ait un sens

X v.a. continue

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, f_X(x) dx$$

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x) \, f_X(x) dx$$

à condition que l'intégrale converge

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{6} i \, \mathbb{P}(X = i) \\
&= \frac{1}{36} + \frac{2 \times 3}{36} + \ldots + \frac{6 \times 11}{36} \\
\mathbb{E}[X^2] &= \sum_{i=1}^{6} i^2 \, \mathbb{P}(X = i) \\
&= 1^2 \times \frac{1}{36} + 2^2 \times \frac{3}{36} + \ldots + 6^2 \times \frac{11}{36}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[X] &= \int_{\mathbb{R}} x \, f_X(x) dx \\
&= \int_0^{+\infty} x \, e^{-x} dx. \\
\mathbb{E}[X^2] &= \int_{\mathbb{R}} x^2 \, f_X(x) dx \\
&= \int_0^{+\infty} x^2 \, e^{-x} dx.
\end{aligned}$$

Propriétés

Soient X et Y deux v.a.r.

- **Propriété de linéarité :** Pour tout $(\alpha, \beta) \in \mathbb{R}^2$, on a
$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y].$$
- **Indépendance** X et Y sont indépendantes ($X \perp\!\!\!\perp Y$) si et seulement si pour toutes fonctions g et h à valeurs dans $\mathbb{R}$,
$$\mathbb{E}[g(X) \times h(Y)] = \mathbb{E}[g(X)] \times \mathbb{E}[h(Y)].$$
En particulier, si $X \perp\!\!\!\perp Y$ alors $g(X) \perp\!\!\!\perp h(Y)$.

Définition

Soit $p > 0$. On appelle **moment d'ordre** p de la v.a. X la quantité $\mathbb{E}[X^p]$.

Variance / Ecart-type

Définition

La variance de la v.a. X est définie par

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

L'écart-type de la v.a. X est $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

Propriétés

$\forall a, b \in \mathbb{R}$, $\mathrm{Var}(aX + b) = a^2\ \mathrm{Var}(X)$.

Définition

Si X est une v.a. telle que $\mathbb{E}[X] = m$ et $\mathrm{Var}(X) = \sigma^2$, alors la v.a.

$$Y := (X - m)/\sigma$$

est **centrée** ($\mathbb{E}[Y] = 0$) et **réduite** ($\mathrm{Var}(Y) = 1$).

Covariance

Soient X et Y deux v.a.r. admettant une variance.

Définition

La **covariance** du couple (X, Y) est définie par :
$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - E(X))(Y - E(Y))\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Propriété

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(a, X) = 0$
- $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)$
- Var(X+Y)=Var(X)+Var(Y)+2 Cov(X,Y)

Propriété

$X \perp\!\!\!\perp Y \Rightarrow \mathrm{Cov}(X, Y) = 0$ Attention, la réciproque est fausse !

(et alors $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$)

Coefficient de corrélation linéaire

Soient X et Y deux v.a.r. admettant une variance.

Définition

On définit le **coefficient de corrélation linéaire** de X et Y par :

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}\,.$$

Propriété

- Inégalité de Cauchy-Schwarz : $|\rho(X,Y)| \leq 1$ i.e. $\rho(X,Y) \in [-1,1]$
- Cas d'égalité dans cette inégalité :

$$\rho(X,Y) = \pm 1 \iff \exists a,b,c \in \mathbb{R},\ \mathbb{P}(aX+bY=c)=1$$

Fonction caractéristique

Soit X une v.a.r.

X v.a. discrète

$\forall t \in \mathbb{R}$,

$$\begin{aligned}\Phi_X(t) &= \mathbb{E}[e^{itX}] \\ &= \sum_{x \in X(\Omega)} e^{itx}\ \mathbb{P}(X = x)\end{aligned}$$

X v.a. continue

$\forall t \in \mathbb{R}$,

$$\begin{aligned}\Phi_X(t) &= \mathbb{E}[e^{itX}] \\ &= \int_{\mathbb{R}} e^{itx}\ f_X(x)dx\end{aligned}$$

Proposition

Soient X et Y deux v.a.r.

- La fonction caractéristique caractérise la loi.
- Si $X \perp\!\!\!\perp Y$ alors $\forall t \in \mathbb{R},\ \Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t)$.

Quantile

Définition

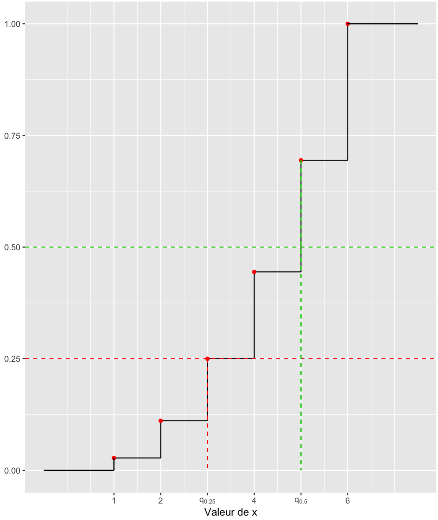
Soit X une v.a.r. de fonction de répartition F_X.
Le **quantile** d'ordre $\alpha \in]0,1[$ de X est le nombre $q_\alpha \in \mathbb{R}$ tel que

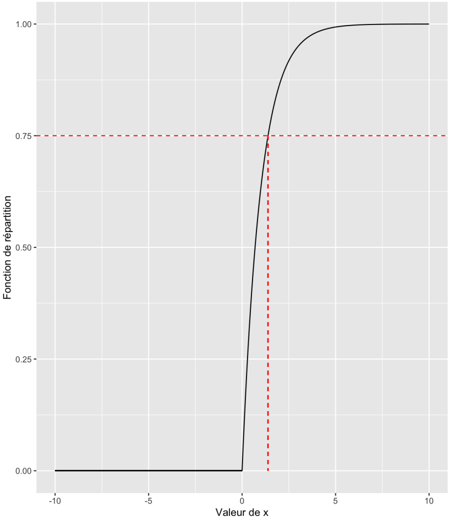
$$q_\alpha = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}.$$

Remarque

Si F_X est strictement croissante alors $F_X(q_\alpha) = \mathbb{P}(X \leq q_\alpha) = \alpha$

Quantile - Exemples





Exemple de lois discrètes

Loi	$X(\Omega)$	Loi de proba $\mathbb{P}(X=k)$	Espérance $\mathbb{E}[X]$	Variance $\mathrm{Var}(X)$
Bernoulli $\mathcal{B}(p)$ $p \in]0,1[$	$\{0,1\}$	$\mathbb{P}(X=1)=p$	p	$p(1-p)$
Binomiale $\mathcal{B}in(n,p)$ $p \in]0,1[, n \in \mathbb{N}$	$\{0,\ldots,n\}$	$C_n^k p^k (1-p)^{n-k}$	np	$np(1-p)$
Géométrique $\mathcal{G}(p)$ $p \in]0,1[$	$\mathbb{N}^*$	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson $\mathcal{P}(\lambda)$ $\lambda > 0$	$\mathbb{N}$	$\frac{\lambda^k}{k!}\, e^{-\lambda}$	λ	λ

Exemples de lois continues

Loi	Densité $f_X(x)$	Espérance $\mathbb{E}[X]$	Variance $\mathrm{Var}(X)$
Loi uniforme $\mathcal{U}[a,b]$ $a<b$	$\frac{1}{b-a}\mathbb{1}_{x\in[a,b]}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Loi exponentielle $\mathcal{E}(\lambda)$ $\lambda>0$	$\lambda e^{-\lambda x}\mathbb{1}_{x\geq 0}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Loi normale $\mathcal{N}(m,\sigma^2)$ $m\in\mathbb{R},\sigma>0$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$	m	σ^2

Exercices

Exercice 1

Soit X et Y deux variables aléatoires indépendantes. X suit une loi géométrique de paramètre $p \in]0,1[$ et Y satisfait la propriété suivante :

$$\mathbb{P}(Y=1) = 1 - \mathbb{P}(Y=-1) = q \text{ avec } q \in]0,1[.$$

1. Déterminez la loi de $Z = Y(X-1)$.
2. Calculez $\mathrm{Cov}(X, Y)$, $\mathrm{Cov}(Z, Y)$ et $\mathrm{Cov}(Z, X)$.

Exercice 2 (Loi de Pareto)

Une variable aléatoire X suit une loi de Pareto de paramètres $a > 0$ et $\alpha > 0$ si elle admet la densité de probabilité suivante

$$f_X(x) = \frac{\alpha a^\alpha}{x^{\alpha+1}} \mathbb{1}_{[a,+\infty[}(x).$$

1. Vérifiez que f_X est bien une densité de probabilité.
2. Déterminez la fonction de répartition F_X de X.
3. Tracez les représentations graphiques de f_X et F_X.
4. Déterminez, si elles existent, $\mathbb{E}[X]$ et $\mathrm{Var}(X)$.

Converge en probabilité

Définition

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. et soit X une autre v.a.
$(X_n)_{n\in\mathbb{N}}$ **converge en probabilité** vers X, noté $X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X$, si

$$\forall \varepsilon > 0, \lim_{n\to+\infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0.$$

Loi des Grands Nombres (LGN)

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. indépendantes et de même loi (i.i.d), admettant une variance. Alors

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\mathbb{P}} \mathbb{E}[X_1]$$

Converge en probabilité

Inégalité de Bienaymé-Tchebychev (B.T.)

