Université de Strasbourg
UFR de Mathématiques et d'Informatique
Han-Ping LI

Année 2021/2022
Statistique
Etude de cas L3

TD 2 Statistiques et estimations

Une statistique est une fonction d'un échantillon, ne dépendant pas de paramètres inconnus.

Exercice 1 :

On souhaite étudier les propriétés de deux statistiques $\overline{X}$ et S_c^2 d'un échantillon gaussien. Pour ce faire, on utilise une matrice via une boucle pour enregistrer $M = 5000$ réalisations d'un échantillon $(X_1, \ldots, X_{15})$ (n =15) de la loi $\mathcal{N}(-10, 1.5^2)$. On sauvegarde 5000 réalisations de la statistique $\overline{X}$ ainsi que celles de S_c^2.

a) Calculer la moyenne et la variance corrigée de la statistique $\overline{X}$ basé sur ces 5000 réalisations. Tracer l'histogramme $\overline{X}$ sur ces 5000 réalisations. Superposer la courbe de la densité normale $\mathcal{N}(-10, 1.5^2/n)$. Commenter vos résultats.

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

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

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

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

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 $\overline{X}$ et S_c^2 basé sur ces 5000 réalisations. Que constatez-vous ?

e) On définit la variable aléatoire $T_s = \dfrac{\sqrt{n}(\overline{X} + 10)}{S_c}$. Tracer l'histogramme T_s basé sur ces 5000 réalisations. Commenter votre résultat.

f) On définit la variable aléatoire $K^2 = \dfrac{(n-1)S_c^2}{1.5^2}$. Tracer l'histogramme de la statistique K^2 basé sur ces 5000 réalisations.. Commenter votre résultat.

Exercice 2 :

On souhaite vérifier si les propriétés de deux statistiques $\overline{X}$ et S_c^2 étudiées dans l'exercice 1 sont valables pour un échantillon non-gaussien. Pour ce faire, on simule $M = 5000$

réalisations d'un échantillon $(X_1, \dots, X_{15})$ (n =15) de la loi exponentielle $\mathcal{E}(\lambda = 1.5)$. On sauvegarde 5000 réalisations de la statistique $\overline{X}$ ainsi que celles de S_c^2.

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

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

i) l'évènement $A = \{0.8 \leq \overline{X} \leq 2\}$;

ii) l'évènement $B = \{1 \leq S_c^2 \leq 3\}$;

iii) l'évènement $C = \{0.8 \leq \overline{X} \leq 2 \text{ et } 1 \leq S_c^2 \leq 3\}$.

c) Commenter vos résultats.

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

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

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

g) Une réalisation $(x_1, \dots, x_{15})$ de l'échantillon $(X_1, \dots, X_{15})$ peut être interprétée comme 15 réalisations de X de la loi exponentielle $\mathcal{E}(\lambda = 1.5)$, ainsi N réalisations de l'échantillon $(X_1, \dots, X_{15})$ peuvent être interprétées comme $15N$ réalisations de X de la loi exponentielle $\mathcal{E}(\lambda = 1)$:

```
X=rbind(Echantillon[,1:15]).
```

Comparer la proportion observée, basée sur ces 75000 réalisations de l'évènement $\{X > 2 + 3 | X > 2\}$ avec celle de l'évènement $\{X > 3\}$. Quelle sont les probabilités correspondantes ? Commenter vos résultats.

Exercice 3 :

Simuler une réalisation des $n = 10000$ variables aléatoires $\big(X_j, j = 1, \dots, 10000\big)$ indépendantes de même loi uniforme $U(0, 2)$. Utilisant "cumsum" pour tracer n points dont les coordonnées sont $\left(k, \dfrac{\sum_{j=1}^{k} x_j}{k}\right), k = 1, \dots, 10000$; ainsi avoir une illustration de la loi des grands nombres. Combien de réalisations a-t-on observé ?

Illustrer la loi des grands nombres en traçant le graphique basé sur une autre loi de votre choix, loi de Poisson par exemple.

Exercice 4 :

On souhaite comparer 4 différents estimateurs du paramètre d'une loi de Poisson.

a) On simule $M = 10000$ réalisations d'un échantillon $(X_1, \dots, X_{10})$ $(n = 10)$ de la loi de Poisson avec paramètre $\lambda = 4.1$ dont la loi est donnée par

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

On n'est pas censé connaître la vraie valeur de $\lambda = 5$.

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

c) On définit quatre estimateurs de λ :

$\widehat{\lambda}_1 = \overline{X}$,

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

$\widehat{\lambda}_3 = \text{médiane}(X)$,

$\widehat{\lambda}_4 = 0.9\overline{X} + 0.1S_c^2$.

