Cours de Statistiques Inférentielles

P. Ribereau

I Echantillonage

1 Modèle statistique

On suppose que les variables aléatoires $X_1, \ldots, X_n$ sont *indépendantes* et *identiquement distribuées* (i.i.d.). tq $X_i \sim \mathbb{P}_\theta$

On note le mod 'ele statistique

$$(\mathbb{R}, \mathcal{B}_\mathbb{R}, (P_\theta,\ \theta \in \Theta \subset \mathbb{R}^p))^n .$$

En dehors de ce cadre, on parle de modèle non paramétrique ou semi-paramétrique si $\Theta \subset \mathbb{R}^k \times U$ où U est non paramétrique.

a. Modèle dominé, vraisemblance

Considérons un modèle statistique $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, Q \in \mathbb{Q})$. Le modèle est *dominé* par une mesure σ-finie μ sur $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ si pour tout $Q \in \mathbb{Q}$, Q est absolument continue par rapport à μ. Alors il existe une fonction de densité f_Q telle que $Q = f_Q \mu$ (i.e $Q(A) = \int_A f_Q(x) d\mu(x)$ pour tout $A \in \mathcal{C}$). La fonction f_Q est la *vraisemblance* (Likelihood en anglais) du modèle notée $L_Q(x) = f_q(x)$. On note $\ell_Q(x) = \log L_Q(x)$ la log vraisemblance.

Dans le cas d'un modèle d'échantillonage réel, la vraisemblance vaut $L_\theta(x_1, \ldots, x_n) = f_\theta(x_1) \times \cdots \times f_\theta(x_n)$.

2 Définition d'une statistique

— On appelle *statistique* toute variable aléatoire définie sur $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ à valeurs dans $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$.

— On utilisera la notation T_n pour une statistique. Elle ne dépend pas de la famille de lois de probabilité $\mathbb{Q}$. En particulier, dans un modèle paramétrique, une statistique T_n ne dépend pas de θ. C'est la loi de probabilité de T_n qui dépend de θ.

Remarque.
On distingue :
— la statistique T_n qui est une variable aléatoire sur $(\mathbb{R}^n, \mathcal{B}, \mathbb{Q})$. Dans l'exemple précédent, c'est l'application moyenne empirique (on est ici au niveau conceptuel, mathématique).
— la variable aléatoire $T_n(X_1, \ldots, X_n)$ qui est une variable aléatoire sur $(\Omega, \mathcal{B})$. Dans l'exemple précédent, c'est la variable aléatoire $\overline{X}_n$ (on est ici au niveau de la statistique inférentielle).
— la valeur observée de cette variable aléatoire : $T_n(x_1, \ldots, x_n) \in \mathbb{R}^k$. Dans l'exemple précédent, c'est le nombre réel $\bar{x}_n$, moyenne empirique de la série statistique $x_1, \ldots, x_n$ (on est au niveau de la statistique descriptive).

3 Quelques notions de base sur les estimateurs

On considère dans cette section un modèle paramétrique réel d'échantillonage indépendant :

$$(\mathbb{R}, \mathcal{B}, (P_\theta, \theta \in \Theta \subset \mathbb{R}^p))^n .$$

a. Définition d'un estimateur

Soit g une fonction définite sur Θ, à valeur dans $\mathbb{R}^p$.
— On appelle *estimateur d'un paramètre* $g(\theta)$ toute statistique à valeur dans $(\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k})$, il sera généralement noté T_n.
— On appelle *estimation* la valeur observée de T_n sur l'échantillon, elle sera notée $T_n(x_1, \ldots, x_n)$.

b. Notion de biais

— On appelle *biais* de l'estimateur T_n pour le paramètre $g(\theta)$ la quantité

$$B_\theta(T_n) = \mathbb{E}_\theta[T_n] - g(\theta).$$

— On appelle *estimateur sans biais de* $g(\theta)$ un estimateur T_n tel que $B(T_n) = 0$.
— Si $B(T_n) \to 0$ quand $n \to +\infty$, on dit que T_n est *asymptotiquement sans biais* pour $g(\theta)$.

c. Convergence d'un estimateur

On dit que T_n est *convergent* pour $g(\theta)$ s'il converge en probabilité vers $g(\theta)$:

$$\forall \varepsilon > 0,\ P(|T_n - g(\theta)| < \varepsilon) \longrightarrow 1 \text{ quand } n \to +\infty.$$

Critères de convergence d'un estimateur
(i) Si T_n est un estimateur (asympt.) sans biais de $g(\theta)$ et si $\mathbb{V}(T_n) \longrightarrow 0$ quand $n \to +\infty$, alors
T_n est un estimateur convergent pour $g(\theta)$.

d. Comparaisons des estimateurs

On utilise le *risque quadratique* pour comparer deux estimateurs du même paramètre $g(\theta)$. Pour l'estimateur T_n de $g(\theta)$, il est défini par :

$$R(T_n, g(\theta)) = \mathbb{E}[(T_n - g(\theta))^2].$$

Propriété. $R(T_n, g(\theta)) = (B(T_n))^2 + \mathbb{V}(T_n)$.

e. Moyenne aléatoire, variance aléatoire

Soit un modèle paramétrique réel d'échantillonage

$$(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (P_\theta,\ \theta \in \Theta \subset \mathbb{R}^p))^n$$

tel que la loi de probabilité P_θ admette pour espérance $\mu < \infty$ et pour variance $0 < \sigma^2 < \infty$.

Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire associé de ce modèle.

On appelle *moyenne aléatoire* la statistique

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

On appelle *variance aléatoire* la statistique

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

on verra ci-dessous que cet estimateur de la variance est biaisé, on lui préférera la *variance estimée* :

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

Proposition I.1 *1. $\overline{X}_n$ est un estimateur sans biais et convergent pour μ.*

2. V_n est un estimateur biaisé mais asymptotiquement sans biais de σ^2.

3. $S_n^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ est un estimateur sans biais de σ^2.

4. V_n et S_n^2 sont des estimateurs convergent de σ^2.

II Estimation

— On reste dans le cadre du modèle :

$$(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (P_\theta,\ \theta \in \Theta \subset \mathbb{R}))^n.$$

— Soit $f(x, \theta)$ la densité de la loi des X_i.

— Soit L la vraisemblance. On a :
$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$.
— Soit $\mathcal{X} \subset \mathbb{R}$ le support de $f(x, \theta)$ où

$$\mathcal{X} = \{x \in \mathbb{R} \mid f(x, \theta) > 0\}.$$

1 Hypothèses fondamentales sur la densité $f(x, \theta)$

(H1) $\mathcal{X}$ est indépendant de θ.
(H2) Θ est un ouvert.

(H3) $\frac{\partial}{\partial\theta} f(x, \theta)$ et $\frac{\partial^2}{\partial\theta^2} f(x, \theta)$ sont définies $\forall (x, \theta) \in \mathcal{X} \times \Theta$.
(H4) On suppose que les dérivés et dérivés secondes de f par rapport à θ sont dominées par des fonctions μ-intégrables sur tout compact inclu dans Θ : pour tout compact $K \subset \Theta$, il existe deux fonctions positive μ-intégrables $\phi(x)$ et $\psi(x)$ telles que pour tout $\theta \in K$ et presque tout $x \in \mathcal{X}$,

$$\left|\frac{\partial}{\partial\theta} f(x, \theta)\right| \leq \phi(x) \text{ et } \left|\frac{\partial^2}{\partial\theta^2} f(x, \theta)\right| \leq \psi(x).$$

2 Information

a. Information de Fisher

Définition II.1 *On appelle fonction score ou score la fonction S définie par :*

$$\begin{array}{rccl} S & : & \mathcal{X} \times \Theta & \longrightarrow & \mathbb{R} \\ & & (x, \theta) & \longmapsto & S(x, \theta) = \frac{\partial}{\partial\theta} \ln(f(x, \theta)) \end{array}$$

Remarques.
— Le score n'est défini que si **(H1)**, **(H2)** et **(H3)** sont vraies.
— On peut définir de même le score à partir de la vraisemblance.

$$S_n(x, \theta) = \frac{\partial}{\partial\theta} \ln(L(x, \theta)) \text{ avec ici } x \in \mathbb{R}^n.$$

Dans le modèle d'échantillonage, on a : $S_n(x, \theta) = \sum_{i=1}^{n} S(x_i, \theta)$.
Propriétés.

Supposons **(H4)** vraie, on a alors :

$$\mathbb{E}[S(X, \theta)] = 0 \quad \text{et} \quad \mathbb{E}[S_n(X, \theta)] = 0.$$

Information au sens de Fisher

Définition II.2 *On appelle information de Fisher la fonction I définie par :*

$$\begin{array}{rccl} I & : & \Theta & \longrightarrow & \mathbb{R}^+ \\ & & \theta & \longmapsto & I(\theta) = \mathbb{E}\left[(S(X, \theta))^2\right] \end{array}$$

Remarques. On peut aussi poser (en terme de vraisemblance) :

$$I_n(\theta) = \mathbb{E}\left[(S_n(X,\theta))^2\right].$$

b. Propriétés.

Si **(H4)** est vraie, alors

$$I(\theta) = \mathbb{V}(S(X,\theta)) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln(f(X,\theta))\right] \quad \text{et} \quad I_n(\theta) = \mathbb{V}(S_n(X,\theta)) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln(L(X,\theta))\right].$$

Dans le modèle d'échantillonage, on a donc :
$I_n(\theta) = nI(\theta)$.

3 Inégalité de Cramer-Rao

a. Hypothèses supplémentaires

(H5) $0 < I_n(\theta) < +\infty$ ou $0 < I(\theta) < +\infty$.
Soit T_n un estimateur. Posons $\mathbb{E}_\theta[T_n] = g(\theta)$.

Théorème II.1 (Inégalité de Cramer-Rao.) *Si les hypothèses* **(H1)**, **(H2)**, **(H3)**, **(H4)** *et* **(H5)** *sont vérifiées, si de plus la fonction g est dérivable, alors on a :*

$$\mathbb{V}(T_n) \geq \frac{[g'(\theta)]^2}{I_n(\theta)}.$$

La partie de droite est appelée *borne inférieure de l'inégalité de Cramer-Rao.* Nous la noterons $K_{T_n}(\theta)$.

Définition II.3
- *On dit que l'estimateur T_n est efficace s'il vérifie $\mathbb{V}(T_n) = K_{T_n}(\theta)$.*
- *Si T_n n'est pas efficace mais que $\dfrac{K_{T_n}(\theta)}{\mathbb{V}(T_n)} \to 1$ quand $n \to +\infty$, on dit que l'estimateur T_n est asymptotiquement efficace.*

b. Relation entre estimateurs efficaces

Propriétés.
(P1) Si T_n est un estimateur efficace de θ, alors $kT_n + b$ est aussi un estimateur efficace de θ, $\forall k \in \mathbb{R}^*, \forall b \in \mathbb{R}$.
(P2) Soient T_{1n} et T_{2n} deux estimateurs sans biais du paramètre θ. S'ils sont tous les deux efficaces, alors $T_{1n} = T_{2n}$ presque sûrement.

c. Dégradation de l'information

Si $h(t,\theta)$ est la vraissemblance de T, on note $I_T(\theta)$ l'information relative à T :

$$I_T(\theta) = \mathbb{E}(S(T,\theta)^2) \text{ où } S(T,\theta) = \frac{\partial \ln h(t,\theta)}{\partial \theta}.$$

$I_T(\theta)$ vérifie les mêmes propriétés (centrée, relation avec $\frac{\partial^2}{\partial \theta^2} h$, inégalité de Cramer-Rao) que $I(\theta)$.

Proposition II.1 *Soit T une statistique, on a $I_T(\theta) \leq I_n(\theta)$ avec égalité si et seulement si la statistique T est exhaustive pour le paramètre θ.*

4 Notion d'exhaustivité

Définition II.4 (Principe de factorisation) *La statistique T_n est ici une variable aléatoire définie sur $(\mathcal{X}, \mathcal{B}_\mathcal{X}, (P_\theta,\ \theta \in \Theta \subset \mathbb{R}))^n$ à valeurs dans $(\Theta, \mathcal{B}_\Theta)$.*

On considère :

- *la vraisemblance (fonction de densité ou probabilité) de T_n que l'on va noter $h(t,\theta)$,*
- *la densité conjointe de l'échantillon $(X_1, \ldots, X_n)$ notée $L(x_1, \ldots, x_n, \theta)$,*

On dira que la statistique T_n est exhaustive pour θ *s'il existe une fonction k sur $\mathcal{X}^n$ telle que*

$$L(x_1, \ldots, x_n, \theta) = h(t,\theta) k(x_1, \ldots, x_n).$$

Théorème II.2 (de factorisation) *Pour qu'une statistique T soit exhaustive, il suffit que la vraissemblance s'écrive :*

$$L(x_1, \ldots, x_n, \theta) = \phi(t,\theta)\psi(x_1, \ldots, x_n).$$

5 Exhaustivité et estimateurs efficaces ; la famille exponentielle

a. Le modèle exponentiel

Définition II.5 *On appelle famille exponentielle à paramètre unidimensionnel θ toute loi de probabilité (discrète ou continue) dont la vraissemblance peut se mettre sous la forme :*

$$f(x,\theta) = \begin{cases} \exp[\alpha(\theta)\beta(x) + \gamma(\theta) + \delta(x)] & si \quad x \in \mathcal{X} \\ 0 & si \quad x \notin \mathcal{X} \end{cases},$$

avec, α et γ des fonctions deux fois différentiables.

Proposition II.2 *Dans la famille exponentielle, toute statistique de la forme $T_n = k \sum_{i=1}^n \beta(X_i)$ est exhaustive pour θ.*

Théorème II.3 (Théorème de Darmois) *Lorsque $\mathcal{X}$ est indépendant de θ, le modèle admet une statistique exhaustive ssi le modèle est exponentiel.*

b. Théorème sur l'efficacité

Théorème II.4 *Si $\mathcal{X}$ ne dépend pas de θ et que le modèle admette un estimateur efficace alors le modèle est exponentiel car l'estimateur est nécessairement exhaustif.*

Dans un modèle exponentiel, alors à une transformation linéaire près, il existe un unique estimateur efficace qui vérifie :

$$T = \frac{1}{n}\sum_{i=1}^{n}\beta(X_i),$$

$$g(\theta) = -\frac{\gamma'(\theta)}{\alpha'(\theta)},$$

$g(\theta) = \mathbb{E}_\theta(T)$, *et* $\mathbb{V}_\theta(T) = \frac{g'(\theta)}{n\alpha'(\theta)}$.