Soit X une v.a. admettant une variance.

$$\forall \varepsilon > 0,\ \mathbb{P}\left(|X - \mathbb{E}[X]| > \varepsilon\right) \leq \frac{\text{Var}[X]}{\varepsilon^2}.$$

Exemple

Soit $(X_i)_{i \in \mathbb{N}}$ une suite i.i.d de loi $\mathcal{B}(1/2)$.

- Avec LGN :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n \to +\infty]{\mathbb{P}} \mathbb{E}[X_1] = \frac{1}{2}$$

- Avec l'inégalité de B.T. :
  Comme $n\bar{X}_n \sim \mathcal{B}in(n, 1/2)$, on a $\mathbb{E}[\bar{X}_n] = \frac{1}{2}$ et $\text{Var}(\bar{X}_n) = \frac{1}{4n}$ donc

$$\forall \varepsilon > 0,\ \mathbb{P}\left(\left|\bar{X}_n - \frac{1}{2}\right| > \varepsilon\right) \leq \frac{1}{4\ n\ \varepsilon^2} \underset{n \longrightarrow +\infty}{\longrightarrow} 0 \text{ donc } \bar{X}_n \xrightarrow[n \to +\infty]{\mathbb{P}} \frac{1}{2}$$

Converge en loi

Définition

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a.r et soit X une autre v.a.
$(X_n)_{n\in\mathbb{N}}$ **converge en loi** vers X, noté $X_n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} X$, si

$$\lim_{n\to+\infty} F_{X_n}(t) = F_X(t)$$

pour tout t point de continuité de F_X.

Propriété

$$X_n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} X \quad \Longleftrightarrow \quad \lim_{n\to+\infty} \Phi_{X_n}(t) = \Phi_X(t), \quad \forall t \in \mathbb{R}.$$

Propriété

$$X_n \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} X \quad \Longrightarrow \quad X_n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} X$$

Attention réciproque fausse sauf si la limite est une constante.

Théorème de la Limite Centrale (TLC)

Théorème

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. indépendantes et de même loi, $\mathbb{E}[X_1] < +\infty$ et $\mathrm{Var}(X_1) < +\infty$. Alors

$$\sqrt{n}\,\frac{\bar{X}_n - \mathbb{E}[X_1]}{\sqrt{\mathrm{Var}(X_1)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Exemple

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. i.i.d de loi $\mathcal{B}(p)$, $\mathbb{E}[X_1] = p < +\infty$ et $\mathrm{Var}(X_1) = p(1-p) < +\infty$ alors par le TLC,

$$\sqrt{n}\,\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Lemme de Slutsky

Propriété

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. Soit $g : \mathbb{R} \to \mathbb{R}$ une fonction continue. Alors

$$X_n \xrightarrow[n \to +\infty]{\mathcal{L}} X \Longrightarrow g(X_n) \xrightarrow[n \to +\infty]{\mathcal{L}} g(X)$$

Lemme de Slutsky

Soient $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ telles que $X_n \xrightarrow[n \to +\infty]{\mathcal{L}} X$ et $Y_n \xrightarrow[n \to +\infty]{\mathbb{P}} c$ alors

$$X_n + Y_n \xrightarrow[n \to +\infty]{\mathcal{L}} X + c \text{ et } X_n Y_n \xrightarrow[n \to +\infty]{\mathcal{L}} cX$$

Exercice 1

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. i.i.d de loi $\mathcal{B}(p)$. Montrez que

$$\sqrt{n}\,\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Exercice 2

Soit $X_1,\ldots,X_n$ des v.a. i.i.d de loi uniforme sur $[0,1]$.
On pose $Z_n = n\min(X_1,\ldots,X_n)$.
Déterminez pour tout $n\geq 1$ la fonction de répartition de Z_n et étudiez la convergence en loi de $(Z_n)_{n\geq 1}$.

Indication : Vous pouvez vérifier que la fonction de répartition d'une loi exponentielle de paramètre 1 vaut

$$F(t) = \left(1-e^{-t}\right)\mathbb{1}_{t>0}.$$

Loi normale sur $\mathbb{R}$

Définition

Soit $m \in \mathbb{R}$ et $\sigma > 0$. On dit que X suit une **loi normale** (ou gaussienne) de paramètres (m, σ^2) si la loi de X a pour densité

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Notation : $X \sim \mathcal{N}(m, \sigma^2)$

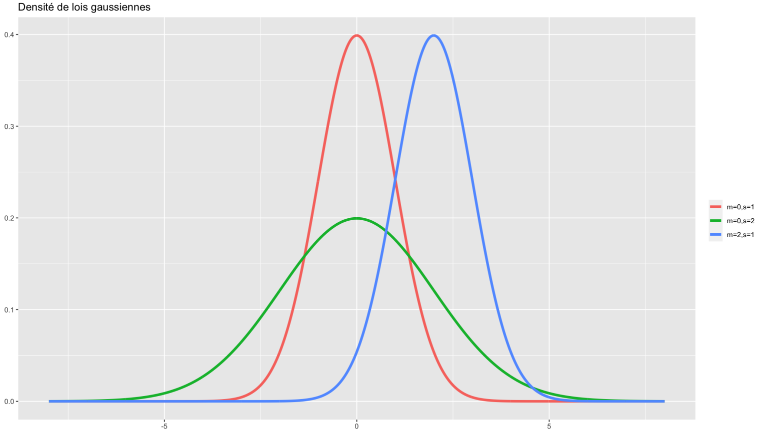
Propriétés

- $\mathbb{E}[X] = m$ et $\text{Var}(X) = \sigma^2$
- $\Phi_X(t) = \exp\left[itm - \frac{t^2\sigma^2}{2}\right], \quad t \in \mathbb{R}$
- la fonction de répartition de X n'a pas d'expression analytique $\Longrightarrow$ table de la loi $\mathcal{N}(0,1)$ pour des valeurs approchées de F_X et des valeurs de quantiles.

Loi normale sur $\mathbb{R}$

Propriétés

- Si $X \sim \mathcal{N}(m, \sigma^2)$ alors $Y = \frac{X-m}{\sigma} \sim \mathcal{N}(0, 1)$
- Si $Y \sim \mathcal{N}(0, 1)$ alors $X = m + \sigma Y \sim \mathcal{N}(m, \sigma^2)$.



Loi du Khi-deux $\chi^2(d)$

Définition

Soient $Y_1, \ldots, Y_d$ des v.a. i.i.d $\mathcal{N}(0,1)$. La loi de

$$Y_1^2 + \ldots + Y_d^2$$

est appelée **loi du khi-deux** à d degrés de liberté, et notée $\chi^2(d)$.

Propriétés

- Si $V \sim \chi^2(d)$ alors $\mathbb{E}[V] = d$ et $\text{Var}(V) = 2d$.
- Si $V_1 \sim \chi^2(d_1)$, $V_2 \sim \chi^2(d_2)$ et $V_1 \perp\!\!\!\perp V_2$ alors

$$V_1 + V_2 \sim \chi^2(d_1 + d_2).$$

Loi de Student $\mathcal{T}(d)$

Définition

Soient U et V deux v.a. telles que
$U \sim \mathcal{N}(0,1)$, $V \sim \chi^2(d)$ et $U \perp\!\!\!\perp V$. Alors la loi de

$$\frac{U}{\sqrt{V/d}}$$

est appelée **loi de Student** à d degrés de liberté, notée $\mathcal{T}(d)$.

Propriétés

La loi du Student est une loi symétrique (i.e. T et $-T$ ont même loi).
En particulier, $\mathbb{E}[T] = 0$.

Loi de Fisher $\mathcal{F}(d_1, d_2)$

Définition

Soit $V_1 \sim \chi^2(d_1)$, $V_2 \sim \chi^2(d_2)$ et $V_1 \perp\!\!\!\perp V_2$. La loi de

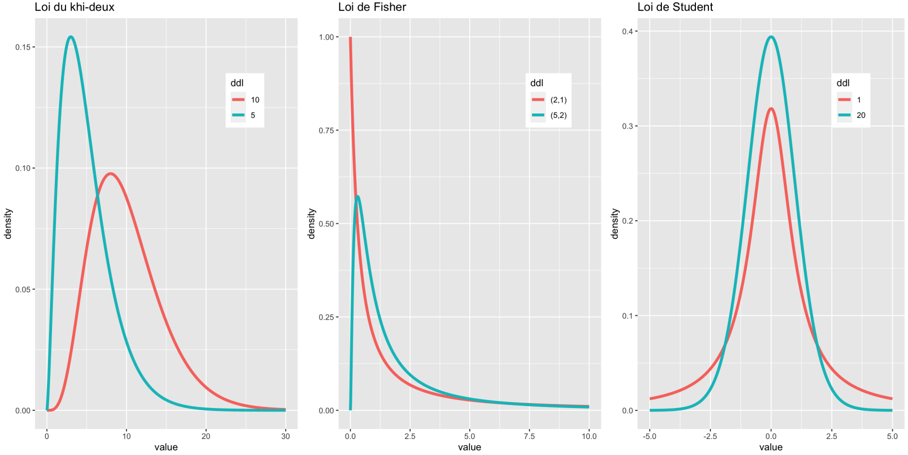
$$\frac{V_1/d_1}{V_2/d_2}$$

est appelée **loi de Fisher** de paramètres (d_1, d_2), notée $\mathcal{F}(d_1, d_2)$.

