

TD 2 Statistiques et estimation

Exercice 1 :

On simule $M = 5000$ réalisations d'un échantillon (X_1, \dots, X_{15}) ($n=15$) de la loi $\mathcal{N}(-10, 1.5^2)$. On sauvegarde 5000 réalisations de la statistique \bar{X} ainsi que celles de S_c^2 .

a) Calculer la moyenne et la variance corrigée de la statistique \bar{X} basé sur ces 5000 réalisations. Tracer l'histogramme \bar{X} sur ces 5000 réalisations. Commenter vos résultats.

b) Calculer la proportion observée, basée sur ces 5000 réalisations, de

i) l'évènement $A = \{-10.5 \leq \bar{X} \leq -9.5\}$;

ii) l'évènement $B = \{2 \leq S_c^2 \leq 4\}$;

iii) l'évènement $C = \{-10.5 \leq \bar{X} \leq -9.5 \text{ et } 2 \leq S_c^2 \leq 4\}$.

c) Que valent les probabilités correspondantes $\mathbb{P}(A)$, $\mathbb{P}(B)$, $\mathbb{P}(C)$ et $\mathbb{P}(A \cap B)$? Commenter vos résultats.

d) Calculer le coefficient de corrélation entre les statistiques \bar{X} et S_c^2 basé sur ces 5000 réalisations. Que constatez-vous?

e) On définit la variable aléatoire $T = \frac{\sqrt{n}(\bar{X} + 10)}{S_c}$. Tracer l'histogramme T basé sur ces 5000 réalisations. Commenter votre résultat.

f) On définit la variable aléatoire $K^2 = \frac{(n-1)S_c^2}{1.5^2}$. Tracer l'histogramme de la statistique T basé sur ces 5000 réalisations.. Commenter votre résultat.

Exercice 2 :

On simule $M = 5000$ réalisations d'un échantillon (X_1, \dots, X_{15}) ($n=15$) de la loi exponentielle $\mathcal{E}(\lambda = 1)$. On sauvegarde 5000 réalisations de la statistique \bar{X} ainsi que celles de S_c^2 .

a) Calculer la moyenne et la variance corrigée de la statistique \bar{X} basé sur ces 5000 réalisations. Tracer l'histogramme \bar{X} sur ces 5000 réalisations. Commenter vos résultats.

b) Calculer la proportion observée, basée sur ces 5000 réalisations, de

- i) l'évènement $A = \{0.8 \leq \bar{X} \leq 1.5\}$;
- ii) l'évènement $B = \{1 \leq S_c^2 \leq 3\}$;
- iii) l'évènement $C = \{0.8 \leq \bar{X} \leq 1.5 \text{ et } 1 \leq S_c^2 \leq 3\}$.

c) Commenter vos résultats.

d) Calculer le coefficient de corrélation entre les statistiques \bar{X} et S_c^2 basé sur ces 5000 réalisations. Que constatez-vous ?

e) On définit la variable aléatoire $T = \frac{\sqrt{n}(\bar{X} - 1)}{S_c}$. Tracer l'histogramme T basé sur ces 5000 réalisations. Commenter votre résultat.

f) On définit la variable aléatoire $K^2 = \frac{(n-1)S_c^2}{1^2}$. Tracer l'histogramme de la statistique T basé sur ces 5000 réalisations.. Commenter votre résultat.

Exercice 3 :

Simuler une réalisation des $n = 5000$ variables aléatoires $(X_j, j = 1, \dots, 5000)$ indépendantes de même loi uniforme $U(0, 2)$. Utilisant "cumsum" pour tracer n points dont les coordonnées sont $(k, \frac{\sum_{j=1}^k x_j}{k})$, $k = 1, \dots, 5000$, ainsi avoir une illustration de la loi des grands nombres.

Exercice 4 :

On utilise une boucle pour simuler $M = 3000$ réalisations d'un échantillon (X_1, \dots, X_{15}) ($n=15$) de la loi de Poisson avec paramètre $\lambda = 5$ dont la loi est donnée par

$$\mathbb{P}_\lambda(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda).$$

a) Déterminer l'estimateur du Maximum de Vraisemblance de λ .

b) On définit trois estimateurs de λ :

$$\hat{\lambda}_1 = \bar{X},$$

$$\hat{\lambda}_2 = S_c^2,$$

$$\hat{\lambda}_3 = \text{médiane}(X).$$

Evaluer l'espérance et la variance de ces trois estimateurs. Quelles sont les qualités et défauts de chaque estimateur ?