6 Quelques méthodes usuelles d'estimation

a. Méthode empirique

Si le paramètre θ considéré représente une quantité particulière pour le modèle (par exemple, l'espérance ou la variance), on peut naturellement choisir comme estimateur la quantité empirique correspondante pour l'échantillon $X_1, \ldots, X_n$.

b. Méthode des moindres carrés

Définition II.6 *On appelle estimateur des moindres carrés de θ la statistique*

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\sum_{i=1}^{n}(X_i - h(\theta))^2.$$

c. Méthode des moments

On note

— Soit X une variable aléatoire réelle.

On appelle *moment (théorique) d'ordre r* : $M_r = \mathbb{E}[X^r]$.

On appelle *moment (théorique) centré d'ordre r* : $\overline{M}_r = \mathbb{E}[(X - \mathbb{E}[X])^r]$.

— Soit $(x_1, \ldots, x_n)$ les valeurs observées d'un échantillon de taille n.

On appelle *moment empirique d'ordre r* : $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i)^r$.

On appelle *moment empirique centré d'ordre r* : $\overline{m}_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^r$.

Principe. Supposons le paramètre θ de dimension p. La méthode consiste à poser un système d'équations en égalant moments théoriques (centrés ou non) et moments empiriques :

$$\left\{\begin{array}{rcl} M_1(\theta) & = & m_1 \\ & \vdots & \\ M_p(\theta) & = & m_p \end{array}\right.$$

d. Méthode du maximum de vraisemblance : principe

Définition II.7 *On appelle estimateur du maximum de vraisemblance (EMV) du paramètre θ la statistique $\hat{\theta}_n$ rendant maximale, selon θ, la fonction de vraisemblance du modèle $L(X_1, \ldots, X_n, \theta)$, soit :*

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} L(X, \theta).$$

Propriété.
Si le modèle vérifie les propriétés **(H1)**, **(H2)** et **(H3)**,
alors pour que $\hat{\theta}_n$ soit un EMV de θ il est nécessaire que

- $\left.\dfrac{\partial \ln L(X,\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_n} = 0$ soit $S_n(X, \hat{\theta}_n) = 0$ (équation de vraisemblance),
- $\left.\dfrac{\partial^2 \ln L(X,\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}_n} < 0.$

Propriété (lien avec l'exhaustivité).
S'il existe une statistique exhaustive T_n pour θ, alors l'EMV de θ ne dépend que de T_n.

7 Exercices

Exercice 1. On considère la loi normale $\mathcal{N}(\mu, \sigma^2)$.

a) On suppose σ^2 connue et l'on considère le modèle paramétrique réel d'échantillonnage suivant

$$\left(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (\mathcal{N}(\mu, \sigma^2),\ \mu \in \Theta = \mathbb{R})\right)^n.$$

L'estimateur $\overline{X}_n$ est-il un estimateur efficace de μ ?

b) Sans supposer μ connue, on pose $\theta = \sigma^2$ et l'on considère le modèle paramétrique réel d'échantillonnage suivant

$$\left(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (\mathcal{N}(\mu, \sigma^2),\ \sigma^2 \in \Theta = \mathbb{R}_+^*)\right)^n.$$

L'estimateur S_n^2 est-il un estimateur efficace de σ^2 ?

Exercice 2.

a) Montrer que la loi de Poisson appartient à la famille exponentielle.

b) Montrer que la loi de Cauchy n'appartient pas à la famille exponentielle.

Exercice 3. Déterminer l'EMV du paramètre λ d'une loi de Poisson. En étudier les propriétés (biais, convergence, efficacité, exhaustivité).

Exercice 4. Ecrire la vraisemblance d'une loi de Bernoulli de paramètre $p \in]0,1[$.

Déterminer l'EMV de p. Etudier ses propriétés (biais, convergence, efficacité, exhaustivité).

Exercice 5.

a) Déterminer l'EMV $\hat{\theta}_n$ du paramètre θ de la loi uniforme sur $[0,\theta]$ avec $\theta \in \mathbb{R}_+^*$.
b) Déterminer la densité de probabilité de $\hat{\theta}_n$.
c) Calculer $\mathbb{E}[\hat{\theta}_n]$ et $\mathbb{V}(\hat{\theta}_n)$.
d) Etudier les propriétés de $\hat{\theta}_n$ (biais, convergence, efficacité).
e) Proposer un estimateur T_n de θ sans biais et convergent.
f) Choisir entre $\hat{\theta}_n$ et T_n au moyen du risque quadratique.
g) Montrer que l'estimateur de θ obtenu par la méthode des moindres carrés est identique à l'estimateur des moments. On notera U_n cet estimateur.
h) Etudier les propriétés de U_n (biais et convergence) et le comparer à T_n.
i) Commenter.

8 Généralisation au cas d'un paramètre multidimensionnel

On considère dans ce chapitre un modèle paramétrique réel d'échantillonnage :

$$(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (P_\theta,\ \theta \in \Theta \subset \mathbb{R}^p))^n, \quad \text{avec } p \geq 2.$$

a. Généralisation des définitions sur les estimateurs

Estimateurs

Un estimateur est une statistique T_n définie sur $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (P_\theta,\ \theta \in \Theta \subset \mathbb{R}^p))^n$ à valeurs dans $(\Theta, \mathcal{B}_\Theta)$. C'est donc un vecteur aléatoire de dimension $p : T_n = (T_{n,1}, \ldots, T_{n,p})$.

Estimateur sans biais.

— **Biais de** T_n : $B(T_n) = \mathbb{E}(T_n) - \theta \ \in \mathbb{R}^p$.
— On dit que T_n est *sans biais pour* θ si $B(T_n) = 0_p$,

— On dit que T_n est *asymptotiquement sans biais pour* θ si $B(T_n) \longrightarrow 0_p$ pour $n \to +\infty$,
autrement dit si $\forall j = 1, \ldots, p,\ \mathbb{E}[T_{n,j}] \longrightarrow \theta_j$ pour $n \to +\infty$.

Estimateur convergent.

— T_n est *convergent pour* θ si et seulement si $T_n \longrightarrow_{proba} \theta$ pour $n \to +\infty$,

c'est à dire : $\forall j = 1, \ldots, p,\ T_{n,j} \longrightarrow_{proba} \theta_j$.
— **Conditions nécessaires et suffisantes de convergence :**

$$||T_n - \theta|| \longrightarrow_{proba} 0 \text{ pour } n \to +\infty \quad \Longleftrightarrow \quad T_n \text{ est convergent pour } \theta$$

où $||.||$ désigne toute norme de $\mathbb{R}^p$.

Risque quadratique.

— Il est défini par : $R(T_n, \theta) = \mathbb{E}\left[(T_n - \theta)'(T_n - \theta)\right] = \sum_{j=1}^{p} \mathbb{E}\left[(T_{n,j} - \theta_j)^2\right]$.

— **Propriété :**

$$R(T_n, \theta) = B(T_n)'B(T_n) + \sum_{j=1}^{p} \mathbb{V}(T_{n,j}).$$

On peut réécrire ceci sous la forme :

$$R(T_n, \theta) = \sum_{j=1}^{p} \left(\mathbb{E}[T_{n,j}] - \theta_j\right)^2 + \sum_{j=1}^{p} \mathbb{V}(T_{n,j}).$$

b. Généralisation de l'inégalité de Cramer-Rao

On continue à noter
— $f(x, \theta)$ la vraisemblance du modèle de dimension 1, ici $x \in \mathbb{R}$ et $\theta \in \mathbb{R}^p$;
— $L(x, \theta)$ la vraisemblance du modèle de dimension n, ici $x \in \mathbb{R}^n$ et $\theta \in \mathbb{R}^p$;
— $\mathcal{X}$ le support de f et $\mathcal{X}^n$ celui de L.

Généralisation des hypothèses de régularité

Les hypothèses **(H1)** et **(H2)** ne sont pas modifiées mais seront ici notée **(H1′)** et **(H2′)**.

(H1′) $\mathcal{X}$ est indépendant de θ.

(H2′) Θ est un ouvert.

On a $f(x, \theta) > 0, \forall (x, \theta) \in \mathcal{X} \times \Theta$.

(H3′) $\forall j = 1, \ldots, p$, $\dfrac{\partial}{\partial \theta_j} f(x, \theta)$ est définie $\forall (x, \theta) \in \mathcal{X} \times \Theta$.

$\forall (j, k) \in \{1, \ldots, p\}$, $\dfrac{\partial^2}{\partial \theta_j \partial \theta_k} f(x, \theta)$ est définie $\forall (x, \theta) \in \mathcal{X} \times \Theta$.

(H4′) $\forall (j, k) \in \{1, \ldots, p\}$

$$\frac{\partial}{\partial \theta_j} f(x, \theta) \text{ et } \frac{\partial^2}{\partial \theta_j \partial \theta_k} f(x, \theta)$$

vérifient la propriété de domination sur tout compact de Θ (par des fonctions de x μ-intégrables).

Fonction de score (ou score)

On suppose les hypothèses **(H1′)**, **(H2′)** et **(H3′)** vérifiées.

Définition. La fonction *score* est définie par :

$$\begin{array}{rccl} S : & \mathcal{X} \times \Theta & \longrightarrow & \mathbb{R}^p \\ & (x, \theta) & \longmapsto & S(x, \theta) = \operatorname{grad}_\theta \ln(f(x, \theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \ln(f(x, \theta)) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln(f(x, \theta)) \end{pmatrix} \end{array}$$

Remarques et propriétés.
— On peut aussi définir le score du modèle de dimension n : $S_n(x, \theta) = \operatorname{grad}_\theta \ln(L(x, \theta))$.
— Sous **(H4′)**, on peut montrer que : $\mathbb{E}[S(X, \theta)] = 0_p = \mathbb{E}[S_n(x, \theta)]$.

Matrice d'information de Fisher

Définition. La *matrice d'information de Fisher* est une matrice carrée $p \times p$ définie par :

$$I(\theta) = \mathbb{E}\left[S(X,\theta)(S(X,\theta))'\right],$$

l'élément (j,k) de la matrice $I(\theta)$ est donnée par
$\mathbb{E}\left[\frac{\partial}{\partial\theta_j}\ln(f(x,\theta))\frac{\partial}{\partial\theta_k}\ln(f(x,\theta))\right]$.
Remarques et propriétés.

— On peut aussi définir la matrice d'information de Fisher par rapport du modèle de dimension n :

$$I_n(x,\theta) = \mathbb{E}\left[S_n(X,\theta)(S_n(X,\theta))'\right].$$

— Dans un modèle d'échantillonnage, on a :
$I_n(x,\theta) = nI(\theta)$.

Généralisation de l'inégalité de Cramer-Rao

— Soit T_n un estimateur de θ. On pose $\mathbb{E}[T_n] = g(\theta)$.
La fonction g définie sur Θ est à valeurs dans $\mathbb{R}^p$, sa $j$ème coordonnée est $g_j(\theta) = \mathbb{E}[T_{n,j}]$.
— Soit $D_g(\theta)$ la matrice jacobienne de g.
— Notons $V_{T_n}(\theta)$ la matrice de variances-covariances de T_n.

Considérons l'hypothèse supplémentaire suivante :
(H5′) $I_n(\theta)$ est une matrice définie positive.
Inégalité de Cramer-Rao. Sous les hypothèses **(H1′)** à **(H5′)**, la matrice

$$V_{T_n}(\theta) - D_g(\theta)\left[I_n(\theta)\right]^{-1}\left(D_g(\theta)\right)'$$

est semi-définie positive.
Définitions.

— La matrice $D_g(\theta)\left[I_n(\theta)\right]^{-1}\left(D_g(\theta)\right)'$ s'appelle la *borne inférieure de l'inégalité de Cramer-Rao.*
— On dit que T_n est *efficace pour* θ s'il vérifie $V_{T_n}(\theta) = D_g(\theta)\left[I_n(\theta)\right]^{-1}\left(D_g(\theta)\right)'$.

Forme générale de la famille exponentielle

Définition. On dit qu'une loi de probabilité appartient à la famille exponentielle (à paramètre multidimensionnel) si sa vraisemblance peut s'écrire sous la forme :

$$f(x,\theta) = \begin{cases} \exp\left[\sum_{j=1}^{p}\alpha_j(\theta)\beta_j(x) + \gamma(\theta) + \delta(x)\right] & \text{si } x \in \mathcal{X}, \\ 0 & \text{sinon,} \end{cases}$$

avec $\mathcal{X}$ indépendant de θ. Les applications α_j et γ vont de $\mathbb{R}^p$ dans $\mathbb{R}$.
Les applications β_j et δ vont de $\mathbb{R}$ dans $\mathbb{R}$.

c. Généralisation de la méthode du maximum de vraisemblance

Définition. L'estimateur du maximum de vraisemblance (EMV) de θ est définie par :

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} L(X_1, \ldots, X_n, \theta).$$

Caractérisation de l'EMV $\hat{\theta}_n$. Si les hypothèses **(H1′)**, **(H2′)** et **(H3′)** sont vérifiées, alors pour déterminer $\hat{\theta}_n$,

i) on résoud $S_n(X, \hat{\theta}_n) = 0$ (équations de vraisemblance),

ii)) on vérifie que la matrice hessienne de $\ln L$ (matrice carrée d'ordre p de terme général $\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ln(L(x,\theta))$ calculée en $\hat{\theta}_n$ est définie négative,

iii) on vérifie que le maximum local est un maximum.