Propriétés

- Si $F \sim \mathcal{F}(d_1, d_2)$, alors $\frac{1}{F} \sim \mathcal{F}(d_2, d_1)$.
- Si q_α est le α-quantile d'une $\mathcal{F}(d_1, d_2)$, alors $\frac{1}{q_\alpha}$ est le $(1-\alpha)$-quantile d'une $\mathcal{F}(d_2, d_1)$.

Représentation des densités



Exercices

Exercice 1
Un jardinier récolte des tomates dont le poids est modélisé par une loi normale de moyenne $m = 200$g et d'écart-type $\sigma = 40$g. Il dépose les tomates cueillies au fur et à mesure dans une caisse qui peut supporter un poids maximal de 2kg.
1) Quelle est la probabilité que le jardinier cueille une tomate de plus de 250g.
2) Déterminer la valeur d telle que la probabilité que l'écart entre le poids d'une tomate et m dépasse d soit égale à 0.1.
3) Déterminer le nombre maximum de tomates que l'on peut cueillir pour que la probabilité de la surcharge de la caisse n'excède pas 0.01.

Exercice 2 : Table de $\mathcal{N}(0,1)$
1) Soit X une v.a.r de loi normale centrée réduite. Calculer $\mathbb{P}(-0.3 < X < 0.1)$ et déterminer t pour que $\mathbb{P}(X > t) = 0.7$.
2) Soit Y une v.a.r de loi normale, d'espérance 1 et de variance 4.
Calculer $\mathbb{P}(-1 < Y < 2)$, $\mathbb{P}(Y^2 < 4)$ et déterminer t pour que $\mathbb{P}(0 < Y < t) = 0.5$.

Exercice 3 Soit X, Y suivent une loi $\mathcal{N}(0,1)$, $Z \sim \mathcal{N}(1,2)$, X, Y et Z sont indépendants.
Déterminez la loi des variables aléatoires suivantes :

$$A = X - Z \qquad B = X^2 + Y^2$$

$$C = \frac{Z-1}{\sqrt{X^2+Y^2}} \qquad D = \frac{(Z-1)^2}{X^2+Y^2}$$

Définition

Un vecteur aléatoire X à valeurs dans $\mathbb{R}^d$ est dit **gaussien** si toute combinaison linéaire de ses composantes est une v.a. gaussienne.
Si $X = (X_1, \cdots, X_d)'$ est un vecteur gaussien, on définit
son **vecteur moyenne** $\mathbb{E}[X]$ par

$$m := \mathbb{E}[X] = (\mathbb{E}[X_1], \cdots, \mathbb{E}[X_d])'$$

sa **matrice de variance-covariance** Σ_X par

$$\Sigma_X = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \ddots & \ldots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathrm{Var}(X_d) \end{pmatrix}.$$

Notation : $X \sim \mathcal{N}_d(m, \Sigma_X)$

Définition

Définition

Soit $X = (X_1, \cdots, X_d)' \sim \mathcal{N}_d(m, \Sigma_X)$.

- La **fonction caractéristique** de X est

$$\forall t \in \mathbb{R}^d,\ \Phi_X(t) = \mathbb{E}\left[\exp(it'X)\right] = \exp\left(it'm - \frac{1}{2}t'\Sigma_X t\right).$$

- La **densité** de X est

$$\forall x \in \mathbb{R}^d,\ \ f_{m,\Sigma_X}(x) = (2\pi)^{-\frac{d}{2}}|\Sigma_X|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-m)'\Sigma_X^{-1}(x-m)\right]$$

- La fonction caractéristique et la densité caractérisent la loi.
- La loi de X est entièrement déterminée par m et Σ_X.

Propriétés

Propriétés

- La matrice Σ_X est symétrique

$$\Sigma_{X,(i,j)} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = \Sigma_{X,(j,i)}.$$

- Si $(X_1, \cdots, X_n)$ n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ alors

$$X = (X_1, \cdots, X_n)' \sim \mathcal{N}(\mu \mathbb{1}_n, \sigma^2 I_n)$$

Propriétés de linéarité

Soit $X \sim \mathcal{N}_d(m, \Sigma_X)$.
Pour toute matrice $A \in \mathcal{M}_{p\,d}(\mathbb{R})$ et pour tout vecteur $b \in \mathbb{R}^p$

$$AX + b \sim \mathcal{N}_{p}(Am + b, A\Sigma_X A').$$

Propriétés

Propriété d'indépendance

Soit $X = (X_1, \cdots, X_d)'$ un vecteur gaussien.
Les composantes $X_1, \cdots, X_d$ sont deux à deux indépendantes
$\Longleftrightarrow$ Σ_X est diagonale : $\forall i \neq j$, $\mathrm{Cov}(X_i, X_j) = 0$.

Remarque :

Si $X = (X_1, \cdots, X_d)'$ est un vecteur gaussien alors les composantes X_i sont des variables aléatoires gaussiennes **mais la réciproque est fausse**.
Soient $X \sim \mathcal{N}(0,1) \perp\!\!\!\perp Y \sim \mathcal{B}(0.5)$. Alors $X_1 = X$ et $X_2 = (2Y-1)X$ sont des v.a. gaussiennes mais $(X_1, X_2)'$ n'est pas un vecteur gaussien.
On note que $\mathrm{Cov}(X_1, X_2) = 0$ mais que X_1 et X_2 ne sont pas indépendantes.

Théorème de Cochran

Théorème de Cochran

Soient $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et $X = (X_1, \ldots, X_n)' \in \mathbb{R}^n$.

Soit $E_1 \oplus E_2 \oplus \ldots \oplus E_p$ une décomposition de $\mathbb{R}^n$ en p sous-espaces orthogonaux de dimensions respectives $r_1, \ldots, r_p$.

On note $P_{E_i}X$ la projection orthogonale de X sur E_i.

Alors les vecteurs $P_{E_1}X, P_{E_2}X, \ldots, P_{E_p}X$ sont indépendants et pour tout i,

$$\|P_{E_i}X\|^2 \sim \sigma^2\chi^2(r_i).$$

Corollaire

Corollaire

Soient $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(m, \sigma^2)$.

- Les variables aléatoires

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

  sont indépendantes.
- $\bar{X}_n \sim \mathcal{N}(m, \sigma^2/n)$,
- $S_n^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$.
- Il en résulte que la variable aléatoire

$$\sqrt{n}\,\frac{\bar{X}_n - m}{S_n} \sim \mathcal{T}(n-1).$$

Très important pour les IC et tests dans le cas gaussien !

Exercice 1

Soient $X_1, X_2, \ldots, X_n$ des v.a. i.i.d de loi $\mathcal{N}(m, \sigma^2)$. Soient

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

1. Montrer à l'aide des propriétés des vecteurs gaussiens que

$$\bar{X}_n \sim \mathcal{N}(m, \frac{\sigma^2}{n})$$

2. En appliquant le théorème de Cochran sur un vecteur gaussien Y bien choisi et en considérant la décomposition orthogonale de $\mathbb{R}^n = E_1 \oplus E_1^{\perp}$ avec $E_1 = \text{Vect}\,((1, 1, \ldots, 1)')$, montrez que
   - $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$
   - $\bar{X}_n \perp\!\!\!\perp S_n^2$

Exercice 2

Soit $U = (X, Y, Z)'$ un vecteur gaussien d'espérance m et de matrice de variance-covariance Σ où

$$m = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \text{ avec } 0 \leq \rho$$

1. Quelle est la loi de Z ?
2. Pour quelle(s) valeur(s) de ρ les variables aléatoires X, Y et Z sont-elles indépendantes ?
3. Soit $V = \begin{pmatrix} X + Y + \alpha \\ X - Y + \sqrt{2\rho}Z \end{pmatrix}$ avec $\alpha \in \mathbb{R}$. Déterminez la loi de V.
4. Que peut-on en conclure sur les coordonnées du vecteur V ?
5. Pour quelles valeurs de α et ρ, le vecteur V est-il centré ?
6. Dans ce cas, on obtient que $V \sim \mathcal{N}_2(0_2, 6I_2)$. Quelle est la loi de $\|V\|_2^2$?

n-échantillon

Définition

On appelle ***n*-échantillon** issu d'une loi P (ou "issu de X" si X est une variable aléatoire de loi P) toute suite de n variables aléatoires $(X_1, X_2, \ldots, X_n)$ **indépendantes et identiquement distribuées (i.i.d)** selon P.
Une fois l'expérience aléatoire effectuée, on obtient un n-échantillon de valeurs observées $(x_1, x_2, \ldots, x_n)$, c'est-à-dire toute réalisation de l'échantillon aléatoire $(X_1, X_2 \ldots, X_n)$.

Exemple du sondage

Les **variables aléatoires** $(X_1, X_2, \ldots, X_n)$ modélisent la réponse binaire de n personnes interrogées
Les **nombres** $(x_1, x_2, \ldots, x_n)$ représentent les réponses binaires de ces personnes une fois la question posée

Par la suite, on considère **un paramètre inconnu** θ (ex : la moyenne, la variance ou encore la proportion) associé à la loi P de X.