Exercice 5 :

Soit (x_1, \dots, x_n) une réalisation d'un échantillon (X_1, \dots, X_n) de loi exponentielle avec paramètre $\theta > 0$ dont la densité est

$$f(x|\theta) = \frac{1}{\theta} \exp(-x/\theta) \mathbf{1}_{\{x>0\}}.$$

Déterminer l'estimateur du Maximum de Vraisemblance de θ .

Exercice 6 :

Soit (x_1, \dots, x_n) une réalisation d'un échantillon (X_1, \dots, X_n) de loi gaussienne $N(\mu, \sigma^2)$.

a) Si on note $v = \sigma^2$, déterminer les estimateurs du Maximum de Vraisemblance de μ et de v .

b) On définit trois estimateurs de $v = \sigma^2$:

$$\hat{v}_1 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})^2,$$

$$\hat{v}_2 = S_c^2,$$

$$\hat{v}_3 = \frac{\sqrt{\sum_{j=1}^n (X_j - \bar{X})^4}}{3n}.$$

On simule $M = 5000$ réalisations d'un échantillon (X_1, \dots, X_{15}) ($n=15$) de la $N(8, 2^2)$. Evaluer l'espérance et la variance de ces trois estimateurs. Quelles sont les qualités et défauts de chaque estimateur ?

Exercice 7 :

On utilise une boucle pour simuler $M = 3000$ réalisations d'un échantillon (X_1, \dots, X_{15}) ($n=15$) de la loi exponentielle avec paramètre $\lambda = 3$ dont la densité est

$$f(x|\lambda) = \lambda \exp(-\lambda x) \mathbf{1}_{\{x>0\}}.$$

On définit quatre estimateurs de λ :

$$\hat{\lambda}_1 = 1/\bar{X},$$

$$\hat{\lambda}_2 = (n-1)/(n\bar{X}),$$

$$\hat{\lambda}_3 = n/((n-1)\bar{X}),$$

$$\hat{\lambda}_4 = 1/\text{médiane}X.$$

Evaluer l'espérance et la variance de ces trois estimateurs. Quelles sont les qualités et défauts de chaque estimateur ?

Exercice 8 :

Soit (x_1, \dots, x_{10}) ($n=10$) une réalisation d'un échantillon (X_1, \dots, X_{10}) de loi uniforme avec paramètre $\theta = 2$ dont la densité est

$$f(x|\theta) = \frac{1}{\theta} \mathbf{1}_{\{0 \leq x \leq \theta\}}.$$

a) Déterminer l'estimateur du MV de θ .

b) On définit quatre estimateurs de λ :

$$\hat{\lambda}_1 = \max_{\{1 \leq j \leq n\}} X_j$$

$$\hat{\lambda}_2 = \frac{n}{n+1} \max_{\{1 \leq j \leq n\}} X_j,$$

$$\hat{\lambda}_2 = \frac{n+1}{n} \max_{\{1 \leq j \leq n\}} X_j,$$

$$\hat{\lambda}_3 = 2\bar{X}.$$

Evaluer l'espérance et la variance de ces trois estimateurs. Quelles sont les qualités et défauts de chaque estimateur ?

Exercice 9 Estimateur sans biais et de variance minimale

Soit (X_1, \dots, X_n) d'un échantillon de loi $N(\mu, \sigma^2)$ avec $\sigma^2 = 4\mu^2$. On souhaite déterminer parmi les 4 estimateurs suivants de μ celui dont l'EQM atteint le minimum (ou la variance atteint le minimum). Les 4 estimateurs sont les suivants :

$$T_1 = (X_1 + X_n), \quad T_2 = \text{médiane}_{1 \leq i \leq n}(X_i), \quad T_3 = \frac{1}{n} \sum_{i=1}^n X_i, \quad T_4 = \frac{S_{n,c}}{2C}$$

$$\text{où } S_{n,c} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}, \quad C = \sqrt{\frac{2}{49} \frac{\Gamma(25)}{\Gamma(24.5)}} = 0,9949113.$$

a) Que représentent la moyenne théorique et la moyenne d'échantillon ? Quelles sont leurs natures ? (ex : paramètre inconnu ou constante connue ou variable aléatoire).