III Comportement asymptotique des estimateurs

1 Propriétés asymptotiques de l'EMV

a. En dimension 1

Deux hypothèses supplémentaires sont nécessaires :

(H6) $\theta \neq \theta' \implies P_\theta \neq P_{\theta'}$.

(H7) $\dfrac{\partial^2}{\partial\theta^2}\ln f(x,\theta)$ est continue en θ, uniformément en x.

Théorème III.1 *Si les hypothèses* **(H1)**, **(H2)**, **(H3)**, **(H4)** *et* **(H6)** *sont vérifiées, alors il existe une suite $\hat{\theta}_n$ d'estimateurs du maximum de vraissemblance qui converge presque sûrement vers θ.*

Théorème III.2 *Sous les hypothèses* **(H1)** *à* **(H7)**, *on a :*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \longrightarrow_{loi} \mathcal{N}(0, I^{-1}(\theta)) \quad \textit{quand } n \to +\infty.$$

b. En dimension supérieure

Les résultats de convergence pour l'EMV en dimension supérieure restent valables :

Théorème III.3 *Si les hypothèses* **(H1′)** *à* **(H7′)** *sont toutes vérifiées, alors on a :*

$$\sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow_{loi} \mathcal{N}_p(0_n, I^{-1}(\theta)).$$

2 Définitions / outils

a. Normalité et efficacité asymptotique

Soit T_n un estimateur de θ.

— Si $\sqrt{n}(T_n - \theta) \longrightarrow_{loi} \mathcal{N}_p(0_p, \Sigma)$,
alors on dit que T_n est *asymptotiquement normal*. La matrice Σ est appelée matrice de variances-covariances aymptotique de T_n. (Cela n'implique pas que $n\mathbb{V}(T_n) \to \Sigma$.)

— Si $\sqrt{n}(T_n - \theta) \longrightarrow_{loi} \mathcal{N}_p(0_p, I^{-1}(\theta))$,
alors on dit que T_n est *asymptotiquement efficace*.

— L'EMV est asymptotiquement normal et efficace.

b. Méthode Delta

Soit g une fonction C^1. On suppose que T_n est un estimateur de θ tel que

$$a_n(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta))$$

avec $a_n \to \infty$. Alors, $g(T_n)$ converge en probabilité vers $g(\theta)$ et

$$a_n(g(T_n) - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, g'(\theta)^2\sigma^2(\theta)).$$

En dimension supérieure, on considère T_n un vecteur aléatoire de $\mathbb{R}^k$, Σ une matrice de covariance. On suppose que

$$a_n(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

avec $a_n \to \infty$. Alors, pour toute fonction g de classe C^1, $g(T_n)$ converge en probabilité vers $g(\theta)$ et

$$a_n(g(T_n) - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, D_g \Sigma D_g^t)$$

où D_g est la matrice Jacobienne de g calculée en θ.

3 Exercices

Exercice 1. On considère le modèle d'échantillonnage normal avec $P_\theta = \mathcal{N}(\mu, \sigma^2)$.

a) Déterminer l'EMV de $\theta = (\mu, \sigma^2)$.

b) Etudier ses propriétés (biais, convergence, efficacité).

c) Quelle fonction $h(\theta)$ peut-on estimer par un estimateur sans biais et efficace ?

Exercice 2. On considère le modèle d'échantillonnage multinomial à $k \geq 3$ catégories :

$$\left(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{B}(p_1, \ldots, p_k), p_i \in [0,1], \sum_{i=1}^{k} p_i = 1\right)^n$$

avec $X \rightsquigarrow \mathcal{B}(p_1, \ldots, p_k)$, $\mathbb{P}(X = a_i) = p_i$.

a) Déterminer l'EMV de $\theta = (p_1, \ldots, p_{k-1})$.

b) Montrer qu'il est sans biais, convergent et efficace.

IV Estimation par intervalle de confiance

1 Introduction

On va considérer dans ce chapitre un modèle statistique réel paramétrique (avec un paramètre unidimensionnel) :

$$(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (P_\theta, \theta \in \Theta \subset \mathbb{R}))^n.$$

Définition.

Soit $\alpha \in [0,1]$. On appelle *intervalle de confiance du paramètre* θ de niveau (de confiance) $1-\alpha$ la donnée de deux statistiques A_n et B_n vérifiant

$$P(A_n \leq \theta \leq B_n) = 1 - \alpha.$$

2 Intervalles de confiance pour les paramètres de la loi normale

On suppose ici que l'on dispose d'un échantillon $(X_1, \ldots, X_n)$ où les X_i sont indépendants et identiquement distribués selon la loi $\mathcal{N}(\mu, \sigma^2)$.

Intervalle de confiance pour μ lorsque σ^2 est connue

L'intervalle de confiance pour μ de niveau de confiance $1-\alpha$ lorsque σ^2 est connue est :

$$\overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

où $z_{1-\alpha/2}$ est le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite $\mathcal{N}(0,1)$.

Intervalle de confiance pour μ lorsque σ^2 est inconnue

L'intervalle de confiance pour μ de niveau de confiance $1-\alpha$ lorsque σ^2 est inconnue est :

$$\overline{X}_n - t_{1-\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}$$

où $t_{1-\alpha/2}$ est le fractile d'ordre $1-\alpha/2$ de la loi de Student $T(n-1)$ et $S_n = \sqrt{S_n^2}$.

Intervalle de confiance pour σ^2 lorsque μ est connue

On se donne ici $\alpha_1 > 0$ et $\alpha_2 > 0$ vérifiant $\alpha_1 + \alpha_2 = \alpha$.

L'intervalle de confiance pour σ^2 de niveau de confiance $1-\alpha$ lorsque μ est connue est :

$$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\tilde{k}_2} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{\tilde{k}_1}$$

où $\tilde{k}_1$ (resp. $\tilde{k}_2$) est le fractile d'ordre α_1 (resp. $1-\alpha_2$) de la loi du chi-deux $\chi^2(n)$.

Intervalle de confiance pour σ^2 lorsque μ est inconnue

On se donne ici à nouveau $\alpha_1 > 0$ et $\alpha_2 > 0$ vérifiant $\alpha_1 + \alpha_2 = \alpha$.

L'intervalle de confiance pour σ^2 de niveau de confiance $1-\alpha$ lorsque μ est inconnue est :

$$\frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{k_2} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{k_1}$$

où k_1 (resp. k_2) est le fractile d'ordre α_1 (resp. $1-\alpha_2$) de la loi du chi-deux $\chi^2(n-1)$.

3 Construction d'intervalles de confiance asymptotiques

Définition IV.1 *Un intervalle de confiance* $[A_n, B_n]$ *est de niveau asymptotique* $1-\alpha$ *si*

$$\mathbb{P}(A_n \leq \theta \leq B_n) \overset{n\to\infty}{\longrightarrow} 1-\alpha.$$

a. Utilisation de la convergence de l'EMV

Supposons vérifiées les hypothèses de régularité **(H1)** à **(H7)**.

Soit $I_n(\theta)$ l'information de Fisher du modèle considéré. Soit $\widehat{\theta}_n$ l'EMV de θ. On a vu que :

$$\sqrt{I_n(\theta)}\left(\widehat{\theta}_n - \theta\right) \longrightarrow_{Loi} \mathcal{N}(0,1) \quad \text{pour } n \to +\infty.$$

donc

$$P\left(-z_{1-\alpha/2} \leq \sqrt{I_n(\theta)}\left(\widehat{\theta}_n - \theta\right) \leq z_{1-\alpha/2}\right) \simeq 1-\alpha.$$

4 Exercices

Exercice 1. Soient $X_1, \ldots, X_{10}$ dix variables aléatoires i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$. On dispose des observations suivantes :

6 8 1 5 6 7 6 6 5 9

Calculer les intervalles de confiance de niveau 95% suivants :
- pour μ, sachant que $\sigma^2 = 4$;
- pour μ, ne connaissant pas σ^2 ;
- pour σ^2, puis pour σ, ne connaissant pas μ.

Exercice 2.

Dans une fabrication en série, on cherche à estimer le taux de pièces défectueuses. Pour cela, on a réalisé, à quatre périodes différentes, quatre prélèvements. Les résultats sont les suivants :

6 pièces défectueuses sur 30,
10 pièces défectueuses sur 50,
20 pièces défectueuses sur 100,
40 pièces défectueuses sur 200.

Déterminer, dans chaque cas, l'intervalle de confiance de niveau 95% de ce taux.

Exercice 3.

Déterminer l'intervalle de confiance de niveau 95% de la proportion p d'un événement E, lorsque sur 80 expériences (indépendantes), l'événement s'est produit 45 fois.

Exercice 4.

En utilisant le théorème central limite, construire un intervalle de confiance de niveau asymptotiquement égal à $1-\alpha$ pour le paramètre λ d'une loi de Poisson.

Application numérique : On compte le nombre de parasites par fruit dans un lot de fruits parasités et on obtient :

x_i : nombre de parasites par fruit	0	1	2	3	4	5
n_i : nombre de fruits contenant x_i parasites	11	29	27	19	10	4

Si l'on suppose que le nombre de parasites suit une loi de Poisson de paramètre λ,

donner l'intervalle de confiance de niveau asymptotiquement égal à 99% pour le paramètre λ.

Exercice 5.

On considère une variable aléatoire réelle continue de densité :

$$f(x) = \begin{cases} 0 & \text{si } x < 2, \\ \theta \exp(-\theta(x-2)) & \text{si } x \geq 2, \end{cases}$$

avec $\theta > 0$.

1. Vérifier que cette loi appartient à la famille exponentielle.
2. En utilisant les propriétés de l'EMV de θ, construire un intervalle de confiance pour θ de niveau asymptotiquement égal à $1-\alpha$.
3. Application numérique : Calculer cet intervalle de confiance pour $n = 200$, $\bar{x}_n = 6,68$ et $\alpha = 5\%$.

Exercice 6.

On considère n_1 variables aléatoires réelles $X_{1,1}, \ldots, X_{1,n_1}$ i.i.d. de loi $\mathcal{N}(\mu_1, \sigma^2)$ et

n_2 variables aléatoires réelles $X_{2,1}, \ldots, X_{2,n_2}$ i.i.d. de loi $\mathcal{N}(\mu_2, \sigma^2)$. On suppose de plus les variables $X_{k,i}$ ($k = 1, 2$ et $i = 1, \ldots, n_k$) mutuellement indépendantes.

1. Soient $\overline{X}_1 = \dfrac{1}{n_1} \displaystyle\sum_{i=1}^{n_1} X_{1,i}$ et

   $\overline{X}_2 = \dfrac{1}{n_2} \displaystyle\sum_{i=1}^{n_2} X_{2,i}$. Quelle est la loi de $\overline{X}_1 - \overline{X}_2$?

2. Soit $S^2 = \dfrac{1}{n_1 + n_2 - 2} \left[\displaystyle\sum_{i=1}^{n_1} (X_{1,i} - \overline{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2,i} - \overline{X}_2)^2 \right]$.

   Quelle est la loi de $(n_1 + n_2 - 2)\dfrac{S^2}{\sigma^2}$?

3. En déduire un intervalle de confiance de niveau $1-\alpha$ pour le paramètre $\mu_1 - \mu_2$.

4. Application numérique : On a observé $n_1 = 10$ et $n_2 = 8$ observations dans chacune des deux populations considérées. Les données obtenues sont les suivantes :

   $x_{1,i}$: 1,36 2,66 2,05 1,85 2,28 1,71 0,75 1,97 1,70 1,68

   $x_{2,i}$: 1,91 2,03 1,31 1,33 2,68 2,04 0,40 3,31

   Calculer l'intervalle de confiance de niveau 95% pour le paramètre $\mu_1 - \mu_2$.

V Généralités sur les tests

1 Problèmes de test

Le but d'un test statistique est de donner un critère permettant de retenir l'hypothèse $H_0 : \theta \in \Theta_0$ ou de retenir une hypothèse alternative $H_1 : \theta \in \Theta_1$, avec $\Theta_1 \subset \Theta_0^c$.

La mise en œuvre du critère du test détermine une zone de rejet ou zone critique W, W^c est la zone d'acceptation ou zone de confiance. On appelle *risque de première espèce*, notée α, la probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie, $\alpha = \mathbb{P}(W|H_0)$. La probabilité, notée $1-\beta$, de retenir l'hypothèse H_0 alors qu'elle est fausse s'appelle *risque de deuxième espèce*, $1-\beta = \mathbb{P}(W^c|H_1)$. La probabilité β s'appelle *puissance du test.*

Définition V.1 *Un test est donné par une fonction $\Phi : E^n \longrightarrow \{0,1\}$, on retiendra H_0 si $\Phi(X_1,\ldots,X_n) = 0$, on rejette H_0 si $\Phi(X_1,\ldots,X_n) = 1$. On appelle zone de rejet l'ensemble $R = \{\Phi(X_1,\ldots,X_n) = 1\}$. Évidemment, étant donnée une zone de rejet $R \subset E^n$, on définit un test en posant $\Phi = \mathbb{1}_R$.*

Lorsque $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$, on parle d'hypothèses simples.

Définition V.2 *Le niveau du test - ou sa sensibilité - est la probabilité de rejeter H_0 à tort :*

$$\alpha = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta((X_1,\ldots,X_n) \in R)).$$

La puissance du test est la fonction $\beta : \Theta_1 \longrightarrow [0,1]$ définie par $\beta(\theta) = \mathbb{P}_\theta((X_1,\ldots,X_n) \in R)$. Le test est dit sans biais si $\beta(\theta) \geq \alpha\ \forall\theta \in \Theta_1$.

2 Tests uniformément plus puissants

Définition V.3 *Étant donnés deux tests Φ_1 et Φ_2 de niveau $\leq \alpha$ pour tester l'hypothèse $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$. Le test Φ_1 est uniformément plus puissant (u.p.p.) que Φ_2 ssi $\forall\theta \in \Theta_1$, $\beta_1(\theta) \geq \beta_2(\theta)$. Dans le cas où $\Theta_1 = \{\theta_1\}$ et $\Theta_0 = \{\theta_0\}$, on parle de test plus puissant (p.p.).*

Dans un premier temps, on supposera que H_0 et H_1 sont des hypothèses simples :

$H_0 \ : \theta = \theta_0$, $H_1 \ : \theta = \theta_1$.