Estimateur - estimation

Définition

On appelle **estimateur** du paramètre θ toute variable aléatoire $\hat{\theta}_n$ fonction du n-échantillon $(X_1, X_2, \ldots, X_n)$ issu de X.
La valeur observée (donc après expérience aléatoire) d'un estimateur est appelée **estimation**.

Exemple du sondage

$\underline{X} = (X_1, \ldots, X_n)$ n-échantillon de loi $\mathcal{B}(p)$. Le paramètre inconnu : $\theta = p$
Un estimateur :

$$\hat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Si $n = 8$ personnes ont été interrogées, et que l'on dispose ainsi des réponses suivantes : $(x_1, \ldots, x_8) = (0, 1, 1, 0, 1, 0, 0, 0)$, alors l'estimation associées : $\frac{1}{8}\sum_{i=1}^{8} x_i = \frac{3}{8}$.

Estimateur consistant

Définition

Un estimateur $\hat{\theta}_n$ de θ est dit **consistant** si $\hat{\theta}_n \xrightarrow[n \to +\infty]{\mathbb{P}} \theta$.

Exemple du sondage

$\underline{X} = (X_1, \ldots, X_n)$ n-échantillon de loi $\mathcal{B}(p)$.
En utilisant la LGN, on a

$$\hat{\theta}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n \to +\infty]{\mathbb{P}} \mathbb{E}[X_1] = p$$

donc $\hat{\theta}_n$ est un estimateur consistant de p.

(asymptotiquement) sans biais / biaisé

Définition

- On appelle **biais** d'un estimateur $\hat{\theta}_n$ de θ la quantité $\mathbb{E}[\hat{\theta}_n] - \theta$.
- $\hat{\theta}_n$ est un estimateur **sans biais** de θ si son biais est nul : $\mathbb{E}[\hat{\theta}_n] = \theta$
  Il est dit **biaisé** sinon.
- $\hat{\theta}_n$ est dit **asymptotiquement sans biais** si $\mathbb{E}[\hat{\theta}_n] \underset{n\to+\infty}{\longrightarrow} \theta$.

Exemple du sondage

$$\mathbb{E}[\hat{\theta}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{np}{n} = p$$

donc $\hat{\theta}_n$ est un estimateur sans biais de p.

Moyenne empirique et variance empirique

Propriétés

- La moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur consistant et sans biais de $\mathbb{E}[X]$.
- La variance empirique
$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} X_i^2\right) - n\left(\bar{X}_n\right)^2\right\}$$
est un estimateur consistant et sans biais de la variance $\mathrm{Var}(X)$.
$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ est également un estimateur consistant de $\mathrm{Var}(X)$ mais biaisé.
On préférera donc l'estimateur non biaisé S_n^2.

écart quadratique moyen

Définition

Soit $\hat{\theta}_n$ un estimateur consistant de θ. **L'écart quadratique moyen** est défini par

$$\mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right] = \mathrm{Var}(\hat{\theta}_n) + \left(\underbrace{\mathbb{E}[\hat{\theta}_n] - \theta}_{\text{biais}}\right)^2 .$$

Propriété

étant donnés deux estimateurs $\hat{\theta}_{n,1}$ et $\hat{\theta}_{n,2}$ de θ, on choisira celui qui a le plus faible écart quadratique moyen.

Définition

étant donnés deux estimateurs **sans biais** $\hat{\theta}_{n,1}$ et $\hat{\theta}_{n,2}$ de θ, on dira que $\hat{\theta}_{n,1}$ est **plus efficace** que $\hat{\theta}_{n,2}$ si $\mathrm{Var}(\hat{\theta}_{n,1}) \leq \mathrm{Var}(\hat{\theta}_{n,2})$.

Méthode des moments

Estimateur des moments

La **méthode des moments** consiste à déterminer une fonction g continue et inversible, et une fonction φ continue telles que

$$g(\theta) = \mathbb{E}[\varphi(X_1)].$$

L'estimateur des moments est alors défini par

$$\hat{\theta} = g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\right)$$

Exemples

- $\theta = \mathbb{E}[X_1]$, $g = \varphi = Id$ et $\hat{\theta} = \bar{X}_n$
- $X_1, \ldots, X_n$ i.i.d de loi exponentielle $\mathcal{E}(\theta)$.
  $\mathbb{E}[X_1] = \frac{1}{\theta}$ donc $g(x) = \frac{1}{x}$ et $\varphi = Id$. On obtient donc $\hat{\theta} = (\bar{X}_n)^{-1}$.

Vraisemblance

Définition

Soit $\{P_\theta, \theta \in \Theta\}$ une famille de loi de probabilité. Pour tout $\theta \in \Theta$,

- $f(.;\theta)$ la densité / à la mesure de Lebesgue si P_θ continue
- $f(x;\theta) = \mathbb{P}_\theta(X = x),\ \forall x \in X(\Omega)$ si P_θ discrète

On suppose qu'il existe $\theta^* \in \Theta$ tel que les $X_i \sim P_{\theta^*}$.
On appelle **vraisemblance** de l'échantillon $\underline{X} = (X_1, \ldots, X_n)$ la fonction

$$\theta \mapsto L(\underline{X};\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

Exemple

Soit $\underline{X} = (X_1, \ldots, X_n)$ n-échantillon de loi de $\mathcal{B}(p)$ La vraisemblance de $\underline{X}$ est

$$p \in [0,1] \mapsto \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{S_n}(1-p)^{n-S_n} \text{ avec } S_n = \sum_{i=1}^{n} X_i$$

Estimateur du maximum de vraisemblance (EMV)

Définition

On appelle **estimateur du maximum de vraisemblance** (EMV) de θ associé à l'échantillon $\underline{X}$ toute v.a. $\widehat{\theta}_{\text{MV}}$ définie, quand elle existe, par :

$$\widehat{\theta}_{\text{MV}} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\underline{X}; \theta).$$

Remarques

- $\widehat{\theta}_{\text{MV}}$ peut ne pas exister, ou ne pas être unique si elle existe.
- Pour déterminer $\widehat{\theta}_{\text{MV}}$, plus facile de maximiser la **logvraisemblance** $\ell(\underline{X}; \theta) = \ln\left[L(\underline{X}; \theta)\right]$.

Exemple

Soit $(X_1, \ldots, X_n)$ i.i.d de loi de $\mathcal{B}(p)$. La logvraisemblance :
$\ell(\underline{X}; p) = S_n \ln(p) + (n - S_n) \ln(1 - p)$ avec $S_n = \sum_{i=1}^n X_i$

Score - Information de Fisher

Définitions

- On appelle **fonction score** la fonction

$$\underline{x} \mapsto S(\underline{x};\theta) = \frac{\partial \ell(\underline{x};\theta)}{\partial \theta}$$

- On appelle **information de Fisher** apportée par l'échantillon $\underline{X}$ sur le paramètre θ la quantité suivante, si elle existe,

$$I_n(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln L(\underline{X};\theta)}{\partial \theta}\right)^2\right] = \mathbb{E}\left[\left(\frac{\partial \ell(\underline{X};\theta)}{\partial \theta}\right)^2\right] \geq 0$$

Propriétés

Pour un modèle paramétrique régulier (admis), on a

- $\mathbb{E}_\theta[S(\underline{X};\theta)] = 0$ et $\mathrm{Var}_\theta(S(\underline{X};\theta)) = I_n(\theta)$
- $I_n(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\underline{X};\theta)}{\partial \theta^2}\right]$.

Borne de Cramer-Rao

Propriété / Définition

Soit $\underline{X} = (X_1, \ldots, X_n)$ issu d'un modèle paramétrique régulier.
Soit $T_n(\underline{X})$ un estimateur **sans biais** de $g(\theta)$.
Sous certaines hypothèses,

$$\mathrm{Var}(T_n(\underline{X})) \geq \underbrace{\frac{g'(\theta)^2}{I_n(\theta)}}_{:=B_n(\theta)}.$$

$B_n(\theta)$ est appelée la **borne de Cramer-Rao**.
Dans le cas particulier où $g(\theta) = \theta$, on a $B_n(\theta) = (I_n(\theta))^{-1}$.

Définition

On dit que $T_n(\underline{X})$ est un estimateur **efficace** de $g(\theta)$ si

$$\mathrm{Var}(T_n(\underline{X})) = B_n(\theta)$$

Exercice

Exercice 1

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$.

1. Déterminez l'EMV $\hat{\mu}$ pour le paramètre μ.
2. L'estimateur $\hat{\mu}$ est-il sans biais ?
3. L'estimateur $\hat{\mu}$ est-il consistant ?
4. L'estimateur $\hat{\mu}$ est-il efficace ?

Exercice 2

Soit $(Y_1, \ldots, Y_n)$ un n-échantillon de densité de probabilité

$$f_Y(y) = \frac{|y|}{\mu} \exp\left(-\frac{y^2}{\mu}\right)$$

avec $\mu > 0$ paramètre inconnu.
Reprenez les questions de l'exercice 1.
Indication : vérifiez que $X = Y^2$ suit une loi exponentielle $\mathcal{E}(\frac{1}{\mu})$.

Estimation par intervalle

En donnant une estimation ponctuelle de $\hat{\theta}_n$, on ne fait alors aucune évaluation de l'erreur d'estimation commise.

Le principe de l'estimation par intervalle de confiance est au contraire d'associer au n-échantillon observé un intervalle de valeurs possibles, avec une évaluation de l'erreur commise.
On se fixe **un risque** $\alpha \in [0, 1]$ (en général 1%, 5% ou 10%).