b) Générer $M = 3000$ réalisations de l'échantillon (X_1, \dots, X_n) de taille $n = 20$, de loi $N(\mu, 4\mu^2)$ avec $\mu = 5$ en utilisant `rnorm(20, mean=5, sd=10)`.

c) En déduire 3000 réalisations des estimateurs T_1, \dots, T_4 . Tracer les histogrammes de T_3 et T_4 sur la même sortie graphique.

d) Pour ces 4 estimateurs, que pouvez-vous dire sur leurs biais et leurs variances ? Basé sur ces 3000 réalisations, calculer leurs moyennes et variances observées. Sont-ils sans biais ? Lequel a la plus petite variance ?

e) En fait, les estimateurs T_3 et T_4 sont sans biais, et leurs variances théoriques sont 2 et 0,25 respectivement. Sachant que T_3 et T_4 sont indépendants, déterminer la constante α pour que $T_\alpha = \alpha T_3 + (1 - \alpha) T_4$ ait la plus petite variance.

Exercice 10 Estimation d'une proportion θ :

Soit (X_1, \dots, X_n) un échantillon de taille $n = 5$ de loi de Bernoulli $B(1, \theta)$ avec

$$\mathbf{P}(X_i = 1) = 1 - \mathbf{P}(X_i = 0) = \theta, \quad \theta \in [0, 15; 0, 85]$$

étant le paramètre inconnu. On souhaite comparer les deux estimateurs suivants du paramètre θ :

$$T_1 = \frac{1}{5} \sum_{i=1}^5 X_i, \quad T_2 = 0,7 \times T_1 + 0,3 \times \frac{1}{2}.$$

On rappelle que l'EQM d'un estimateur quelconque T du paramètre θ est défini par

$$\text{EQM}(T) = \left(\mathbf{E}(T) - \theta \right)^2 + \mathbf{E}(T - \mathbf{E}(T))^2. \quad (0)$$

On souhaite évaluer puis comparer $f_i(\theta) = \text{EQM}(T_i)$, $i = 1, 2$ dans la suite.

a) Montrer les résultats suivants :

$$\begin{aligned} \mathbf{E}(X_i) &= \theta; & \text{Var}(X_i) &= \theta(1 - \theta) \\ \mathbf{E}(\sum_{i=1}^k X_i) &= k\theta & \text{Var}(\sum_{i=1}^k X_i) &= k\theta(1 - \theta) \end{aligned}$$

b) Que représentent respectivement les deux termes du côté droit de la définition (0) ci-dessus ?

c) Montrer que

$$f_1(\theta) = \frac{\theta(1 - \theta)}{5}.$$

d) Tracer le graphique de $f_1(\theta)$, comme suit :

```
x = seq(0.15, 0.85, length = 100) # représentant  $\theta \in [0.15, 0.85]$ 
f1 = fonction(x){ ... } # vous devez coder l'expression de  $f_1$ 
y = f1(x) # définissant  $\theta \rightarrow f_1(\theta)$ 
plot(x, y, type = "l", col = "blue", ylim = c(0.01, 0.085)) # tracer le graphique
```

e) Montrer que

$$f_2(\theta) = 0.7^2 \frac{\theta(1 - \theta)}{5} + 0.3^2 \left(\frac{1}{2} - \theta \right)^2.$$

5) Superposer le graphique de $f_2(\theta)$ sur le premier graphique comme suit :

```
f2 = fonction(x){ ... } # coder l'expression de  $f_2$ 
y2 = f2(x) # définissant  $\theta \rightarrow f_2(\theta)$ 
lines(x, y2, col = "red", lwd = 2) # Superposition
```

f**) En étudiant la variation de la fonction

$$D(\theta) = f_1(\theta) - f_2(\theta) = \frac{(1 - 0.7^2)}{5} \theta(1 - \theta) - 0.3^2 \left(\frac{1}{2} - \theta \right)^2,$$

montrer que

$$f_1(\theta) > f_2(\theta) \quad \forall \theta \in [0.15; 0.85].$$

7) Quel est le meilleur estimateur de θ parmi les trois selon vous ? Pourquoi ?