Soit $L(x_1, \ldots, x_n, \theta)$ la fonction de vraisemblance de $(X_1, \ldots, X_n)$.

Définition V.4 *On considère des hypothèses simples $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$. Soit*

$$V_{\theta_0,\theta_1}(x) = \frac{L(x, \theta_1)}{L(x, \theta_0)}$$

le rapport de vraissemblance. On considère la famille de tests Φ_k dont la région critique R_k est de la forme :

$$V_{\theta_0,\theta_1}(x) > k.$$

On appellera test de Neyman et Pearson *tout test de cette forme.*

Théorème V.1 (Lemme de Neyman-Pearson)

1. *Soit $\alpha > 0$, si Φ_k est un test de niveau α alors il est p.p. que tout autre test de niveau $\leq \alpha$, de plus il est sans biais.*
2. *Si $\alpha \in]0, 1[$,si les lois $\mathbb{P}_\theta$ sont absoluement continues, il existe $k_\alpha \in \mathbb{R}$ tel que Φ_{k_α} est de niveau α. Si les lois $\mathbb{P}_\theta$ sont discrètes alors il existe un plus petit k_α tel que Φ_{k_α} est de niveau $\leq \alpha$.*
3. *Soit Φ un test p.p. de niveau α alors $\forall \theta \in \{\theta_0, \theta_1\}$,*

$$\mathbb{P}_\theta(\Phi(X) \neq \Phi_{k_\alpha}(X) \text{ et } V(X) \neq k) = 0.$$

Ce résultat permet de construire des tests optimisant la puissance dans le cas d'hypothèses simples. Le point 3. montre que les tests de Neymann et Pearson sont les seuls tests p.p. (dans le cas absoluement continu). Dans le cas d'hypothèses composites (par exemple H_1 n'est pas réduite à un point), le lemme de Neyman et Pearson permet d'obtenir des tests u.p.p., si les rapports de vraisemblance sont monotones.

Proposition V.1 *Si T est une statistique exhautive dont la fonction de vraissemblance $g(t, \theta)$ vérifie : pour $\theta > \theta_0$,*

$$\frac{g(t, \theta)}{g(t, \theta_0)} \text{ est une fonction croissante de } t.$$

Alors le test de zone de rejet $\widetilde{R_k} = \{T > k\}$ est u.p.p. pour tester $\theta = \theta_0$ contre $\theta > \theta_0$.

3 Tests fondés sur le rapport du maximum de vraisemblance

Définition V.5 *On appellera test du maximum de vraisemblance tout test fondé sur la région critique*

$$W = \left\{ x \in \mathbb{R}^n \ / \ \frac{\sup_{\theta\in\Theta_1} L(x,\theta)}{\sup_{\theta\in\Theta_0} L(x,\theta)} > k_\alpha \right\},$$

où k_α est choisit tel que $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(W) = \alpha$.

Dans le cas où $\Theta_1 = \Theta_0^c$, on considèrera le test de région critique :

$$R = \left\{ x \in \mathbb{R}^n \ / \ \lambda(x) = \frac{\sup_{\theta\in\Theta} L(x,\theta)}{\sup_{\theta\in\Theta_0} L(x,\theta)} > k_\alpha \right\}.$$

Décision et "p-value"

Lorsqu'on procède à un test, on fixe l'hypothèse H_0, par exemple $\theta = \theta_0$, on choisit une hypothèse alternative H_1 :

$H_1 : \theta \neq \theta_0$ hypothèse bilatérale

$H_1 : \theta < \theta_0$ ou $H_1 : \theta > \theta_0$ hyothèses unilatérale.

On se fixe un risque de première espèce α, on détermine la région critique (i.e. pour un test basé sur le rapport de vraissemblance, la valeur de k_α), on calcule la valeur expérimentale de la statistique du test Z_{exp}, si Z_{exp} est dans la région critique, on rejette H_0 (et on retient H_1), sinon on retient H_0.

La plupart des logiciels de statistique permettent de calculer la valeur de k_α mais fournissent aussi un autre renseignement : la "p-value" p. Si le test est de région critique $f(Z) > k$ où Z est la statistique du test ($f(x) = x$ ou $f(x) = |x|$), la p-value est la probabilité : $p_{\text{value}} = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(f(T) > Z_{\text{exp}})$, on remarque que $Z_{\text{exp}} \geq k \Leftrightarrow p_{\text{value}} \leq \alpha$.

Si $p < \alpha$, on rejette H_0 (et on retient H_1), sinon, on retient H_0. On prendra garde que certains logiciels ne fournissent la valeur de p que pour des hypothèses alternatives bilatérales. Dans le cas de distributions symétriques (normale, Student), on passe du $p_{\text{bilatéral}}$ au $p_{\text{unilatéral}}$ en divisant par 2.

4 Tests asymptotiques

Définition V.6 *On considére une suite $(E^{(N)}, \mathcal{B}^{(N)}, (\mathbb{P}_\theta^{(N)})_{\theta\in\Theta})$ de modèles d'échantillonages paramétriques ayant le même espace de paramètres Θ. On note $X^{(N)}$ le vecteur aléatoire correspondant.*

Le niveau asymptotique d'une suite de tests de Θ_0 contre Θ_1, de région de rejet $R^{(N)}$ est la limite (lorsqu'elle existe)

$$\alpha = \lim_{N\to\infty} \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(X^{(N)} \in R_{(N)}).$$

On dit que la suite de tests est convergente si

$$\forall \theta \in \Theta_1 \lim_{N\to\infty} \mathbb{P}_\theta(X^{(N)} \in R_{(N)}) = 1.$$

a. Propriétés asymptotiques des tests du maximum de vraissemblance

On se place dans le cas où $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta \neq \theta_0\}$. On considère alors

$$\lambda_n(X) = \frac{\sup_{\theta\in\Theta} L_n(X,\theta)}{L_n(X,\theta_0)}.$$

Théorème V.2 *On suppose que les hypothèses de régularité du modèle $(H1)-(H7)$ sont satisfaites et que $\Theta \subset \mathbb{R}$ (paramètre de dimension 1). La suite de test de région critique :*

$$R_n = \{2\ln\lambda_n > K\},$$

où K est le $1-\alpha$ quantile d'une loi $\chi^2(1)$ est de sensibilité asymptotique α et convergente.

b. Tests de Wald et du score

Proposition V.2 *On considère $\widetilde{\theta}_n$ une suite d'estimateurs asymptotiquement efficace d'un paramètre $\theta \in \mathbb{R}^d$, on suppose de plus que $I(\widetilde{\theta}_n) \xrightarrow{\mathbb{P}_\theta} I(\theta)$. Soit $g : \Theta \longrightarrow \mathbb{R}^k$ une fonction C^1 de matrice jacobienne $D(\theta)$ de rang k. On considère $\Theta_0 = \{\theta \in \Theta \ / \ g(\theta) = 0\}$. Le test de Wald de région de rejet :*

$$R_n \ : \xi_n > \chi^2_{1-\alpha}(k) \text{ avec } \xi_n = n g(\widetilde{\theta}_n)^t \left(D(\widetilde{\theta}_n) I(\widetilde{\theta}_n)^{-1} D(\widetilde{\theta}_n)^t\right)^{-1} g(\widetilde{\theta}_n).$$

est de sensibilité asymptotique α et convergent pour $\Theta_1 = \{\theta \in \Theta \ / \ g(\theta) \neq 0\}$.

Lemme V.3 *Soit X un vecteur aléatoire gaussien de $\mathbb{R}^d$, d'espérance μ et de matrice de covariance Σ. On suppose que Σ est inversible. Alors*

$$D^2 = (X-\mu)^t \Sigma^{-1} (X-\mu)$$

suit une loi du $\chi^2(d)$.

Lemme V.4 *Soit X_n une suite de vecteurs aléatoires de $\mathbb{R}^d$ telle que $X_n \xrightarrow{\mathcal{L}} X$ et A une matrice $d \times k$. Alors la suite de vecteurs aléatoires de $\mathbb{R}^k$ AX_n converge en loi vers AX.*

Les tests du score sont basés sur la région de rejet :

$$\xi_n^S > \chi^2_{1-\alpha}(k) \text{ avec } \xi_n^S = \frac{1}{n} DL_n(\widehat{\theta}_{0,n})^t I(\widehat{\theta}_{0,n})^{-1} DL_n(\widehat{\theta}_{0,n}),$$

où DL_n est le gradient de la vraisemblance et $\widehat{\theta}_{0,n}$ l'EMV de θ sur Θ_0.

VI Tests paramétriques classiques

Notations :

P : proportion sur un échantillon aléatoire pour une variable de Bernouilli : X prend les valeurs 0 ou 1 avec probabilité $1-\pi$ et π,

$P = \frac{X_1+\cdots+X_n}{n}$.

$\overline{X}$: moyenne sur un échantillon aléatoire de taille d'espérance μ.

S^2 est la variance estimée définie par :

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

Lorsque $\mathbb{E}(X) = \mu$ est connue, on utilise aussi :

$$D^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

1 Tests gaussiens

Test	Hypothèses	Stat. du test (Z) et cond. d'appl.	Loi de Z sous H_0	Remarques
conform. d'une moy.	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ ou $H_1 : \mu > \mu_0$ ou $H_1 : \mu < \mu_0$	$\frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$, la variable étudiée doit suivre une loi normale	Student à $n-1$ d.d.l.	Si σ est connu, on peut utiliser directement $\overline{X}$ qui suit une loi normale de paramètre $\mathcal{N}(\mu_0, \sigma)$ sous H_0.
comp. de deux moy. pour des ech. indép.	$H_0 : \mu_1 = \mu_2 = \mu_0$ $H_1 : \mu_1 \neq \mu_2$ ou $H_1 : \mu_1 < \mu_2$ ou $H_1 : \mu_1 > \mu_2$	$\frac{\overline{X^1} - \overline{X^2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}} \times \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, la variable étudiée doit suivre une loi normale. Il faut que les écart-types σ_1 et σ_2 soient les mêmes.	Student à $n_1 + n_2 - 2$ d.d.l.	Avant de faire ce test, on doit tester l'égalité des variances avec un test de Fisher-Snedecor. Si on accepte l'hypothèse $\sigma_1 = \sigma_2$, on estime la valeur commune $\sigma_1 = \sigma_2 = \sigma$ (voir ci-dessous). Si on refuse l'hypothèse $\sigma_1 = \sigma_2$, on ne peut pas faire le test.
comp. de deux moy. pour des ech. appariés, on pose $Y = X_1 - X_2$	$H_0 : \mu_Y = 0$ $H_1 : \mu_Y \neq 0$ ou $H_1 : \mu_Y < 0$ ou $H_1 : \mu_Y > 0$	$\frac{\overline{X_Y}}{\frac{S_Y}{\sqrt{n}}}$, la variable étudiée doit suivre une loi normale.	Student à $n-1$ d.d.l.	
conf. d'une variance	$H_0 : \sigma = \sigma_0$ $H_1 : \sigma \neq \sigma_0$ ou $H_1 : \sigma > \sigma_0$ ou $H_1 : \sigma < \sigma_0$	$\frac{(n-1)S^2}{\sigma_0^2}$ La variable étudiée doit suivre une loi normale	χ^2 à $(n-1)$ d.d.l.	Il faut tester la normalité. Si μ est connue, on peut remplacer S^2 par D^2 et on a une loi $\chi^2(n)$.
comp de deux variances	$H_0 : \sigma_1 = \sigma_2$ $H_1 : \sigma_1 \neq \sigma_2$ ou $H_1 : \sigma_1 > \sigma_2$ ou $H_1 : \sigma_1 < \sigma_2$	$\frac{S_1^2}{S_2^2}$. **La variable étudiée doit suivre une loi normale**	Fisher Snedecor à $(n_1 - 1, n_2 - 1)$ d.d.l.	Lorsqu'on accepte H_0, on estime l'écart-type commun par $\sigma = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

2 Tests asymptotiques

Test	Hypothèses	Stat. du test (Z) et cond. d'appl.	Loi de Z sous H_0	Remarques
comp. de deux prop. pour des ech. indep.	$H_0 : \pi_1 = \pi_2 = \pi_0$ $H_1 : \pi_1 \neq \pi_2$ ou $H_1 : \pi_1 < \pi_2$ ou $H_1 : \pi_1 > \pi_2$	$\frac{P_1 - P_2}{\sqrt{\hat{\pi}(\hat{\pi}-1) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, $n_1 \geq 30$ et $n_2 \geq 30$, avec $\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$	$\mathcal{N}(0,1)$	Il s'agit d'un test asymptotique.
comp. de deux moy. pour des ech. indép.	$H_0 : \mu_1 = \mu_2 = \mu_0$ $H_1 : \mu_1 \neq \mu_2$ ou $H_1 : \mu_1 < \mu_2$ ou $H_1 : \mu_1 > \mu_2$	$\frac{\overline{X^1} - \overline{X^2}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}} \times \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, $n_1 \geq 30$ et $n_2 \geq 30$	$\mathcal{N}(0,1)$	Pour les grands échantillons, il n'est pas nécessaire d'avoir l'égalité des variances. Il s'agit d'un test asymptotique.
comp. de deux moy. pour des ech. appariés, on pose $Y = X_1 - X_2$	$H_0 : \mu_Y = 0$ $H_1 : \mu_Y \neq 0$ ou $H_1 : \mu_Y < 0$ ou $H_1 : \mu_Y > 0$	$\frac{m_Y}{\frac{S_Y}{\sqrt{n}}}$, $n \geq 30$	$\mathcal{N}(0,1)$	Il s'agit d'un test asymptotique.
conform. d'une prop.	$H_0 : \pi = \pi_0$ $H_1 : \pi \neq \pi_0$ ou $H_1 : \pi > \pi_0$ ou $H_1 : \pi < \pi_0$	$\frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$, $n \geq 30$	$\mathcal{N}(0,1)$	Il s'agit d'un test asymptotique
conform. d'une moy.	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ ou $H_1 : \mu > \mu_0$ ou $H_1 : \mu < \mu_0$	$\frac{\overline{X} - \mu_0}{\frac{S_n}{\sqrt{n}}}$, $n \geq 30$	$\mathcal{N}(0,1)$	Il s'agit d'un test asymptotique

VII Quelques tests non paramétriques

1 Tests du χ^2.

a. Loi multinômiale

On considère r évènements A_1, ..., A_r de probabilité p_1, ..., p_r. On suppose que les A_i forment un système complet d'évènement (i.e. ils sont disjoints et leur union est Ω), en particulier, $\sum_{i=1}^r p_i = 1$.