Définition

Un **intervalle de confiance** (IC) pour θ de niveau de confiance $1 - \alpha$ (ou au risque α) est un intervalle $IC_{1-\alpha}(\theta)$ dont les bornes dépendent du n-échantillon $(X_1, \ldots, X_n)$ issu de X, et tel que

$$\mathbb{P}\left(\theta \in IC_{1-\alpha}(\theta)\right) = 1 - \alpha.$$

IC pour la moyenne d'une population gaussienne

- Contexte : Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$.
- But : μ paramètre inconnu. Construire $IC_{1-\alpha}(\mu)$.
- 2 cas à distinguer :
  - La variance σ^2 est connue
  - La variance σ^2 est inconnue

IC à variance connue

Il faut savoir faire le raisonnement, ne pas apprendre la formule finale !

- Estimateur pour μ : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$
- Loi de $\bar{X}_n$: $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ donc

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

Cette statistique ne dépend que du paramètre inconnu cible μ, le reste est connu. On peut donc continuer !

- Soit $z_{1-\alpha/2}$ est le $1-\alpha/2$ quantile de la loi $\mathcal{N}(0,1)$:
si $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{P}(|Z| \leq z_{1-\alpha/2}) = 1-\alpha \Longleftrightarrow \mathbb{P}(Z \leq z_{1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

IC à variance connue

- Ainsi

$$\begin{aligned}
& \mathbb{P}\left(\left|\sqrt{n}\,\frac{\bar{X}_n-\mu}{\sigma}\right| \le z_{1-\alpha/2}\right) = 1-\alpha \\
\Longleftrightarrow\quad & \mathbb{P}\left(-z_{1-\alpha/2} \le \sqrt{n}\,\frac{\bar{X}_n-\mu}{\sigma} \le z_{1-\alpha/2}\right) = 1-\alpha \\
\Longleftrightarrow\quad & \mathbb{P}\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \le \mu \le \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right) = 1-\alpha \\
\Longleftrightarrow\quad & \mathbb{P}(\mu \in IC_{1-\alpha}(\mu)) = 1-\alpha
\end{aligned}$$

avec

$$IC_{1-\alpha}(\mu) = \left[\bar{X}_n \pm \underbrace{\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}}_{\text{marge d'erreur } d_n}\right]$$

Rem : Lorsque n augmente, la longueur de l'IC ($2d_n$) diminue et l'on gagne en précision.

IC à variance inconnue

Il faut savoir faire le raisonnement, ne pas apprendre la formule finale !

- Estimateur pour μ : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$
- Loi de $\bar{X}_n$: $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ donc

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

mais σ^2 INCONNU !

- On estime σ^2 par $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$
- D'après le **Théorème de Cochran** (lien),

$$(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1) \text{ et } S_n^2 \perp\!\!\!\perp \bar{X}_n$$

donc

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n} \sim \mathcal{T}(n-1).$$

Cette fois le seul paramètre inconnu dans cette statistique est la cible μ, on peut continuer !

IC à variance inconnue

- Soit $t_{n-1,1-\alpha/2}$ est le $1-\alpha/2$ quantile de la loi $\mathcal{T}(n-1)$: si $T \sim \mathcal{T}(n-1)$,

$$\mathbb{P}(|T| \leq t_{n-1,1-\alpha/2}) = 1-\alpha \Longleftrightarrow \mathbb{P}(T \leq t_{n-1,1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

- Ainsi

$$\begin{aligned}
& \mathbb{P}\left(\left|\sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n}\right| \leq t_{n-1,1-\alpha/2}\right) = 1-\alpha \\
\Longleftrightarrow\ & \mathbb{P}\left(-t_{n-1,1-\alpha/2} \leq \sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n} \leq t_{n-1,1-\alpha/2}\right) = 1-\alpha \\
\Longleftrightarrow\ & \mathbb{P}\left(\bar{X}_n - \frac{S_n}{\sqrt{n}}\,t_{n-1,1-\alpha/2} \leq \mu \leq \bar{X}_n + \frac{S_n}{\sqrt{n}}\,t_{n-1,1-\alpha/2}\right) = 1-\alpha \\
\Longleftrightarrow\ & \mathbb{P}\left(\mu \in \left[\bar{X}_n \pm \frac{S_n}{\sqrt{n}}\,t_{n-1,1-\alpha/2}\right]\right) = 1-\alpha
\end{aligned}$$

IC pour la moyenne d'une population non gaussienne

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi non gaussienne.

On désire construire un IC pour $\mathbb{E}[X]$ de niveau de confiance $1 - \alpha$.

Par le Théorème de la Limite Centrale (lien) :

$$\sqrt{n}\, \frac{\bar{X}_n - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

On peut alors construire un intervalle de confiance de **niveau asymptotique** $1 - \alpha$, c'est-à-dire que $\mathbb{P}(\mathbb{E}[X] \in IC) \underset{n \to +\infty}{\longrightarrow} 1 - \alpha$, en différenciant les cas de variance connue et inconnue.

IC pour la moyenne d'une population non gaussienne

Cas de la variance connue

$$IC_{1-\alpha}(\mathbb{E}[X]) = \left[\bar{X}_n \pm \sqrt{\frac{\text{Var}(X)}{n}}\, z_{1-\alpha/2}\right].$$

Cas de la variance inconnue

Comme S_n^2 est un estimateur consistant de $\text{Var}(X)$, par le **lemme de Slutsky** (lien) on conserve la propriété de convergence en loi

$$\sqrt{n}\left(\frac{\bar{X}_n - \mathbb{E}[X]}{S_n}\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

donc

$$IC_{1-\alpha}(\mathbb{E}[X]) = \left[\bar{X}_n \pm \frac{S_n}{\sqrt{n}}\, z_{1-\alpha/2}\right].$$

IC pour une proportion p

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi $\mathcal{B}(p)$, p inconnu.
On désire construire un IC de niveau de confiance $1 - \alpha$ pour $p = \mathbb{E}[X_1]$.

- Par le TLC (lien),

$$\sqrt{n}\left(\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

- $\bar{X}_n \xrightarrow[n \to +\infty]{\mathbb{P}} p$ donc $\bar{X}_n(1-\bar{X}_n) \xrightarrow[n \to +\infty]{\mathbb{P}} p(1-p)$
- Par le lemme de Slutsky (lien),

$$\sqrt{n}\left(\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Exemple pour une proportion p

- Soit $z_{1-\alpha/2}$ le $1-\alpha/2$ quantile d'une loi $\mathcal{N}(0,1)$.
- Ainsi,

$$\mathbb{P}\left(\left|\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\right| \leq z_{1-\alpha/2}\right) \underset{n\to+\infty}{\longrightarrow} \mathbb{P}(|Z| \leq z_{1-\alpha/2}) = 1-\alpha$$

$$\mathbb{P}\left(p \in \left[\bar{X}_n \pm z_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]\right) \underset{n\to+\infty}{\longrightarrow} 1-\alpha$$

donc

$$IC_{1-\alpha}(p) = \left[\bar{X}_n \pm z_{1-\alpha/2}\frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}\right].$$

IC pour la variance

Il faut savoir faire le raisonnement, ne pas apprendre la formule finale !

- Estimateur pour σ^2 : $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$
- Loi de S_n^2 : par le théorème de Cochran (lien)

$$(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$$

- La loi du $\chi^2(n-1)$ n'étant pas symétrique, on prend $q_{\alpha/2}$ et $q_{1-\alpha/2}$ les $\alpha/2$ et $1-\alpha/2$ quantiles de la loi du $\chi^2(n-1)$ d'où, si $U \sim \chi^2(n-1)$,

$$\mathbb{P}(q_{\alpha/2} \leq U \leq q_{1-\alpha/2}) = 1-\alpha$$

- Ainsi,

$$\mathbb{P}(q_{\alpha/2} \leq (n-1)S_n^2/\sigma^2 \leq q_{1-\alpha/2}) = 1-\alpha$$

$$\iff \quad \mathbb{P}\left(\frac{(n-1)S_n^2}{q_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{q_{\alpha/2}}\right) = 1-\alpha$$

Exercice 1

Le responsable qualité d'une usine de boissons gazeuses s'intéresse à la stabilité du système de remplissage des bouteilles. Il décide de prélever $n = 16$ bouteilles remplies par la machine et mesure la hauteur du liquide dans chacune des bouteilles. On pose X_i la hauteur de liquide dans la $i$ème bouteille et on suppose que $(X_1, \ldots, X_{16})$ forment un 16-échantillon de loi $\mathcal{N}(m, \sigma^2)$. Les résultats numériques obtenus sont $\sum_{i=1}^{16} x_i = 376$ et $\sum_{i=1}^{16} x_i^2 = 8837$.

1. Proposez un estimateur pour m, précisez sa loi et donnez une estimation pour m.
2. On suppose dans cette question que $\sigma^2 = 0.05$. Construisez un intervalle de confiance pour m au niveau de confiance de 0,95.
3. Le responsable qualité met en doute la valeur de σ^2.
   a) Proposez un estimateur et une estimation pour σ^2.
   b) Construisez un intervalle de confiance pour m au niveau de confiance de 0,95 en supposant σ^2 inconnu.

