Université Paris-Dauphine L3 CPES

# DST2 – Statistique

Polycopié et calculatrice autorisés. Tout autre document interdit.
Durée 2h
Date : 7 décembre 2022

Dans tout le sujet pour tout $\alpha \in ]0,1[$, $q_\alpha$ désigne le quantile d'ordre $\alpha$ de la loi normale $\mathcal{N}(0,1)$.
Approximations à $10^{-2}$ près : $q_{0.9} = 1.28$, $q_{0.95} = 1.64$, $q_{0.975} = 1.96$, $q_{0.99} = 2.33$, $q_{0.999} = 3.09$.

**Exercice 1 (questions de cours (/5))**

1. Nous souhaitons tester l'indépendance entre le nombre d'années d'étude et le revenu dans une population donnée. Le nombre de modalités de la variable nombre d'années d'étude est 8 et la variable revenu est regroupée en 8 classes. Donner le nombre de degrés de liberté de la loi du $\chi^2$ utilisée pour effectuer le test du $\chi^2$ d'indépendance.

   **Solution:** Le nombre de degrés de liberté est $(J-1)(K-1) = (8-1)(8-1) = 49$.

2. Expliquer en une phrase pourquoi un test qui rejette tout le temps l'hypothèse nulle $\mathcal{H}_0$ n'est pas conforme au principe de Neyman.

   **Solution:** Un tel test à un niveau de 100%, il n'est donc pas de niveau $\alpha$ pour $\alpha < 1$.

3. Soit $\widehat{\theta}$ un estimateur d'un paramètre $\theta$ tel qu'il existe $\sigma > 0$ tel que

$$\frac{\sqrt{n}}{\sigma}(\widehat{\theta} - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

   De quoi avons-nous besoin pour construire un intervalle de confiance asymptotique pour le paramètre $\theta$ ?

   **Solution:** Nous devons trouver un estimateur convergent de $\sigma$.

4. Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction. Quelles propriétés $f$ doit-elle vérifier pour être une densité de probabilité ?

   **Solution:** Elle doit être positive et vérifier $\int_{\mathbb{R}} f(t)dt = 1$.

5. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. (indépendantes et identiquement distribuées) selon la densité $f_\theta$ de paramètre $\theta$. Définir le score des observations.

**Solution:** Le score des observations est la variable aléatoire

$$\frac{\partial \ell}{\partial \theta}(\theta; X_1, \ldots, X_n),$$

où $\ell(\theta; x_1, \ldots, x_n) = \log\left(\Pi_{i=1}^n f_\theta(x_i)\right)$ est la log-vraisemblance des données.

**Exercice 2 (saturation des antenne-relais mobiles (/4))** À un instant donné, $n = 2800$ personnes utilisent leur téléphone dans le secteur d'une antenne-relais téléphonique. Parmi ces personnes-là,

- $n_1 = 2500$ personnes utilisent leur téléphone pour une communication téléphonique, avec un débit moyen de $\mu_1 = 94.8$ Kb.s$^{-1}$ (écart-type $\sigma_1 = 20$ Kb.s$^{-1}$),
- $n_2 = 330$ personnes utilisent leur téléphone pour visionner une vidéo, avec un débit moyen de $\mu_2 = 150$ Mb.s$^{-1}$ (écart-type $\sigma_2 = 6$ Mb.s$^{-1}$).

Nous souhaitons estimer la probabilité que l'antenne soit saturée, c'est-à-dire la probabilité que le débit total des utilisateurs excède la capacité de l'antenne-relais 50 000 Mb.s$^{-1}$.

On rappelle que 1 Mb.s$^{-1}$ $= 10^6$ b.s$^{-1}$ et 1 Kb.s$^{-1}$ $= 10^3$ b.s$^{-1}$.

Soient $X_1, \ldots, X_{n_1}$ le débit des personnes qui téléphonent en Mb.s$^{-1}$ et $Y_1, \ldots, Y_{n_2}$ celui des personnes qui visionnent des vidéos en Mb.s$^{-1}$. On suppose que

$$X_1, \ldots, X_{n_1} \sim_{i.i.d.} \mathcal{N}(\mu_1, \sigma_1^2)$$

$$Y_1, \ldots, Y_{n_2} \sim_{i.i.d.} \mathcal{N}(\mu_2, \sigma_2^2)$$

et que $X_1, \ldots, X_{n_1}$ est indépendant de $Y_1, \ldots, Y_{n_2}$.