On répète n fois, de manière indépendante, l'expérience aléatoire dont le résultat est l'un des A_i (penser à un tirage avec remise de n boules dans une urne qui contient des boules de r couleurs différentes, p_i est alors la proportion de boules de couleur i).

On note N_i la variable aléatoire qui donne le nombre de fois (parmi les n expériences) où l'évènement A_i se produit.

N_i suit une loi Binômiale $\mathcal{B}(n, p_i)$.

La loi du vecteur $(N_1, \ldots, N_r)$ est donnée par :

$$\mathbb{P}(N_1 = n_1, \ldots, N_r = n_r) = \frac{n!}{n_1! \cdots n_r!} p_1^{n_1} \times \cdots \times p_r^{n_r},$$

pour $(n_1, \ldots, n_r) \in \mathbb{N}^r$ avec $\sum_{i=1}^r n_i = n$. En particulier, les N_i **ne** sont **pas** indépendants.

b. Loi asymptotique

Théorème VII.1 *Soit*

$$D^2 = \sum_{i=1}^r \frac{(N_i - np_i)^2}{np_i}.$$

Alors, D^2 converge en loi (quand $n \to \infty$) vers une loi du χ^2 à $r-1$ degrés de liberté.

c. Test du χ^2 d'adḱuation à une loi

On présente ici le test du χ^2 d'adéquation à une loi théorique.

On se demande si une variable aléatoire Y suit une loi donnée notée $\mathbb{P}_0$.

Soit Y_1,..., Y_n un échantillon aléatoire indépendant de la loi de Y. On fixe une partition de $\mathbb{R}$ à r éléments $\mathbb{R} = C_1 \cup \cdots \cup C_r$. On note N_1 le nombre d'indices i tels que $Y_i \in C_1$, ... N_r le nombre d'indices i tels que $Y_i \in C_r$. Soient $p_i^0 = \mathbb{P}_0(Y \in C_i)$ les probabilités théoriques, pour une loi $\mathbb{P}$, on note $p_i = \mathbb{P}(Y \in C_i)$ et on teste :

H_0 : pour tout $i = 1, \ldots, r$, $p_i = p_i^0$

H_1 : il existe i tel que $p_i \neq p_i^0$.

Si on retient H_0, on concluera que la loi de Y est P_0.

La statistique du test est

$$D^2 = \sum_{i=1}^{r} \frac{(N_i - np_i^0)^2}{np_i^0}.$$

On rejette H_0 si $D^2 > \chi^2_{r-1,1-\alpha}$.

Si la loi $\mathbb{P}_0$ appartient à une famille paramétrique, $\mathbb{P}_0 = \mathbb{P}_{\theta_0}$, $\theta \in \mathbb{R}^d$, si on connait θ_0, il n'y a pas de différence avec le cas considéré ci-dessus. Si on ne connait pas θ_0 - par exemple, on se demande si la loi de Y est normale - on doit alors estimer θ_0. Soit $\hat{\theta}$ un estimateur du maximum de vraissemblance de θ,

Théorème VII.2 *Soit*

$$\widetilde{D}^2 = \sum_{i=1}^{r} \frac{(N_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}.$$

Alors, sous $\mathbb{P}_\theta$, $\widetilde{D}^2$ converge en loi (quand $n \to \infty$) vers une loi du χ^2 à $r-d-1$ degrés de liberté.

On rejette alors H_0 si $\widetilde{D}^2 > \chi^2_{r-d-1,1-\alpha}$.

D'un point de vue pratique, on considère que l'approximation donnée par le théorème limite ci-dessus est bonne si $n \geq 30$ et que les effectifs théoriques $np_i(\hat{\theta})$ sont supérieures à 5, $i = 1, \ldots, r$. Si cette dernière condition n'est pas vérifiée, on procède à des regroupements de classes.

d. Test du χ^2 d'indépendance

On considère $X = (Y, Z)$, et $X_i = (Y_i, Z_i)$ $i = 1, \ldots, n$, un échantillon aléatoire de loi P_X. Y_i et Z_i sont des variables discrètes prenant respectivement les valeurs : $\{y_1, \ldots, y_\ell\}$ et $\{z_1, \ldots, z_m\}$. On veut tester l'indépendance de Y et Z. Le test se base sur le fait que Y et Z sont indépendants si et seulement si $\mathbb{P}_X = \mathbb{P}_Y \otimes \mathbb{P}_Z$.

Si l'hypothèse d'indépendance est satisfaite, $p_{i,j} = q_i r_j$ avec $p_{i,j} = \mathbb{P}(X = (y_i, z_j))$, $q_i = \mathbb{P}(Y = y_i)$, $r_j = \mathbb{P}(Z = z_j)$. On est dans le cadre ci-dessus avec le paramètre $\theta = (q_1, \ldots, q_{\ell-1}, r_1, \ldots r_{m-1}) \in \mathbb{R}^{\ell+m-2}$. On estime q_i par $\frac{N_{i\cdot}}{n}$ et r_j par $\frac{N_{\cdot j}}{n}$. Soient

$$D_1^2 = n \sum_{i=1}^{\ell} \sum_{j=1}^{m} \frac{\left(N_{i,j} - \dfrac{N_{i\cdot} N_{\cdot j}}{n}\right)^2}{N_{i\cdot} N_{\cdot j}} \quad D_2^2 = \sum_{i=1}^{\ell} \sum_{j=1}^{m} \frac{\left(N_{i,j} - \dfrac{N_{i\cdot} N_{\cdot j}}{n}\right)^2}{N_{i,j}}$$

D_1^2 et D_2^2 convergent en loi vers une loi $\chi^2(\ell-1)(m-1)$ ($(\ell-1)(m-1) = \ell \times m - 1 - (\ell-1) - (m-1)$). Les tests associés aux régions de rejet

$$\{D_1^2 > k\} \text{ et } \{D_2^2 > k\}$$

avec k le $1-\alpha$ quantile d'une loi $\chi^2((\ell-1)(m-1))$ sont de sensibilité asymptotique α et convergents pour les hypothèses

$H_0 : p_{i,j} = q_i r_j$ pour tout (i,j)

H_1 : il existe (i,j) tel que $p_{i,j} \neq q_i r_j$.

Si on retient H_0, on retient l'hypothèse d'indépendance.

Exercice 1 *On souhaite procéder à un test de conformité à une loi de Poisson.*

1. *On rapelle que X suit une loi de Poisson de paramètre $\lambda > 0$ si pour tout $n \in \mathbb{N}$,*

$$\mathbb{P}(X = n) = e^{-\lambda}\frac{\lambda^n}{n!}.$$

Déterminer $E(X)$ et $Var(X)$.

2. *On considère un échantillon aléatoire X_1, ..., X_n, indépendant de loi de Poisson de paramètre λ.*

Quel est l'estimateur du maximum de vraissemblance de λ ?

3. *Proposer un test pour tester :*

H_0 : X suit une loi de Poisson

H_1 : X ne suit pas une loi de Poisson.

4. *Application. Pendant* 100 *intervalles de* 10 *minutes, on a compté le nombre X d'ouvriers se présentant à un magasin pour emprunter des outils. Le tableau suivant donne les valeurs observées pour ces* 100 *mesures et les effectifs correspondants.*

x_i	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
n_i	1	0	1	2	1	3	5	6	9	10	11	9	8	9	7	5	4	3	1	1	1

Peut-on conclure que X suit une loi de Poisson ?

Exercice 2 *On procède à un sondage téléphonique, il est demandé aux sondés s'ils sont optimistes ou non quant à leur capacité d'achat pour les années à venir. Les résultats sont présentés par catégories d'âge.*

Age	*Optimistes*	*Pas optimistes*
$[20, 40[$	237	392
$[40, 60[$	326	298
≥ 60	362	258

Peut-on considérer que le fait d'être optimiste quand à sa capacité d'achat est indépendante de l'âge ?

Un des inconvénients du test d'adéquation du χ^2 est le choix des classes. Cet inconvénient n'est plus présent pour le test de Kolmogorov-Smirnov.

2 Test de Kolmogorov-Smirnov

On considère X_1, ..., X_n un échantillon aléatoire indépendant de même loi que X. Soit F la fonction de répartition de X. On défini la fonction de répartition empirique :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}},$$

Attention, c'est une variable aléatoire.

Soit

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

Pour tout $x \in \mathbb{R}$, $F_n(x)$ s'écrit comme une somme de variables aléatoires indépendentes de loi $\mathcal{B}(F(x))$. On en déduit :

— $F_n(x)$ converge presque sûrement (et en probabilité) vers $F(x)$,

— $\sqrt{n}(F_n(t) - F(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(t)(1 - F(t)))$.

Théorème VII.3 (Théorème de Glivenko-Cantelli) *D_n converge vers 0 presque sûrement.*

Proposition VII.1 *La loi de D_n ne dépend pas de la loi de X. Plus présicément, dans le cas où F est strictement croissante et continue, on a les égalités en loi :*

$$\begin{aligned} D_n &\overset{\mathcal{L}}{=} \sup_{t \in [0,1]} |H_n(t) - t| \\ &\overset{\mathcal{L}}{=} \max \left\{ \left| \frac{i}{n} - t \right| \ / \ U_{(i)} \leq t < U_{(i+1)} \right\}, \end{aligned}$$

où H_n est la fonction de répartition empirique d'une suite $(U_i)_{i=1,...,n}$ de vaiid de loi uniforme $\mathcal{U}([0,1])$ et $(U_{(i)})_{i=1,...,n}$ désigne les statistiques de l'ordre associées à $(U_i)_{i=1,...,n}$ i.e. les $U_{(i)}$ vérifient

$$U_{(1)} < U_{(2)} \cdots < U_{(n)} \text{ et } U_{(1)} = \min_{i=1,...,n} U_i \ , \ U_{(n)} = \max_{i=1,...,n} U_i.$$

La loi de $K_n = \sqrt{n}D_n$ (loi de Kolomogorov-Smirnov à 1 échantillon) est tabulée et converge en loi vers une variable aléatoire K elle aussi tabulée. Ce résultat permet de tester si un échantillon provient d'une loi théorique connue. Attention : le résultat n'est pas valable si les paramètres de la loi sont estimés. Avec R : `ks.test`.

Si Y_1, ..., Y_m est un échantillon de la loi de Y. On note $G_m(x)$ la fonction de répartition empirique associée. Soit

$$D_{n,m} = \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)|.$$

Si la loi de X est la même que celle de Y, la loi de $D_{n,m}$ est la même que la loi de $\sup_{t\in[0,1]} |H_n(t) - I_m(t)|$ où H_n et I_m sont les fonction de répartition empiriques de suites de variables aléatoires uniformes $\mathcal{U}([0,1])$, on a aussi l'égalité en loi :

$$D_{n,m} \overset{\mathcal{L}}{=} \max\left\{\left|\frac{i}{n} - \frac{j}{m}\right|,\ U_{(i)} < V_{(j)} < U_{(i+1)}\right\},$$

où $(U_{(i)})_{i=1,\dots,n}$ désigne les statistiques de l'ordre associées à $(U_i)_{i=1,\dots,n}$, $U_i \rightsquigarrow \mathcal{U}([0,1])$ et $(V_{(j)})_{j=1,\dots,m}$ désigne les statistiques de l'ordre associées à $(V_j)_{j=1,\dots,m}$, $V_j \rightsquigarrow \mathcal{U}([0,1])$. Pour

$$c_{n,m} = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}},$$

$K_{n,m} = c_{n,m}D_{n,m}$ suit une loi de Kolmogorov-Smirnov à deux échantillons et converge en loi vers une variable aléatoire K_2 elle aussi identifiée. On peut donc tester si deux échantillons proviennent de la même loi (avec R : `ks.test`).

3 Test de Shapiro-Wilk

Le test de Shapiro-Wilk permet de tester la normalité d'un échantillon, quel que soit sa taille et sans estimer les paramètres de la loi.

a. Droite de Henry

Il s'agit de représenter les quantiles théoriques d'une loi "connue"en fonction des données x_i.

Soit F_i les fréquences cumulées empiriques, on note $u_i^\star$ le quantile de la loi théorique correspondant : $\mathbb{P}(Z \leq u_i^\star) = F_i$. Si le graphe $(x_i, u_i^\star)$ est quasiment une droite, alors, la loi empirique est proche d'une transformation affine de la loi théorique. En particulier, si la loi théorique condidérée est une loi $\mathcal{N}(0,1)$ alors la loi empirique est proche d'une loi normale (avec R : `qqnorm`, la commande `qqline` rajoute une droite qui passe par les premiers et troisièmes quantiles, la commande `qqplot` trace la droite de Henry pour deux échantillons).

Exemple : la distribution suivante qui donne des résultats d'essais de fatigue d'un matériau (nombre de cycles avant rupture) :

225	31	400	62	850	39	89	580	115	442	270	125	342	251	140

b. Test de Shapiro-Wilk

Ce test est spécifique à la loi normale. Son principal avantage est qu'il ne requière pas d'estimation préalable des paramètres. L'idée est de tester la proximité du nuage de points des écarts inter-quartiles empiriques et des écarts inter-quartiles d'une loi normale centrée réduite, à la droite des moindres carrés correspondante. Si $(U_1, \ldots, U_n)$ est un échantillon aléatoire indépendant de loi $\mathcal{N}(0,1)$, on note $V = (V_1, \ldots, V_n)$ l'échantillon ordonné : $V_1 = \min_{i=1,\ldots,n} U_i$, ..., $V_n = \max_{i=1,\ldots,n} U_i$ (voir ci-dessous pour les détails sur les satistiques de l'ordre). μ est le vecteur d'espérance de V, $\Sigma = \mathbb{E}[(V-\mu)(V-\mu)^t]$, enfin

$$a^t = \mu^t \Sigma^{-1} (\mu^t \Sigma^{-2} \mu)^{-\frac{1}{2}},$$

$a = (a_1, \ldots, a_n)$.