Exercice 2

La durée de vie d'un composant électronique suit une loi exponentielle de paramètre $\frac{1}{\theta}$ avec $\theta > 0$. On observe les durées de vie de n composants notées $X_1, \ldots, X_n$.

1. Proposez un estimateur pour θ. Calculez l'espérance et la variance de cet estimateur.
2. Construisez un intervalle bilatéral asymptotique de niveau de confiance 0.95 pour θ.

Contexte

- On se donne un modèle statistique, c'est-à-dire une famille de probabilité : $\{P_\theta,\ \theta \in \Theta\}$
- On considère un n-échantillon $\underline{X} = (X_1, \ldots, X_n)$ dont la loi est supposée appartenir à $\{P_\theta,\ \theta \in \Theta\}$.
- Soit $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$ tels que $\Theta_0 \cap \Theta_1 = \varnothing$.
- A partir de l'échantillon $\underline{X}$, on souhaite construire **une règle de décision (zone de rejet)** pour décider entre **deux hypothèses** :

$$\mathcal{H}_0 : \theta \in \Theta_0 \text{ contre } \mathcal{H}_1 : \theta \in \Theta_1 .$$

Zone de rejet - Erreur de première espèce

Définition

Faire **un test de niveau** $\alpha \in]0,1[$ de l'hypothèse $\mathcal{H}_0 : \theta \in \Theta_0$ contre $\mathcal{H}_1 : \theta \in \Theta_1$ à partir de $\underline{X}$ c'est se donner une **zone de rejet** $\mathcal{R}_\alpha$ tel que

$$\underbrace{\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left\{ \underline{X} \in \mathcal{R}_\alpha \right\}}_{\text{Taille du test}} \leq \alpha.$$

On applique alors la **règle de décision** : on rejette $\mathcal{H}_0$ si $\underline{X}^{obs} \in \mathcal{R}_\alpha$.
On parle aussi **d'erreur de première espèce** pour la taille du test.

Erreur de seconde espèce - Puissance

Définition

On appelle **risque de second espèce** du test basé sur la zone de rejet $\mathcal{R}_\alpha$ la fonction

$$\beta : \theta \in \Theta_1 \mapsto \mathbb{P}_\theta(\underline{X} \notin \mathcal{R}_\alpha),$$

et **l'erreur de seconde espèce** correspond à :

$$\sup_{\theta \in \Theta_1} \beta(\theta) = \sup_{\theta \in \Theta_1} \mathbb{P}_\theta(\underline{X} \notin \mathcal{R}_\alpha).$$

Définition

On appelle **fonction puissance** du test basé sur la zone de rejet $\mathcal{R}_\alpha$ l'application

$$\Pi : \theta \in \Theta_1 \mapsto \mathbb{P}_\theta\left(\underline{X} \in \mathcal{R}_\alpha\right) = 1 - \beta(\theta) \in [0, 1].$$

Test plus puissant ...

Parmi les tests de même niveau, on préfère le plus puissant.

Définition

On dira que le test basé sur la région de rejet $\mathcal{R}_\alpha$ est meilleur que celui basé sur la région $\tilde{\mathcal{R}}_\alpha$ s'ils sont tous les deux de niveau α et que

$$\forall \theta \in \Theta_1, \; \Pi(\theta) = \mathbb{P}_\theta\left(\underline{X} \in \mathcal{R}_\alpha\right) \geq \tilde{\Pi}(\theta) = \mathbb{P}_\theta\left(\underline{X} \in \tilde{\mathcal{R}}_\alpha\right).$$

Définition

On dit que le test basé sur la région de rejet $\mathcal{R}_\alpha$ est **uniformément plus puissant (UPP)** au niveau α si :

1. $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\underline{X} \in \mathcal{R}_\alpha) \leq \alpha$
2. Pour toute zone de rejet $\tilde{\mathcal{R}}_\alpha$ telle que $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\underline{X} \in \tilde{\mathcal{R}}_\alpha) \leq \alpha$,

$$\forall \theta \in \Theta_1, \; \Pi(\theta) = \mathbb{P}_\theta(\underline{X} \in \mathcal{R}_\alpha) \geq \tilde{\Pi}(\theta) = \mathbb{P}_\theta(\underline{X} \in \tilde{\mathcal{R}}_\alpha).$$

Importance du choix des hypothèses

Attention : Les erreurs de première et de deuxième espèce ne sont pas de la même importance.
Le choix des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ est donc très important.

Exemple

Soit θ la moyenne du niveau de radioactivité en picocuries par litre d'eau. La valeur $\theta = 5$ est considérée comme la valeur critique entre l'eau potable et l'eau non potable.
On teste $\mathcal{H}_0 : \theta \geq 5$ contre $\mathcal{H}_1 : \theta < 5$.
L'erreur de première espèce a pour conséquence de laisser boire une eau toxique.
L'erreur de deuxième espèce conduit seulement à interdire la consommation alors que l'eau est potable.

Vocabulaire sur les hypothèses

- On parle d'**hypothèse simple** si elle correspond à une seule loi possible ($\Theta_0 = \{\theta_0\}$) et d'**hypothèse composée** sinon.

- Si l'hypothèse $\mathcal{H}_1$ est de type "$\theta > \theta_0$" ou "$\theta < \theta_0$", où θ est le paramètre d'intérêt, le test est dit **unilatéral**.

- Si l'hypothèse $\mathcal{H}_1$ est du type $\theta \neq \theta_0$, le test est dit **bilatéral**.

Démarche générale pour construire un test

- On définit les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$, basées sur un paramètre inconnu θ.
- On considère un estimateur $\hat{\theta}_n$ du paramètre d'intérêt θ.
- On transforme $\hat{\theta}_n$ pour se ramener à une **statistique de test** $T(\underline{X})$ (ne dépend d'aucun paramètre inconnu !) dont on connait la loi sous $\mathcal{H}_0$
- On étudie le comportement de la statistique de test sous $\mathcal{H}_1$ pour déterminer la forme de la zone de rejet.
- On calibre la zone de rejet pour assurer que le test soit de niveau α, ce qui détermine précisément le domaine sur lequel l'hypothèse $\mathcal{H}_0$ sera rejetée.
- On évalue, si possible, la puissance du test.

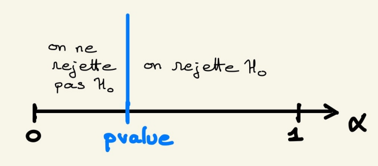
Définition

La **p-valeur** d'un test est le plus petit niveau du test pour lequel on rejette $\mathcal{H}_0$.
Cette quantité est calculée à partir de l'échantillon **observé** $\underline{x}$
Ex : Si la zone de test est de la forme $\mathcal{R}_\alpha = \{T(\underline{X}) > t_\alpha\}$ alors

$$\begin{aligned} \text{Rejet de } \mathcal{H}_0 \quad &\Leftrightarrow \quad T^{obs} := T(\underline{x}) > t_\alpha \\ &\Leftrightarrow \quad \mathbb{P}_{\mathcal{H}_0}(T(\underline{X}) > T^{obs}) \leq \mathbb{P}_{\mathcal{H}_0}(T(\underline{X}) > t_\alpha) := \alpha \end{aligned}$$

donc la pvaleur est $\mathbb{P}_{\mathcal{H}_0}(T(\underline{X}) > T^{obs})$.

Rem : Un logiciel statistique renvoie une p-valeur. On peut alors conclure au test selon le niveau α que l'on souhaite.



Ex d'un test pour la moyenne à variance connue

- Soit $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$. Soit $\mu_0 \in \mathbb{R}$ connu
- On suppose que μ est inconnu et σ^2 connue
- On souhaite tester $\mathcal{H}_0 : \mu = \mu_0$ **contre** $\mathcal{H}_1 : \mu > \mu_0$
- **Statistique de test** :
  - Estimateur pour μ : $\bar{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.
  - Loi de $\bar{X}_n$ sous $\mathcal{H}_0$:

$$\bar{X}_n \underset{\mathcal{H}_0}{\sim} \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right) \text{ donc } T := \sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma} \underset{\mathcal{H}_0}{\sim} \mathcal{N}(0,1)$$

- **Zone de rejet** : Rejet de $\mathcal{H}_0$ si $\bar{X}_n - \mu_0 > x_\alpha$ donc la zone de rejet est de la forme

$$\mathcal{R}_\alpha = \{\bar{X}_n - \mu_0 > x_\alpha\}.$$

Ex d'un test pour la moyenne à variance connue

- **Calibrage du test :** déterminer x_α pour assurer un test de niveau α :

$$\begin{aligned}\mathbb{P}_{\mu_0}(\bar{X}_n - \mu_0 > x_\alpha) &= \mathbb{P}_{\mu_0}\left(\sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma} > \sqrt{n}\,\frac{x_\alpha}{\sigma}\right)\\ &= \mathbb{P}\left(Z > \sqrt{n}\,\frac{x_\alpha}{\sigma}\right) = \alpha \text{ avec } Z \sim \mathcal{N}(0,1)\end{aligned}$$

donc $\mathbb{P}(Z \leq \sqrt{n}\,\frac{x_\alpha}{\sigma}) = 1 - \alpha$.
Soit $z_{1-\alpha}$ le $(1-\alpha)$ quantile de la loi $\mathcal{N}(0,1)$. En identifiant, on a que

$$\sqrt{n}\,\frac{x_\alpha}{\sigma} = z_{1-\alpha} \Longleftrightarrow x_\alpha = \frac{z_{1-\alpha}\,\sigma}{\sqrt{n}}.$$