1. Donner, en fonction de $\mu_1$, $n_1$, $\sigma_1^2$ et $\mu_2$, $n_2$, $\sigma_2^2$, la loi du débit total

$$D = X_1 + \ldots + X_{n_1} + Y_1 + \ldots + Y_{n_2}.$$

**Solution:** La somme de variables aléatoires normales indépendantes suit également une loi normale et nous avons

$$\begin{aligned}\mathbb{E}[D] &= \mathbb{E}[X_1 + \ldots + X_{n_1} + Y_1 + \ldots + Y_{n_2}] \\ &= \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_{n_1}] + \mathbb{E}[Y_1] + \ldots + \mathbb{E}[Y_{n_2}] \\ &= n_1\mu_1 + n_2\mu_2.\end{aligned}$$

et

$$\begin{aligned}\mathrm{Var}(D) &= \mathrm{Var}(X_1 + \ldots + X_{n_1} + Y_1 + \ldots + Y_{n_2}) \\ &= \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_{n_1}) + \mathrm{Var}(Y_1) + \ldots + \mathrm{Var}(Y_{n_2}) \\ &= n_1\sigma_1^2 + n_2\sigma_2^2.\end{aligned}$$

Donc

$$D \sim \mathcal{N}(n_1\mu_1 + n_2\mu_2, n_1\sigma_1^2 + n_2\sigma_2^2).$$

2. En déduire que

$$D \sim \mathcal{N}(\mu_D, \sigma_D^2),$$

où $\mu_D = 49\ 737$ Mb.s$^{-1}$ et $\sigma_D = 109$ Mb.s$^{-1}$.

**Solution:** On calcule $\mu_D$ et $\sigma_D$ à partir de la question précédente.

$$\mu_D = n_1\mu_1 + n_2\mu_2 = 2500 \times \frac{94.8}{1000} + 330 \times 150 = 49\ 737.$$

et

$$\sigma_D = \sqrt{n_1\sigma_1^2 + n_2\sigma_2^2} = \sqrt{2500 \times \left(\frac{20}{1000}\right)^2 + 330 \times 6^2} = 109.$$

3. Montrer que

$$\mathbb{P}(D > 50000) \leq \mathbb{P}(Z > 2.4),$$

où $Z \sim \mathcal{N}(0, 1)$.

**Solution:** D'après 2.

$$\mathbb{P}(D > 50000) = \mathbb{P}\left(\frac{D - \mu_D}{\sigma_D} > \underbrace{\frac{50000 - \mu_D}{\sigma_D}}_{\geq 2.4}\right) \leq \mathbb{P}(Z > 2.4),$$

car $\frac{D-\mu_D}{\sigma_D}$ et $Z$ suivent la même loi.

4. En déduire que la probabilité que l'antenne soit saturée est inférieure à 1%.

**Solution:** D'après la question précédente, la probabilité que l'antenne soit saturée est

$$\mathbb{P}(D > 46000) \leq \mathbb{P}(Z > 2.4) \leq \mathbb{P}(Z > q_{0.99}) = 1 - F_Z(q_{0.99}) = 1 - 0.99 = 0.01,$$

où $F_Z$ est la fonction de répartition de $Z$.

**Exercice 3 (Estimation et censure (/11))**

Une assurance souhaite modéliser la durée des contrats d'assurance vis pris par ses clients. Pour cela, l'entreprise suit $n$ clients pendant une certaine durée.

Nous modélisons la durée du contrat pris par le $i$-ème client par la variable $T_i$ et la durée d'observation du $i$-ème client par la variable $C_i$. Si $T_i \geq C_i$, nous n'observons pas $T_i$ (le contrat a pris fin après la fin de la période d'observation) de sorte que nous observons seulement

$$X_i = \min\{T_i, C_i\}.$$

Nous supposons que $T_1, \ldots, T_n$ est i.i.d. selon une loi exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda > 0$ inconnu et $C_1, \ldots, C_n$ est i.i.d. selon une loi exponentielle $\mathcal{E}(\lambda_0)$ de paramètre $\lambda_0 > 0$ connu et que $T_1, \ldots, T_n$ est indépendant de $C_1, \ldots, C_n$.

L'objectif de cet exercice est de proposer un ou deux estimateurs du paramètre $\lambda$ et de donner des intervalles de confiance asymptotique.

1. Nous commençons par déterminer la loi de $X_1, \ldots, X_n$. Soient $T \sim \mathcal{E}(\lambda)$ et $C \sim \mathcal{E}(\lambda_0)$, $T$ indépendant de $C$. Nous posons $X = \min\{T; C\}$.