Étant donné un échantillon aléatoire indépendant $X = (X_1, \ldots, X_n)$ et $Y = (Y_1, \ldots, Y_n)$ sa statistique de l'ordre, on définit :

$$T_n = \frac{1}{n-1} \left[\sum_{i=1}^{[\frac{n}{2}]} a_{n-i+1} (Y_{n-i+1} - Y_i) \right]^2 .$$

Sous H_0 : les X_i suivent des lois normales $\mathcal{N}(\nu, \sigma)$, T_n est un estimateur asymptotiquement sans biais de σ^2.

La statistique du test de Shapiro-Wilk est :

$$SW = \frac{T_n}{S_n^2}.$$

La loi de SW est indépendante de ν et σ, cette statistique est aussi indépendante de $\overline{X}n$ et de S_n^2.

Pour mettre en œuvre ce test, on dispose de tables qui donnent les a_i et les valeurs critiques de la statistique SW.

n **j**	2	3	4	5	6	7	8	9	10	-
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739	-
2		0	0,1677	0,2413	0,2806	0,3031	0,3164	0,3244	0,3291	-
3				0	0,0875	0,1401	0,1743	0,1976	0,2141	-
4						0	0,0561	0,0947	0,1224	-
5								0	0,0399	-
n **j**	11	12	13	14	15	16	17	18	19	20
1	0,5601	0,5475	0,5359	0,5251	0,515	0,5056	0,4963	0,4886	0,4808	0,4734
2	0,3315	0,3325	0,3325	0,3318	0,3306	0,329	0,3273	0,3253	0,3232	0,3211
3	0,226	0,2347	0,2412	0,246	0,2495	0,2521	0,254	0,2553	0,2561	0,2565
4	0,1429	0,1586	0,1707	0,1802	0,1878	0,1939	0,1988	0,2027	0,2059	0,2085
5	0,0695	0,0922	0,1099	0,124	0,1353	0,1447	0,1524	0,1587	0,1641	0,1686
6	0	0,0303	0,0539	0,0727	0,088	0,1005	0,1109	0,1197	0,1271	0,1334
7			0	0,024	0,0433	0,0593	0,0725	0,0837	0,0932	0,1013
8					0	0,0196	0,0359	0,0496	0,0612	0,0711
9							0	0,0163	0,0303	0,0422
10									0	0,014
n **j**	21	22	23	24	25	26	27	28	29	30
1	0,4643	0,459	0,4542	0,4493	0,445	0,4407	0,4366	0,4328	0,4291	0,4254
2	0,3185	0,3156	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2968	0,2944
3	0,2578	0,2571	0,2563	0,2554	0,2543	0,2533	0,2522	0,251	0,2499	0,2487
4	0,2119	0,2131	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,215	0,2148
5	0,1736	0,1764	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1064	0,187
6	0,1399	0,1443	0,148	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,163
7	0,1092	0,115	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395	0,1415
8	0,0804	0,0878	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192	0,1219
9	0,053	0,0618	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002	0,1036
10	0,0263	0,0368	0,0459	0,0539	0,061	0,0672	0,0728	0,0778	0,0822	0,0862
11	0	0,0122	0,0228	0,0321	0,0403	0,0476	0,054	0,0598	0,065	0,0697
12			0	0,0107	0,02	0,0284	0,0358	0,0424	0,0483	0,0537
13					0	0,0094	0,0178	0,0253	0,032	0,0381
14							0	0,0084	0,0159	0,0227
15									0	0,0076

N	5%	1%	N	5%	1%
3	0,767	0,753	25	0,918	0,888
4	0,748	0,687	26	0,92	0,891
5	0,762	0,686	27	0,923	0,894
6	0,788	0,713	28	0,924	0,896
7	0,803	0,73	29	0,926	0,898
8	0,818	0,749	30	0,927	0,9
9	0,829	0,764	31	0,929	0,902
10	0,842	0,781	32	0,93	0,904
11	0,85	0,792	33	0,931	0,906
12	0,859	0,805	34	0,933	0,908
13	0,856	0,814	35	0,934	0,91
14	0,874	0,825	36	0,935	0,912
15	0,881	0,835	37	0,936	0,914
16	0,837	0,844	38	0,938	0,916
17	0,892	0,851	39	0,939	0,917
18	0,897	0,858	40	0,94	0,919
19	0,901	0,863	41	0,941	0,92
20	0,905	0,868	42	0,942	0,922
21	0,908	0,873	43	0,943	0,923
22	0,911	0,878	44	0,944	0,924
23	0,914	0,881	45	0,945	0,926
24	0,916	0,884	46	0,945	0,927
47	0,946	0,928	48	0,947	0,929
49	0,947	0,929	50	0,947	0,93

Tester la normalité de la distribution suivante qui donne des résultats d'essais de fatigue d'un matériau (nombre de cycles avant rupture) :

225	31	400	62	850	39	89	580	115	442	270	125	342	251	140

Quel autre ajustement pourrait-on proposer ?

4 Tests de rang

Il s'agit de tests non paramétriques de comparaison. De manière générale, on préfère effectuer des tests paramétriques, en effet, les tests non paramétriques sont moins sensibles ; c'est à dire que, pour un test non paramétrique, la probabilité d'accepter H_0 alors que H_0 est fauusse est plus importante, par contre lorsque l'on rejette H_0, on peut être raisonablement confiant quand à cette conclusion).

Dans les tests du rang, les valeurs observées sont remplacées par leurs rangs au sein des échantillons. L'idée du test est la suivante : on ordonne toutes les valeurs observées (i.e. les valeurs de tous les échantillons concernés), si le facteur étudié a une influence, les valeurs d'un des échantillons seront "dans les premiers" parmi les valeurs ordonnées.

a. Statistiques de l'ordre, de rang

Si $(X_1, \ldots, X_n)$ est un échantillon aléatoire indépendant i.d., on lui associe le vecteur aléatoire $X_o = (X_{(1)}, \ldots, X_{(n)})$: échantillon ordonné.

$$X_{(1)} = \min_{i=1,\ldots,n} X_i \leq X_{(2)} \leq \cdots < X_{(n)} = \max_{i=1,\ldots,n} X_i.$$

La loi de $X_{(1)}$ a pour fonction de répartition :

$$F_1(t) = 1 - [1 - F(t)]^n \text{ où } F \text{ est la fonction de répartition de } X.$$

La loi de $X_{(n)}$ a pour fonction de répartition :

$$F_n(t) = [F(t)]^n.$$

Plus généralement, on obtient :

$$\mathbb{P}(X_{(k)} < t) = \sum_{i=k}^{n} C_n^i [F(t)]^i (1 - F(t))]^{n-i}.$$

Définition VII.1 *Le rang de X_i dans la liste $X_1, \ldots, X_n$ est :*

$$R_i = 1 + \sum_{j \neq i} \mathbb{1}_{X_j < X_i}.$$

C'est le rang occupé par X_i dans la suite ordonnée $X_{(1)} < \cdots < X_{(n)}$.

b. Le test de Wilcoxon

Il s'agit de comparer deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$, indépendants. Sont-ils issus de la même loi ? Soit $N = n + m$ et

$$(Z_1, \ldots, Z_N) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

l'échantillon concaténé. On considère les statistiques d'ordre et du rang attachées à cet échantillon :

$$Z_{(1)} < \ldots < Z_{(N)}, \; R_Z(i) = 1 + \sum_{j \neq i} \mathbb{1}_{Z_j < Z_i}.$$

Si X et Y ont même loi, alors la variable aléatoire R_Z, à valeur dans l'ensemble des permutations de $\{1, \ldots, N\}$ est uniforme ($\mathbb{P}(R_Z = \sigma) = \frac{1}{N!}$), cette loi est indépendante de la loi commune de X et Y.

On note W_X la somme des rangs des X_i : $W_X = \sum_{i=1}^{n} R_Z(i)$. On montre que

$$\mathbb{E}(W_X) = \frac{n(N+1)}{2} \text{ et } \mathrm{Var}(W_X) = \frac{nm(N+1)}{12}.$$

La loi de $W_X - \frac{n(n+1)}{2}$ est tabulée et permet de construire un test de comparaison de deux échantillons.

Exemple : on veut comparer les performances de deux groupes d'élèves à des tests d'habilité manuelle. Les performances en minutes sont les suivantes :

Groupe I	22	31	14	19	24	28	27	29		
Groupe II	25	13	20	11	23	16	21	18	17	26

On se demande s'il y a une différence significative entre les deux groupes (avec R : `wilcox.test`).

VIII Exemples d'estimation non paramétrique

1 Estimation d'une densité de probabilité

On considère X_1, ..., X_n un échantillon aléatoire indépendant de même loi que X.

a. Histogramme empirique

Une première approximation de la densité est fournie par l'histogramme. Pour cela, on choisit des classes : $[x_0, x_1]$, $]x_1, x_2]$, ..., $]x_{k-1}, x_k]$, l'histogramme est constitué pour chaque classe d'un rectangle de hauteur $\widehat{f_i} = \frac{N_i}{n(x_i - x_{i-1})}$, où

$$N_i = \sum_{j=1}^{n} \mathbb{1}_{]x_{i-1}, x_i]}(X_j).$$

Il s'agit d'une approximation de l'histogramme théorique ($\widehat{f_i}$ converge vers $\frac{\mathbb{P}(x_{i-1} < X \leq x_i)}{x_i - x_{i-1}}$). Si x_i et x_{i-1} convergent vers x alors ce rapport converge vers $f(x)$. Considérons des classes toutes de même taille h. Alors on considère $\widehat{f_n}(x) = \frac{N_i}{nh}$ si $x_{i-1} < x \leq x_i$. Le problème est de choisir les x_i.

b. Fenêtres mobiles

Une réponse à ce problème du choix des x_i est donnée par les fenêtres mobiles : pour $x \in \mathbb{R}$, $I_x = [x - \frac{h}{2}, x + \frac{h}{2}]$, soit

$$N_x = \sum_{j=1}^{n} \mathbb{1}_{\{X_j \in I_x\}},$$

et

$$\widehat{f_n}(x) = \frac{1}{nh} N_x = \frac{1}{nh} \sum_{j=1}^{n} \mathbb{1}_I \left(\frac{x - X_j}{h} \right)$$

avec $I = [-\frac{1}{2}, \frac{1}{2}]$. On peut montrer que si $h \to 0$ et $nh \to \infty$ alors $\widehat{f_n}(x)$ converge vers $f(x)$. On a aussi un théorème de la limite centrale fonctionnel.

c. Versions lisses

L'approximation ci-dessus est assez irrégulière (à cause de la fonction $\mathbb{1}_I$). Pour obtenir un estimateur plus régulier, on peut remplacer $\mathbb{1}_I$ par une fonction régulière K appelée noyau. Par exemple :

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x^2 \right) \text{ (noyau gaussien)},$$

$$K(x) = \frac{3}{4 \times \sqrt{5}} \left(1 - \frac{u^2}{5} \right) \text{ si } |u| < \sqrt{5} \text{ (noyau d'Epanechnikov)}.$$

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right).$$

2 Estimation des quantiles

L'estimation des quantiles revêt un intérêt particulier. Par exemple, la VaR (Value at Risk) utilisée comme indicateur de risque dans de nombreux domaines, n'est rien d'autre qu'un quantile.

La fonction quantile d'une distribution de probabilités est l'inverse généralisé de la fonction de distribution F :

$$F^{-1}(p) = \inf\{x \in \mathbb{R},\ F(x) \geq p\}.$$

a. Quantiles empiriques

On définit alors la fonction quantile empirique F_n^{-1} comme l'inverse généralisé de la fonction de répartition empirique F_n.

On admettra que $F_n^{-1}(p)$ converge vers $F^{-1}(p)$ en tout point de continuité de F^{-1} si et seulement si $F_n(t)$ converge vers $F(t)$ en tout point de continuité de F.

b. Lien avec les statistiques d'ordre

Étant donné un échantillon aléatoire iid $(X_1, \ldots, X_n)$, on note $(X_{(1)}, \ldots, X_{(n)})$ la statistique d'ordre associée.

On a la relation suivante :

$$\forall p \in \left]\frac{i-1}{n}, \frac{i}{n}\right],\ F_n^{-1}(p) = X_{(i)}.$$

c. Résultats asymptotiques

On a le résultat asymptotique suivant dans le cas où la fonction de répartition F est différentiable : pour tout $p \in]0, 1[$,

$$\sqrt{n}(F_n^{-1}(p) - F^{-1}(p)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(p))$$

avec

$$\sigma^2(p) = \frac{p(1-p)}{f(F^{-1}(p))^2}.$$

Ce résultat pose ainsi la question de l'estimation de la densité.

L'utilisation de la *transformation par quantiles* peut permettre de trouver des intervalles de confiance pour les quantiles, sans passer par l'estimation de la densité, dans le cas d'une fonction de répartition strictement croissante et continue. On note $U_1 = F(X_1), \ldots, U_n = F(X_n)$. Les U_i sont i.i.d. de loi uniforme sur $[0, 1]$. On note $U_{(1)}$, ..., $U_{(n)}$ les statistiques d'ordre associées. On a alors

$$\mathbb{P}(X_{(k)} < F^{-1}(p) \leq X_{(\ell)}) = \mathbb{P}(U_{(k)} < p \leq U_{(\ell)}).$$

On admet que pour

$$\frac{k}{n} = p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \quad \frac{\ell}{n} = p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

on a :

$$\mathbb{P}(U_{(k)} < p \leq U_{(l)}) \longrightarrow 1 - \alpha.$$

On peut alors choisir $X_{(k)}$ et $X_{(\ell)}$ comme bornes d'un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour $F^{-1}(p)$.