Finalement, la zone de rejet vaut

$$\mathcal{R}_\alpha = \left\{\bar{X}_n - \mu_0 > \frac{z_{1-\alpha}\sigma}{\sqrt{n}}\right\} = \left\{\underbrace{\sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma}}_{:=T(X)} > z_{1-\alpha}\right\}.$$

Ex d'un test pour la moyenne à variance connue

- **pvaleur** : pval $= \mathbb{P}_{\mu_0}\left(T(\underline{X}) > T^{obs}\right) = \mathbb{P}\left(Z > T^{obs}\right)$ où $T^{obs} = T(\underline{x}) = \sqrt{n}\,\frac{(\bar{X}_n)^{obs} - \mu_0}{\sigma}$.
- **Evaluation de la puissance :** Pour tout $\mu > \mu_0$,

$$\begin{aligned}
\Pi(\mu) &= \mathbb{P}_\mu\left(\sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma} > z_{1-\alpha}\right) \\
&= \mathbb{P}_\mu\left(\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} > \frac{\sqrt{n}}{\sigma}\,(\mu_0 - \mu) + z_{1-\alpha}\right) \\
&= \mathbb{P}\left(Z > \frac{\sqrt{n}}{\sigma}(\mu_0 - \mu) + z_{1-\alpha}\right) \\
&= 1 - F_Z\left(\frac{\sqrt{n}}{\sigma}(\mu_0 - \mu) + z_{1-\alpha}\right)
\end{aligned}$$

La fonction puissance est croissante de μ.

Ex d'un test pour la moyenne à variance connue

- Si on teste $\mathcal{H}_0 : \mu \leq \mu_0$ contre $\mathcal{H}_1 : \mu > \mu_0$:
  il faut reprendre le contrôle du niveau du test

$$\forall \mu \leq \mu_0, \ \mathbb{P}_\mu(\bar{X}_n - \mu_0 > x_\alpha) \leq \alpha$$

$$\iff \sup_{\mu \leq \mu_0} \mathbb{P}_\mu(\bar{X}_n - \mu_0 > x_\alpha) \leq \alpha$$

$$\iff \sup_{\mu \leq \mu_0} \mathbb{P}_\mu\left(\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} > \frac{\sqrt{n}}{\sigma}\,(x_\alpha + \underbrace{\mu_0 - \mu}_{\geq 0})\right) \leq \alpha$$

$$\iff \sup_{\mu \leq \mu_0} \mathbb{P}\left(Z > \frac{\sqrt{n}}{\sigma}(x_\alpha + \underbrace{\mu_0 - \mu}_{\geq 0})\right) \leq \alpha$$

$$\iff \mathbb{P}\left(Z > \frac{\sqrt{n}}{\sigma} x_\alpha\right) \leq \alpha$$

Ex d'un test pour la moyenne à variance connue

- Si on teste $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_1 : \mu \neq \mu_0$:
  Il faut reprendre la forme de la zone de rejet

$$\mathcal{R}_\alpha = \left\{ |\bar{X}_n - \mu_0| > y_\alpha \right\}$$

avec y_α tel que

$$\begin{aligned} \mathbb{P}_{\mu_0}\left(|\bar{X}_n - \mu_0| > y_\alpha\right) = \alpha \quad &\Longleftrightarrow \quad \mathbb{P}\left(|Z| > \frac{\sqrt{n}}{\sigma}\, y_\alpha\right) = \alpha \\ &\Longleftrightarrow \quad \mathbb{P}\left(Z \leq \frac{\sqrt{n}}{\sigma}\, y_\alpha\right) = 1 - \alpha/2 \end{aligned}$$

donc $\frac{\sqrt{n}}{\sigma} y_\alpha = z_{1-\alpha/2}$ le $(1-\alpha/2)$ quantile de la loi $\mathcal{N}(0,1)$.

Ex d'un test pour la moyenne à variance inconnue

- La variance σ^2 **est maintenant supposée inconnue**.
- On désire tester $\mathcal{H}_0 : \mu = \mu_0$ **contre** $\mathcal{H}_1 : \mu \neq \mu_0$.
- **Statistique de test** :
  - Estimateur pour μ : on considère $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ donc

$$\sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma} \underset{\mathcal{H}_0}{\sim} \mathcal{N}(0,1).$$

    On ne peut utiliser cette statistique pour construire le test car elle dépend du paramètre inconnu σ^2 !

  - On estime σ^2 par

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

    Et comme l'échantillon est gaussien, on a par Cochran (lien) que

$$T := \sqrt{n}\,\frac{\bar{X}_n - \mu_0}{S_n} \underset{\mathcal{H}_0}{\sim} \mathcal{T}(n-1).$$

Ex d'un test pour la moyenne à variance inconnue

- **Zone de rejet** : On rejette $\mathcal{H}_0$ quand T s'écarte trop de 0 à gauche comme à droite donc

$$\mathcal{R}_\alpha = \{|T| > t_\alpha\}.$$

- **Test de niveau** α :

$$\mathbb{P}_{\mu_0}(\sqrt{n}\,|\bar{X}_n - \mu_0|/S_n > t_\alpha) = \mathbb{P}(|V| > t_\alpha) = \alpha$$

où $V \sim \mathcal{T}(n-1)$, ce qui est équivalent à

$$\mathbb{P}(V \leq t_\alpha) = 1 - \frac{\alpha}{2}.$$

t_α est donc le $(1-\alpha/2)$ quantile de la loi de Student à $(n-1)$ degrés de liberté.

- **pvaleur** : pval $= \mathbb{P}_{\mu_0}\left(|T| > |T^{obs}|\right) = \mathbb{P}(|V| > |T^{obs}|)$

1. Formalisme mathématique
2. Test des paramètres d'un échantillon gaussien
   - Test sur la moyenne
   - Test sur la variance
3. Test sur la moyenne en non gaussien
4. Comparaison de deux populations gaussiennes
   - Test d'égalité des variances
   - Test d'égalité des moyennes
5. Exercices

Ex. de test sur la variance

- $X_1, \ldots, X_n$ sont supposées i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$.
- On souhaite tester $\mathcal{H}_0 : \sigma \leq \sigma_0$ contre $\mathcal{H}_1 : \sigma > \sigma_0$.
- **Statistique de test** : On estime σ^2 par

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1).$$

- **Zone de rejet :** $\mathcal{R}_\alpha = \{S^2 > v\}$ avec v tel que

$$\begin{aligned}
\forall\, \sigma \leq \sigma_0,\ \mathbb{P}_\sigma(S^2 > v) \leq \alpha \quad & \Leftrightarrow \quad \sup_{\sigma \leq \sigma_0} \mathbb{P}_\sigma \left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)v}{\sigma^2} \right) \leq \alpha \\
& \Leftrightarrow \quad \sup_{\sigma \leq \sigma_0} \mathbb{P} \left(\chi^2_{(n-1)} \geq \frac{(n-1)v}{\sigma^2} \right) \leq \alpha \\
& \Leftrightarrow \quad \mathbb{P} \left(\chi^2_{(n-1)} \geq \frac{(n-1)v}{\sigma_0^2} \right) \leq \alpha
\end{aligned}$$

On pose donc $v = \frac{\sigma_0^2}{n-1} u_{1-\alpha}$ où $u_{1-\alpha}$ est le $(1-\alpha)$ quantile de la loi du $\chi^2(n-1)$.

Ex. de test sur la variance

- **Zone de rejet :** $\mathcal{R}_\alpha = \{S^2 > \frac{\sigma_0^2}{n-1} u_{1-\alpha}\} = \{\frac{n-1}{\sigma_0^2} S^2 > u_{1-\alpha}\}$
- **pvaleur** pval $= \mathbb{P}\left(V > \frac{n-1}{\sigma_0^2} (S^2)^{obs}\right)$ où $V \sim \chi^2_{(n-1)}$.
- **Evaluation de la puissance :** $\forall \sigma > \sigma_0$,

$$
\begin{aligned}
\Pi(\sigma) &= \mathbb{P}_\sigma\left(S^2 > \frac{\sigma_0^2}{n-1} u_{1-\alpha}\right) \\
&= \mathbb{P}_\sigma\left(\frac{(n-1)S^2}{\sigma^2} > u_{1-\alpha} \frac{\sigma_0^2}{\sigma^2}\right). \\
&= \mathbb{P}\left(V > u_{1-\alpha} \frac{\sigma_0^2}{\sigma^2}\right)
\end{aligned}
$$

La fonction puissance est croissante de σ.

Test sur la moyenne en non gaussien

- $X_1, \ldots, X_n$ i.i.d d'espérance $m = \mathbb{E}[X_1] < +\infty$ et $\mathrm{Var}(X_1) < +\infty$.
- On veut tester $\mathcal{H}_0 : m = m_0$ contre $\mathcal{H}_1 : m \neq m_0$
- Si on n'a pas d'information sur la loi de $\bar{X}_n$ sous $\mathcal{H}_0$, par le TLC (lien)

$$\text{Sous } \mathcal{H}_0, \sqrt{n}\,\frac{\bar{X}_n - m_0}{\sqrt{\mathrm{Var}(X_1)}} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

- Si S^2 est un estimateur consistant de $\mathrm{Var}(X_1)$, par Slutsky (lien)

$$\text{Sous } \mathcal{H}_0, T := \sqrt{n}\,\frac{\bar{X}_n - m_0}{\sqrt{S^2}} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

- Zone de rejet : $\mathcal{R}_\alpha = \left\{|T| > z_{1-\alpha/2}\right\}$ avec

$$\mathbb{P}_{m_0}\left(|T| > z_{1-\alpha/2}\right) \underset{n\to+\infty}{\to} \mathbb{P}\left(|Z| > z_{1-\alpha/2}\right) = \alpha$$

On construit ainsi un test **asymptotique** de niveau α.