Évaluer l'espérance et la variance de ces quatre estimateurs grâce à la loi des grands nombres. 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

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

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

b) On simule $M = 10000$ réalisations d'un échantillon $(X_1, \dots, X_n)$ $(n = 20)$ de la loi exponentielle avec paramètre $\theta = 4$. On définit quatre estimateurs de λ :

$\widehat{\lambda}_1 = \overline{X}$,

$\widehat{\lambda}_2 = (1 + 1/n)\sqrt{\frac{\sum_{i=1}^n X_i^2}{2n}}$,

$\widehat{\lambda}_3 = (1 + 1/n)S_c$.

$\widehat{\lambda}_4 = \frac{\text{médiane}(X)}{\ln(2)(1 + 1/n)}$,

Évaluer l'espérance et la variance de ces quatre estimateurs grâce à la loi des grands nombres. Quelles sont les qualités et défauts de chaque estimateur ?

Exercice 6 :

On souhaite comparer 4 différents estimateurs du paramètre d'une loi exponentielle. On simule d'abord $M = 10000$ réalisations d'un échantillon $(X_1, \ldots, X_n)$ (n =18) de cette 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 :

$\widehat{\lambda}_1 = 1/\overline{X}$,

$$\widehat{\lambda}_2 = \frac{(n-1)}{n\overline{X}},$$

$$\widehat{\lambda}_3 = \frac{4n}{(4n+11)S_c},$$

$$\widehat{\lambda}_4 = \frac{\ln(2)(n-4)}{(n-3)\text{médiane}(X)}.$$

Evaluer l'espérance et la variance de ces quatre estimateurs basé sur $M = 5000$ réalisations grâce à la loi des grands nombres. Quelles sont les qualités et défauts de chaque estimateur ?

Exercice 7 :

Soit $(x_1, \ldots, x_{10})$ (n =10) une réalisation d'un échantillon $(X_1, \ldots, X_{10})$ 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 θ comme suit :

$$\widehat{\theta}_1 = \max_{\{1\leq j\leq n\}} X_j$$

$$\widehat{\theta}_2 = \frac{n}{n+1} \max_{\{1\leq j\leq n\}} X_j,$$

$$\widehat{\theta}_3 = \frac{n+1}{n} \max_{\{1\leq j\leq n\}} X_j,$$

$$\widehat{\theta}_4 = 2\overline{X}.$$

Evaluer l'espérance et la variance de ces quatre estimateurs basé sur $M = 5000$ réalisations. Quels sont les qualités et défauts de chaque estimateur ?

Exercice 8 :

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 $v = \sigma^2$:

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

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

$$\widehat{v}_3 = \sqrt{\frac{\sum_{j=1}^{n}(X_j - 8)^4}{3n}}.$$

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 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 :

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

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

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 $M = 10000$ réalisations de l'échantillon $(X_1, \dots, X_n)$ de taille $n = 50$, de loi $N(\mu, 4\mu^2)$ avec $\mu = \mathbf{5}$ (**que vous n'êtes pas censés connaître**) en utilisant `rnorm(50, 5, 10)`.

c) En déduire 10000 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 $(X_1, \ldots, X_n)$ 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 marametre θ est défini par

$$\text{EQM}(T) = \Big(\mathbf{E}(T) - \theta\Big)^2 + \mathbf{E}\big(T - \mathbf{E}(T)\big)^2. \tag{0}$$

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

a) Montrer les résultats suivants :

$$\begin{aligned} E(X_i) &= \theta; \quad \mathbf{V}ar(X_i) = \theta(1-\theta) \\ \mathbf{E}(\textstyle\sum_{i=1}^k X_i) &= k\theta \quad \mathbf{V}ar(\textstyle\sum_{i=1}^k X_i) = k\theta(1-\theta) \end{aligned}$$

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 θ ∈ [0.15, 0, 85]
f1 = function(x){ ... } # vous devez coder l'expression de f1
y= f1(x) # définissant θ → f1(θ)
plot(x,y,type = "l", col="blue",ylim=c(0.01, 0.085)) # tracer le graphique
```

e) Montrer que

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

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

```
f2 = function(x){ ... } # coder l'expression de f2
y2=f2(x) # définissant θ → f2(θ)
lines( x,y2, col="red", lwd = 2) # Superposition
```

g**) En étudiant la variation de la fonction

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

montrer que

$$f_1(\theta) > f_2(\theta) \ \ \forall\ \theta \in [0,15\,;0,85].$$

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