IX TP

ISFA- M1
Statistique inférentielle

Fiche de TP n^o 1

Une introduction au logiciel R

*** Important ***

Vous enregistrerez toutes vos commandes dans un document que vous nommerez "TP1.R" et que vous sauverez dans un répertoire "IntroR".

1 Commandes et fonctions élémentaires

Commande	description	Commande	Description
`a+b`	addition	`x<-a`	assigner x à la valeur a
`-`	soustraction	`ls()`	lister les objets du repertoire
`*`	multiplication	`x`	afficher la valeur de x
/	division	`rm(x)`	supprimer l'objet x
$\wedge$	puissance	`pi`	valeur de π
`log(a)`	logarithme népérien	`sin, cos, tan`	les fonctions circulaire
`exp(a)`	exponentielle	`asin, acos, atan`	réciproques des fonctions circulaires
`sqrt(a)`	racine carré	`?commande`	obtenir de l'aide

2 Vecteurs et matrices

1. Taper les commandes de la colonne "Commande", du tableau suivant. Au vu du résultat, essayez de comprendre chacune d'entre-elles (au besoin, utiliser l'aide en ligne) et complétez la colonne "Description" du tableau.

Commande	Descri.	Commande	Descri
`> x<-c(1,2,5)`		`> t<-rep(T,4)`	
`> y <- c("a","b","c")`		`> S=array(1.2,5)`	
`> z <- c(x,y)`		`> A<-array(2,c(5,5))`	
`> x<-seq(1,5,by=0.5)`		`> A%*%T`	
`> y<-rep(c(1,"abc"),4)`		`> t(T)%*%S`	
`> z<-sample(1 : 100,30)`		`> max(abs(T-S))`	
`> r<-sample(1 : 100,30,rep=T)`		`> t<-matrix(t,nrow=5,byrow=T)`	
`> T<-1:5`			

Notons que les opérations arithmétiques vues précédemment pour les scalaires s'appliquent de la même manière sur des vecteurs. Par exemple, on peut additionner deux vecteurs a et b, et effectuer les produits scalaires ou membre à membre classiques :

2. — Créez un vecteur d'entiers allant de 1 à 10 de trois façons différentes
   — Créez un vecteur caractère avec successivement les noms de 5 villes de France, puis 5 des départements où elles appartiennent.
   — Créez un vecteur numérique contenant dans l'ordre la population approximatif des ces villes et départements.
   — Créez un vecteur caractère contenant 5 fois "homme" puis 5 fois "femme" en vous servant de la fonction `rep()`.
   — Gardez ces trois derniers vecteurs en mémoire, pensez à leur donner des noms explicites.

3 Quelques fonctions R

Nous avons vu quelques fonctions comme `rep(), seq(), sample()` etc., d'autres fonctions sont utiles pour manipuler des données. La composition des fonction est valable sous R.

`rev()`: permet de renverser l'ordre d'une séquence

```
> X <- rep(seq(1:5),2)
> rev(X)
```

De nombreuses fonctions permettent également de classer les éléments (ou les indices) d'un vecteur, de les sommer, calculer le max le min etc. Que vous renvoient les commandes suivantes ?

`> a <- c(1,3,2,7,4)  > sort(a)` ; `order(a)  > sum(a)` ;`length(a)  > min(a)` ;`max(a)`

Les éléments d'un vecteur peuvent avoir des noms. La fonction `names()` permet en effet d'associer une étiquette à chacun des éléments d'un vecteur. Taper le programme suivant :

```
> x <- 1:5
> names(x) <- c("a","b","c","d","e")
> v=c(1,2,3,4)
> names(v)=c('alpha','beta','gamma','delta')
> v['beta']
```

Les fonctions `cbind` et `rbind` permettent de manipuler des vecteurs de manière à former une matrice par concaténation sur les colonnes ou sur les lignes. Taper et commenter le code suivant :

```
> x <- seq(1:5)
> y <- 2*x
> cbind(x,y)
> xy <- rbind(x,y)
> M<-matrix(1:20, nrow=5, byrow=T)
> M[1,]; M[,1]; M[1,1]
> M[1,]==1;M[1,]>=1;M[1,]<=1
> dim(M)
> M1<-M[,c(2,1,2)]
> M2<-M[-1,]
```

```
> d<-det(M[1:2,1:2]);D<-solve(M[1:2,1:2])
```

4 Graphiques

Reproduire et commenter les programmes suivants

1. ```
   > x=(1:100)/100
   > plot.new()
   > lines(x,x^2)
   > axis(1)
   > axis(2)
   > title("Fonction carré")
   > z=(1:20)/20
   > op<-par(col="red")
   > points(z,z^2)
   > plot.new()
   > plot(x,x^2)
   > plot(x,x^3,'l')
   > par(op)
   ```
2. ```
   > T<-seq(-10,10,by=0.05)
   > D1<-dnorm(T,mean=0,sd=sqrt(5))
   > D2<-dexp(T,rate=0.15)
   > par(mfrow=c(1,2))
   > plot(T, D1, type="o", pch=3, xlab="x", ylab="Densité", col =
   + 3,lty=1, main="Normale-Exponentielle")
   > lines(T,D2, type="o", pch=4,col ="red")
   > legend(-10, 0.15, c( "norm", "expo"), col = c(3,"red"), pch=
   +c(3,4), text.col="green4", lty = c(1,2))
   > plot(T,sin(T),type="l",col=5,lwd=0.5)
   > abline(h=0.5)
   > abline(h=0,lwd=3,lty=2)
   > abline(0,0.1)
   > abline(v=5)
   ```

5 Programmation et création de fonctions

Voici des programmes R, taper-les et comprendre ce qu'ils calculent.

1. ```
   > x<-array(0,5)
   > for (i in 1:5)
   + x[i]=5-i
   > x
   ```
2. ```
   f = function (x){
   + return(x^2+1)}
   > f(8)
   ```

3.
```
> g<-function(x,y)
+ { z=matrix(ncol=length(x),nrow=length(y))
+ z[1,]=0
+ for(i in 2:length(y))
+ for(j in 1:length(x))
+ {
+ z[i,j]=z[(i-1),j]+1
+ }
+ return(z)
}
> g(1:3,rnorm(4,0,1))
```

— Ecrire deux fonctions appelées `beta` et `Beta` prenant deux arguments x et n et calculant les valeurs :

$$\texttt{beta}(x,n) = \frac{(\ln\ln n)^{\frac{x}{5}}}{n^{1-\frac{x}{5}}} \sum_{k=3}^{n} \left(\frac{k}{\ln\ln k}\right)^{-\frac{x}{5}} \ (n \in \mathbb{N} \text{ fixé })$$

et

$$\texttt{Beta}(x,n) = beta(x,k), k = 3,\ldots,n.$$

Calculer et représenter graphiquement sur la même fenêtre graphique les valeurs :

$$\texttt{Beta}(-1,100),\ \texttt{Beta}(0,100), \texttt{Beta}(0.5,100) \text{ et } \texttt{Beta}(1,100),$$

sur l'intervalle $[1:100]$. Mettre un titre sur chaque graphique.

— Écrire une fonction appelée `MoyenneMobile5`, qui prend en argument un vecteur x de taille supérieur à 5 et qui calcule les moyennes empiriques de ses composantes successives considérées par groupes de 5. Appliquer la à une suite de 50 nombres aléatoires simulés suivant une loi normale centrée réduite.

6 Étude de variables quantitatives discrètes

On appelle variable quantitative discrète une variable ne prenant que des valeurs entières.

a. Les données

Nous considérerons comme données un échantillon aléatoire issu d'une loi de Poisson de paramètre $\lambda = 30$. La taille de l'échantillon est 100. Ce échantillon étant aléatoire, vous n'aurez pas le même que celui de votre voisin...
Créer le code suivant :

```
> x <- rpois(n=100, lambda=30); x
```

Une première approche consiste à classer la série statistique brute par valeurs croissantes. On obtient ainsi la *série statistique ordonnée*. Il apparaît souvent que certaines valeurs de la série se répètent. c'est en se basant sur ces occurences que le

tableau de representation appelé *tableau statistique*, et la "representation tige-et-feuille" sont construits.

Commenter les commandes suivantes et leurs sorties :

```
> n <- length(x)
> xl <- sort(unique(x))
> nl <- table(x)
> fl <- nl/n
> r <- length(xl)
> Nl=cumsum(nl); Nl
> Fl=cumsum(fl); Fl
> tabx <- cbind(xl,nl,Nl,fl,Fl)
> dimnames(tabx) <- list(1:r,c("xl","nl","Nl","fl","Fl"))
> tabx
> tabx <- as.data.frame(tabx)
> stem(x)
```

b. Représentations graphiques

Les informations recueillies dans le tableau statistique peuvent être représentées sur un graphique pour avoir une meilleure vue d'ensemble des données. Les deux principaux graphiques pour une variable quantitative discrète sont le diagramme en bâton, basé sur les effectifs (ou les fréquences ou les pourcentages de fréquences) et le diagramme cumulatif, basé sur les effectifs cumulés (ou les fréquences cumulées ou les pourcentages de fréquences cumulées).

Commenter les commandes suivantes du caractère et leurs sorties

```
> plot(nl,type='h')
> barplot(nl,main="Diagramme en colonnes")
> barplot(fl, main="Diagramme des fréquences")
```

c. Caractéristiques numériques

Les caractéristiques de tendance centrale Les caractéristiques numériques de tendance centrale, dites aussi de position ou de localisation, ont pour objectif de fournir un ordre de grandeur de la série statistique, d'en situer le centre, le milieu.

Commenter les commandes suivantes et leurs sorties :

```
> sum(x)/n
> mean(x)
> min(xl[Fl>=0.5])
> median(x)
```

Les caractéristiques de dispersion Les caractéristiques de dispersion servent à préciser la variabilité de la série, *i.e.* à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

Commenter les commandes suivantes et leurs sorties :

```
> min(x)
> max(x)
> range(x)
> diff(range(x))
> sum((x-mean(x))^2)/n
> var(x)
> var(x)*(n-1)/n
> sd(x)
> summary(x)
> quantile(x)
> quantile(x,0.75)-quantile(x,0.25)
> sum(abs(x-mean(x)))/n
> sum(abs(x-median(x)))/n
> boxplot(x,range=0)
```

d. Autour de la loi de poisson de paramètre λ=30.

Une variable aléatoire X suit une loi de poisson de paramètre λ si :

$$\Omega(X) = \{1, 2, 3 \ldots\} \text{ et } \forall k \in \Omega(X), P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

1. calculer l'espérance, la variance et l'écart-type de X.
2. pour $\lambda = 30$, comparer ces valeurs, respectivement, avec `mean(x)`, `var(x)` et`sd(x)` (trouvés dans la section c.)
3. comparer l'espérance et la variance avec `sum(x)/n`, `sum`$((\texttt{x-mean(x)})^2/\texttt{n}$.

e. Loi des grands nombres

1. L'idée est de représenter empiriquement la loi des grands nombres. Commencer par simuler un échantillon de taille $n = 200$ de n'importe quelle loi (normale, Poisson, exponentielle...)
2. Pour $k = 1, ..., 200$, calculer la moyenne partielle

$$x_k = \frac{1}{k} \sum_{i=1}^{k} x_i$$

et tracer la courbe $(k, x_k)_k$.

3. Recommencer l'opération 100 fois en traçant les courbes sur le même graphique.

f. Théorème Centrale Limite

1. Simuler un échantillon de taille de 1000 moyennes prises sur k réalisations d'une loi de Poisson de paramètres λ.
2. Transformer l'échantillon en

$$\sqrt{k}\frac{x_i - \lambda}{\sqrt{\lambda}} \quad i = 1, ..., 1000$$

3. tracer l'histogramme (en utilisant la commande `hist(...,probability=T)`) et superposer la densité d'une loi gaussienne centrée réduite.
4. Recommencer pour différentes valeurs de k et d'autres distributions en soustrayant l'espérance et en divisant par la variance.

X TD 1

ISFA- 2ème année (M1)
Statistique inférentielle

Fiche de TD n^o 1

Exercice 1.

Dans le cadre d'un modèle paramétrique réel d'échantillonnage, on considère un n-échantillon $X = (X_1, \ldots, X_n)$, de loi de probabilité P_θ.

1. Pour un estimateur T_n de θ donné, montrer les propriétés suivantes :
   a. $R(T_n, \theta) = B^2(T_n) + \mathrm{Var}(T_n)$.
   b. Si T_n est un estimateur sans biais de θ et $\mathrm{Var}(T_n) \to 0$ quand $n \to 0$, alors T_n est un estimateur convergent pour θ.
2. Si P_θ est une loi de poisson de paramètre $\theta > 0$, écrire le modèle statistique correspondant. On considère l'estimateur de θ :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

   N.B : cette notation de la moyenne empirique sera utilisée dans l'ensemble du TD.

   Calculer la moyenne et la variance de $\overline{X}$. Cet estimateur est-il sans biais ? Convergent ?
3. Si P_θ est une loi de Bernoulli de paramètre $\theta \in [0, 1]$, écrire le modèle statistique, puis vérifier que l'estimateur :

$$T_n = \overline{X}\left(1 - \overline{X}\right)$$

   de $\theta(1-\theta)$ est biaisé. Donner un estimateur sans biais de $\theta(1-\theta)$.

Exercice 2.

Soit $X = (X_1, \ldots, X_n)$, un n-échantillon de loi uniforme sur $[0, \theta]$. Écrire le modèle statistique. On considère l'estimateur de θ :

$$X_{\max} := \max_{1 \le i \le n} X_i.$$

1. Déterminer la loi de $X_{\max}$ et calculer son espérance et sa variance.
2. Comparer cet estimateur avec l'estimateur sans biais de θ construit avec $\overline{X}$.

Exercice 3.