Exemple pour une proportion

- Soit $X_1, \ldots, X_n$ de loi $\mathcal{B}(p)$ avec p inconnu
- On veut tester $\mathcal{H}_0 : p = 0.4 \text{ contre } \mathcal{H}_1 : p > 0.4$
- Avec TLC + Slutsky sous $\mathcal{H}_0$

$$T := \sqrt{n}\left(\frac{\bar{X}_n - 0.4}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\right) \underset{n \to +\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,1).$$

- Zone de rejet : $\mathcal{R}_\alpha = \{T > z\}$ avec

$$\mathbb{P}_{p=0.4}(T > z) \underset{n \to +\infty}{\to} \mathbb{P}(Z > z) = \alpha$$

donc z est le $(1-\alpha)$-quantile de la loi $\mathcal{N}(0,1)$.

Contexte

On dispose de deux séries de mesures, notées respectivement $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_p$.

Nous utiliserons la modélisation suivante :

- les variables $X_1, \ldots, X_n$ sont supposées i.i.d. de loi $\mathcal{N}(m_1, \sigma_1^2)$
- les variables $Y_1, \ldots, Y_p$ sont supposées i.i.d. de loi $\mathcal{N}(m_2, \sigma_2^2)$.
- Les deux échantillons sont supposés indépendants.

Les questions suivantes intéressent l'expérimentateur :

1. Les deux échantillons ont-ils la même dispersion :
   a-t-on $\sigma_1 = \sigma_2$ ou $\sigma_1 \neq \sigma_2$?
2. Y a-t-il une différence systématique entre les deux échantillons :
   a-t-on $m_1 = m_2$ ou $m_1 \neq m_2$?

Test d'égalité des variances

- On veut tester $\mathcal{H}_0 : \sigma_1 = \sigma_2$ contre $\mathcal{H}_1 : \sigma_1 \neq \sigma_2$.
- Estimateurs :
  - σ_1^2 est estimé par $S_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \frac{\sigma_1^2}{n-1}\chi^2(n-1)$
  - σ_2^2 est estimé par $S_2^2 = \frac{1}{p-1}\sum_{j=1}^p (Y_j - \bar{Y}_p)^2 \sim \frac{\sigma_2^2}{p-1}\chi^2(p-1)$.
  - S_1^2 et S_2^2 sont indépendantes.
- Loi sous $\mathcal{H}_0$ $(\sigma_1 = \sigma_2)$:

$$\frac{(n-1)S_1^2/\sigma_1^2(n-1)}{(p-1)S_2^2/\sigma_2^2(p-1)} = \frac{S_1^2}{S_2^2} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(n-1, p-1).$$

- Zone de rejet :

$$\mathcal{R}_\alpha = \left\{ \frac{S_1^2}{S_2^2} < a_\alpha \text{ ou } \frac{S_1^2}{S_2^2} > b_\alpha \right\}$$

avec a_α et b_α les $\alpha/2$ et $1-\alpha/2$ quantiles d'une loi $\mathcal{F}(n-1, p-1)$, pour assurer un test de niveau α.

Test d'égalité des moyennes

- On veut tester $\mathcal{H}_0 : m_1 = m_2$ contre $\mathcal{H}_1 : m_1 \neq m_2$
- Estimateurs :
  - m_1 est estimé par $\bar{X} \sim \mathcal{N}(m_1, \frac{\sigma_1^2}{n})$
  - m_2 est estimé par $\bar{Y} \sim \mathcal{N}(m_2, \frac{\sigma_2^2}{p})$
  - Pour comparer m_1 et m_2, on considère donc

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(m_1 - m_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{p}\right)$$

- Loi sous $\mathcal{H}_0$: comme $m_1 = m_2$, on a que

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{p}}} \sim \mathcal{N}(0, 1) \tag{1}$$

mais attention au statut des variances σ_1^2 et σ_2^2 !

Test d'égalité des moyennes

On suppose ici que $\sigma_1 = \sigma_2 = \sigma$ inconnue

- On estime σ^2 par

$$S^2 = \frac{1}{n+p-2}\left\{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{p}(Y_j - \bar{Y}_p)^2\right\} \sim \frac{\sigma^2}{n+p-2}\chi^2(n+p-2)$$

- Sous $\mathcal{H}_0$, on obtient donc, avec les bons arguments d'indépendance, que

$$T := \frac{(\bar{X} - \bar{Y})}{S\sqrt{1/n + 1/p}} \sim \mathcal{T}(n+p-2).$$

- Zone de rejet :

$$\mathcal{R}_\alpha = \left\{|T| > t_{n+p-2,1-\alpha/2}\right\}$$

où $t_{n+p-2,1-\alpha/2}$ est le $1-\alpha/2$ quantile de la loi $\mathcal{T}(n+p-2)$.

Test d'égalité des moyennes

On suppose ici que $\sigma_1 \neq \sigma_2$ inconnues et $n = p$

- LGN : $S_1^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \sigma_1^2$ et $S_2^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \sigma_2^2$ donc $S_1^2 + S_2^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \sigma_1^2 + \sigma_2^2$
- On a également que (Eq (1))

$$\sqrt{n}\,\frac{(\bar{X}-\bar{Y})}{\sqrt{\sigma_1^2+\sigma_2^2}} \sim \mathcal{N}(0,1).$$

- Par Slutsky, il en résulte que

$$\tilde{T} := \sqrt{n}\,\frac{(\bar{X}-\bar{Y})}{\sqrt{S_1^2+S_2^2}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

- Zone de rejet : $\mathcal{R}_\alpha = \{|\tilde{T}| > c_\alpha\}$ avec c_α tel que

$$\mathbb{P}_{\mathcal{H}_0}(\text{ rejeter } \mathcal{H}_0) = \mathbb{P}_{\mathcal{H}_0}(|\tilde{T}| > c_\alpha) \underset{n\to\infty}{\to} \mathbb{P}(|Z| > c_\alpha) = \alpha \text{ où } Z \sim \mathcal{N}(0,1)$$

Donc c_α est le $(1-\alpha/2)$ quantile de la loi $\mathcal{N}(0,1)$.
On obtient un test asymptotique de niveau α.

Exercices

Exercice 1

Un fournisseur commercialise des canettes de soda de 33cl. Ce fournisseur livre une commande à un de ses clients. Lors de la livraison, le fournisseur et le client décident de contrôler la qualité des produits. Ils prélèvent pour cela un même échantillon $X_1, \dots, X_n$ de taille n et mesurent la quantité de soda en cl dans les canettes. On suppose que les X_i suivent une loi $\mathcal{N}(\mu, \sigma^2)$, où l'écart-type σ est supposé connu égal à 0.7.

1. Le fournisseur veut savoir si la moyenne μ est bien égale à 33cl. Il souhaite contrôler le risque de se voir rejeter un lot conforme. Ecrivez les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ associées à cette problématique et construisez un test statistique.
2. Le client a lui besoin d'une quantité minimale de 32cl pour commercialiser les canettes de soda. De son point de vue, il souhaite pouvoir contrôler la probabilité d'accepter un lot non conforme. Par contre, il a des doutes sur la valeur annoncée de σ^2, il suppose donc cette quantité inconnue. Construisez un test statistique adéquate pour le client.
3. Sur un échantillon de taille 10, on trouve $\bar{x}_{10} = 32.5$ et $s_{10}^2 = 0.64$. Quelles sont les conclusions du client et du fournisseur ?

Exercice 2

On s'intéresse à l'influence sur la consommation électrique de l'utilisation d'un adoucisseur d'eau pour la machine à laver. On dispose de $n = 25$ mesures de consommation d'électricité avec adoucisseur et $p = 17$ mesures sans adoucisseur. On suppose que les consommations avec adoucisseur $(X_1, \ldots, X_n)$ et sans adoucisseur $(Y_1, \ldots, Y_p)$ sont gaussiennes de même variance σ^2 inconnue et indépendantes.
Testez l'influence de l'utilisation d'un adoucisseur d'eau sur la consommation électrique d'une machine à laver.

```
Two Sample t-test
data:  X and Y
t = -2.8141, df = 40, p-value = 0.00755
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.16086264 -0.02638442
sample estimates:
mean of x mean of y
0.8052000 0.8988235
```

Exercice 3

On décide de tester l'hypothèse qu'une pièce de monnaie est parfaite, en adoptant la règle de décision suivante : on accepte l'hypothèse de perfection si et seulement si le nombre de faces obtenues dans un échantillon de 100 jets est compris entre 40 et 60.

1. Quelle est la probabilité de rejeter l'hypothèse de perfection alors qu'elle est vraie ?

2. Quelle est la probabilité de ne pas rejeter l'hypothèse de perfection alors que la probabilité p d'avoir un face est de 0.7 ?