Soit un modèle paramétrique réel d'échantillonnage $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, (P_\theta, \theta \in \mathbb{R}))^n$, tel que la loi de probabilité P_θ admette pour densité :

$$f(x, \theta) = e^{-(x-\theta)} \mathbf{1}_{x \geq \theta}, \text{ avec } \theta \in \mathbb{R}.$$

Soit le vecteur aléatoire $X = (X_1, \ldots, X_n)$, associé à ce modèle.

1. Vérifier que $U_1 = X_1 - \theta$ suit sous P_θ une loi exponentielle de paramètre 1.
2. Comparer les estimateurs :

$$Y_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - 1) \text{ et } X_{\min} = \min_{1 \leq i \leq n} X_i.$$

Exercice 4.

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi normale $\mathcal{N}(m, \sigma^2)$. Pour estimer σ^2, on pose :

$$T_n = c(n) \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

1. Quelle est la loi de $\sigma^{-2} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$?
2. calculer $B(T_n)$, et donner un estimateur sans biais de σ^2.
3. Quelle est la fonction $c(n)$ qui minimise le risque quadratique de T_n ?
4. On suppose que $\sigma = 1$ et $m \in [0, 1]$, et on définit l'estimateur :

$$U_n = \begin{cases} 0 & \text{si } \overline{X} < 0 \\ \overline{X} & \text{si } 0 \leq \overline{X} \leq 1 \\ 1 & \text{si } \overline{X} > 1 \end{cases}.$$

   a. Montrer que U_n est un estimateur de m, strictement meilleur que $\overline{X}$, en calculant la différence entre les risques quadratiques.

   b. Montrer que

$$\mathrm{E}(U_n) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{n}\theta}^{\sqrt{n}(1-\theta)} \left(\theta + \frac{t}{\sqrt{n}}\right) e^{-\frac{t^2}{2}} dt + \mathrm{P}\left(\overline{X} > 1\right).$$

   c. En déduire que $E_m(U_n) \to \theta$ lorsque $n \to \infty$.

Exercice 5.

Afin de diminuer les sinistres de ses clients, un assureur décide de financer la construction d'une digue destinée à empêcher les inondations provoquées par les crues d'une rivière. La construction d'une digue de hauteur h coûtera $c_1 h$ à la compagnie d'assurance. En cas de crue, il n'y aura aucun sinistre si la hauteur H de la crue est inférieure à h et un sinistre évalué à $c_2(H - h)$ si $H > h$. La hauteur de la crue H est une variable aléatoire de loi exponentielle de paramètre $1/\theta$. On suppose que $c_2 > c_1 > 0$.

1. On choisit comme fonction de perte :

$$L(\theta, h) = c_1 h + c_2 \mathrm{E}_\theta \left[(H - h) \, \mathbf{1}_{H>h} \right].$$

   Calculer $L(\theta, h)$ si θ est connu, quelle hauteur choisir ?

2. Ayant observé n crues indépendantes de hauteurs $H_1, \ldots, H_n$, l'assureur qui ne connaît pas θ, décide de fixer la hauteur de la digue à $d_n = k\overline{H}$

   a. Évaluer la limite (quand $n \to 0$) du risque de cette décision.

   b. Quelle valeur donneriez-vous à k ?

Exercice 6.

Soient $X = (X_1, \ldots, X_n)$ un n-échantillon de loi de bernoulli de paramètre $\theta \in [0, 1]$, a et b deux constantes positives. On considère les estimateurs de θ :

$$T_{a,b} = \frac{n\overline{X} + a}{n + a + b}.$$

1. Calculer $R(T_{a,b}, \theta)$.
2. Comparer à l'aide des risques quadratiques, les estimateurs $T_{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}$ et $T_{0,0}$ selon les valeurs de θ. Le critère du risque quadratique est-il approprié pour comparer ces estimateurs ?
3. Quel est parmi ces deux estimateurs celui qui minimise le maximum du risque quadratique ? *Cet estimateur est dit meilleur au sens mini-max.*
4. On suppose maintenant que θ est une variable aléatoire suivant une loi uniforme sur $[0, 1]$. Calculer r_1 et r_2 définis par :

$$r_1 = \mathrm{E}\left(R(T_{0,0}, \theta)\right) \quad \text{et} \quad r_2 = \mathrm{E}\left(R(T_{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}, \theta)\right).$$

XI TD 2

ISFA- 2ème année (M1)
Statistique inférentielle

Fiche de TD n^o 2

Exercice 1.

On considère un modèle d'échantillonnage avec $n > 1$ et pour lequel P_λ est une loi de poisson de paramètre $\lambda > 0$. On a donc les X_i iid avec $\forall i = 1, \ldots, n, X_i \sim$ Poisson(λ). On pose :

$$T_n = \sum_{i=1}^{n} X_i \text{ et } T'_n = X_1.$$

1. Montrer que la statistique T_n est exhaustive pour le paramètre λ et que T'_n ne l'est pas.
2. Soit $k \in \mathbb{N}$. On veut estimer $p(\lambda) = \mathrm{P}_\lambda(X_i = k)$. Calculer la moyenne et la variance de N_k/n où :
$$N_k = \sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{X_i=k}\}}$$
est le nombre de X_i égaux à k.
3. Calculer $E\left(\frac{N_k}{n}/T_n\right)$. C'est l'unique estimateur sans biais de $p(\lambda)$ fonction de N_k/n.

Exercice 2. Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi P_θ. Donner la vraisemblance du modèle, vérifier sa régularité et calculer l'information de Fisher dans le cas où :

1. P_θ est une loi de poisson de paramètre θ.
2. P_θ est une loi de Pareto de paramètre $\alpha > 1$ et $\beta > 0$, fixé de densité
$$f(x, \theta, \beta) = \frac{\theta - 1}{\beta} \left(\frac{\beta}{x}\right)^{\theta} \mathbf{1}_{x \geq \beta}$$
3. P_θ est une loi exponentielle de paramètre $\theta > 0$
4. P_θ est une loi uniforme sur $[0, \theta], \theta > 0$.

Exercice 3.

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi P_θ.

1. On suppose que P_θ est une loi de Poisson de paramètre $|\theta|$ si $-1 \leq \theta < 0$ et P_θ est une loi de Bernoulli de paramètre θ si $0 \leq \theta \leq 1$. Calculer la loi conditionnelle de X sachant $n\overline{X}$. En déduire que $n\overline{X}$ n'est pas exhaustive.

2. Vérifier que le modèle est dominé et en appliquant le théorème de factorisation, donner une statistique exhaustive pour $\theta \in [0,1]$.
3. On suppose maintenant que P_θ est une loi uniforme discrète sur $\{1,2,\ldots,\theta\}$, où $\theta \in \mathbb{N}^*$.
   a. Vérifier que le modèle statistique est dominé et donner une vraisemblance du modèle.
   b. Déterminer l'estimateur du maximum de vraisemblance de θ.

Exercice 4.

L'organisateur d'une exposition s'intéresse au rythme d'arrivées de groupes de visiteurs à partir des observations faites au cours des premières journées. Il constate que le temps séparant l'arrivée de deux groupes successifs peut être assimilé à une variable X de loi uniforme sur $[0,\theta]$ et que ces temps inter-arrivées sont indépendantes Il souhaite estimer θ à partir de l'observation d'un n-échantillon $X = (X_1,\ldots,X_n)$ de ces inter-arrivées.

1. Vérifier que le modèle statistique est dominé et donner une vraisemblance du modèle.
2. Déterminer l'estimateur du maximum de vraisemblance T_n de θ et en déduire sa loi.
3. Déterminer β tel que $W_n = \beta T_n$ soit un estimateur sans biais de θ.

Exercice 5.

Soit $X = (X_1,\ldots,X_n)$ un n-échantillon de loi P_θ.

1. On suppose que P_θ est une loi de Poisson de paramètre θ. On veut estimer θ^2.
   a. Vérifier que le modèle statistique est dominé et donner une vraisemblance du modèle.
   b. Déterminer l'estimateur du maximum de vraisemblance de θ^2. Est-il sans biais ?
2. On suppose que P_θ est une loi exponentielle de paramètre $1/\theta$.
   a. Vérifier que le modèle statistique est dominé et donner une vraisemblance du modèle.
   b. Déterminer l'estimateur du maximum de vraisemblance de θ.

Exercice 6.

Soit $X = (X_1,\ldots,X_n)$ un n-échantillon de loi $\mathcal{N}(0,\theta), \theta > 0$.

1. On veut estimer θ par la méthode du maximum de vraisemblance. On pose

$$T_n(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

a. Soit $p(\theta, x_1, \ldots, x_n)$ la fonction de vraisemblance. Calculer :

$$2\left(\ln p(\theta, x_1, \ldots, x_n) - \ln p(T_n(x_1, \ldots, x_n), x_1, \ldots, x_n)\right).$$

b. En utilisant l'inégalité $\ln x \leq x - 1$, montrer que T_n est l'estimateur du maximum de vraisemblance de θ.

c. Montrer que T_n est sans biais et calculer $R(T_n, \theta)$.

2. Calculer l'information de Fisher pour l'estimation de θ
3. L'estimateur T_n est-il efficace de θ ?
4. On veut maintenant estimer θ par la méthode de Bayes relatif à la loi à priori ayant pour densité :

$$g(\theta) = \frac{\lambda^a}{\Gamma(a)} \frac{1}{\theta^{a+1}} \exp\left(-\frac{\lambda}{\theta}\right) \mathbf{1}_{\{\theta>0\}},$$

où $\lambda > 0, a > 1$ et $\Gamma(a+1) = a\Gamma(a)$. Les estimateurs de Bays sont définis par :

$$T_{\lambda,a} = \frac{nT_n + 2\lambda}{2a + n - 2}.$$

Montrer qu'ils sont asymptotiquement sans biais.

5. On veut maintenant comparer l'estimateur du maximum de vraisemblance avec les estimateurs bayésiens.

a. Peut-on les comparer au sens minimax ?

b. Montrer que la limite de $R(T_{\lambda,a}, \theta) - R(T_n, \theta)$ lorsque $\theta \to 0$ est strictement positive.

c. Pour tous λ et a trouver un point $\theta^*_{\lambda,a}$ où $R(T_{\lambda,a}, \theta^*_{\lambda,a}) < R(T_n, \theta^*_{\lambda,a})$.

d. Conclure.

Exercice 7.

1. Montrer que si T_n est un estimateur efficace de θ, alors $kT_n + b$ est aussi un estimateur efficace de $\theta, \forall k \in \mathbb{R}^*, \forall b \in \mathbb{R}$. On considère la loi normale $\mathcal{N}(\mu, \sigma^2)$.
2. On suppose σ^2 connue et l'on considère le modèle paramétrique réel d'échantillonnage suivant :

$$\left(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mathcal{N}\left(\mu, \sigma^2\right), \mu \in \Theta = \mathbb{R}\right)^n.$$

L'estimateur $\overline{X}$ est-il un estimateur efficace de μ ?

3. Sans supposer μ connue, on pose $\theta = \sigma^2$ et l'on considère le modèle paramétrique réel d'échantillonnage suivant :

$$\left(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mathcal{N}\left(\mu, \sigma^2\right), \sigma^2 \in \Theta = \mathbb{R}_+^*\right)^n.$$

L'estimateur $S^2 = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ est-il un estimateur efficace de σ^2 ?

4. Écrire la vraisemblance du modèle. Vérifier que le couple $(\overline{X}, S^2)$ forment une statistique exhaustive pour le paramètre (μ, σ^2).

XII TD 3

ISFA- M1SAFIR
Statistique inférentielle

Fiche de TD n^o 3

Exercice 1.

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi uniforme sur $[0, \theta]$.

1. Écrire la vraisemblance du modèle. Donner une statistique exhaustive pour θ.
2. Donner un estimateur sans biais de θ fonction de la statistique exhaustive.
3. En déduire sans faire de calcul, la valeur de $\mathrm{E}\left(\overline{X}/\max_{1\leq i\leq n} X_i\right)$.

Exercice 2.

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi exponentielle décentrée sur $[\theta, +\infty[$.

1. Déterminer l'estimateur du maximum de vraisemblance de θ. Est-il biaisé ? Est-ce une statistique exhaustive ?
2. En déduire un estimateur sans biais de θ fonction d'une statistique exhaustive.

Exercice 3.

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi uniforme sur $[\theta - \frac{1}{2}, \theta + \frac{1}{2}], \theta \in \mathbb{R}$. On pose :

$$X_{\max} = \max_{1\leq i\leq n} X_i \text{ et } X_{\min} = \min_{1\leq i\leq n} X_i.$$

1. Écrire la vraisemblance du modèle et montrer que le couple $(X_{\min}, X_{\max})$ est exhaustif.
2. Calculer la loi de probabilité de $X_{\max} - X_{\min}$.

Exercice 4.

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi F_θ de densité :

$$f(x, \theta) = (k+1)\frac{x^k}{\theta^{k+1}}\mathbf{1}_{[0,\theta]}(x),$$

où $k \geq 1$ est connu et $\theta > 0$ est le paramètre inconnu.

1. Calculer $\mathrm{E}_\theta(X_1)$.
2. Donner une statistique exhaustive pour θ.
3. On considère l'estimateur

$$S = \frac{n(k+1)+1}{n(k+1)} \max(X_1, \ldots, X_n).$$

Quelles sont ses propriétés ? En déduire sans calcul $\mathrm{E}_\theta(X_1/\max_{1\leq i\leq n} X_i)$.

Exercice 5.

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi F_θ de densité :

$$f(x, \theta) = (\theta + 1)x^\theta \mathbf{1}_{[0,1]}(x),$$

où $\theta > 0$ est le paramètre inconnu.

1. Écrire la vraisemblance du modèle Donner une statistique exhaustive.
2. Quelle est la loi de $-\ln X_1$ et de $-(\theta + 1)\ln X_1$.
3. Donner un estimateur sans biais de $-1/(\theta + 1)$. On pose :

$$S_n = \frac{\max_{1 \leq i \leq n} \ln X_i}{\min_{1 \leq i \leq n} \ln X_i}.$$

Montrer que la loi de S_n ne dépend pas de θ. En déduire que S_n et $\prod_{i=1}^n X_i$ sont indépendantes.