   (a) Montrer que, pour tout $t > 0$, la fonction de répartition de $T$ vérifie

   $$F_T(t) = 1 - e^{-\lambda t}.$$

   **Solution:**

   $$F_T(t) = \mathbb{P}(T \leq t) = \int_0^t \lambda e^{-\lambda t} dt = \left[\frac{\lambda e^{-\lambda t}}{-\lambda}\right]_0^t = -e^{-\lambda t} + 1.$$

   (b) Montrer que, pour tout $x > 0$,

   $$\mathbb{P}(X > x) = \mathbb{P}(T > x)\mathbb{P}(C > x).$$

   **Solution:** Nous avons

   $$\mathbb{P}(X > x) = \mathbb{P}(\min\{T; C\} > x) = \mathbb{P}(T > x \text{ et } C > x) = \mathbb{P}(T > x)\mathbb{P}(C > x),$$

   par indépendance de $T$ et $C$.

(c) En déduire que la fonction de répartition de $X$, vérifie, pour tout $x > 0$,

$$F_X(x) = 1 - e^{-(\lambda+\lambda_0)x}.$$

**Solution:** D'après la question précédente

$$F_X(x) = 1 - \mathbb{P}(X > x) = 1 - \mathbb{P}(T > x)\mathbb{P}(C > x).$$

D'après 1.(a),

$$\mathbb{P}(T > x) = 1 - F_T(x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}.$$

De même,

$$\mathbb{P}(C > x) = e^{-\lambda_0 x}.$$

D'où

$$F_X(x) = 1 - e^{-\lambda x}e^{-\lambda_0 x} = 1 - e^{-(\lambda+\lambda_0)x}.$$

(d) En déduire que la densité de $X$ est

$$f_X(x) = (\lambda + \lambda_0)e^{-(\lambda+\lambda_0)x}\mathbf{1}_{\{x>0\}}$$

et que

$$\mathbb{E}[X] = \frac{1}{\lambda + \lambda_0}.$$

**Solution:**

**Méthode 1** On remarque que, d'après (c) et (a), $F_X$ est la fonction de répartition d'une loi exponentielle de paramètre $\lambda + \lambda_0$, on en déduit donc la densité et l'espérance.

**Méthode 2** Pour une loi continue, la densité est la dérivée de la fonction de répartition donc, d'après (c), pour tout $x > 0$,

$$f_X(x) = F'_X(x) = (\lambda + \lambda_0)e^{-(\lambda+\lambda_0)x}.$$

Et d'autre part

$$\mathbb{E}[X] = \int_0^{+\infty} x f_X(x)dx = (\lambda + \lambda_0)\int_0^{+\infty} xe^{-(\lambda+\lambda_0)x}dx.$$

Par intégration par parties ($u(x) = x$, $v'(x) = e^{-(\lambda+\lambda_0)x}$), on obtient le résultat voulu.

2. Nous définissons dans un premier temps un estimateur des moments de $\lambda$.

(a) Proposer un estimateur des moments $\widehat{\lambda}^{Mom}$ de $\lambda$.

**Solution:** D'après 1.(d),

$$\lambda = \frac{1}{\mathbb{E}[X]} - \lambda_0.$$

Donc, en posant $\mu_1 = \mathbb{E}[X]$, $\lambda = g(\mu_1)$ avec $g(t) = \frac{1}{t} - \lambda_0$. On définit donc,

$$\widehat{\lambda}^{Mom} = g(\widehat{\mu}_1) = \frac{1}{\widehat{\mu}_1} - \lambda_0,$$

où

$$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

qui est bien défini car $\widehat{\mu}_1 > 0$.

(b) Montrer que $\widehat{\lambda}^{Mom}$ est convergent.

**Solution:** La fonction $g$ est continue sur $]0, +\infty[$ et $\mu_1 \in ]0, +\infty$, donc, d'après le cours $\widehat{\lambda}^{Mom}$ est convergent.

(c) Montrer que $\widehat{\lambda}^{Mom}$ est asymptotiquement normal et exprimer sa variance asymptotique en fonction de $\lambda$ et $\lambda_0$.

**Solution:** $X$ suit une loi exponentielle, donc elle admet un moment d'ordre 2,

$$\mu_2 = \mathbb{E}[X^2] = \mathrm{Var}(X) + \mu_1^2 = \frac{1}{(\lambda+\lambda_0)^2} + \left(\frac{1}{\lambda+\lambda_0}\right)^2 = \frac{2}{(\lambda+\lambda_0)^2}.$$

La fonction $g$ est de classe $\mathcal{C}^1$ sur $]0, +\infty[$ et $\mu_1 \in ]0, +\infty[$, de plus

$$g'(\mu_1) = -\frac{1}{\mu_1^2} \neq 0.$$

Donc, d'après le cours (résultat issu de la méthode $\Delta$), $\widehat{\lambda}^{Mom}$ est asymptotiquement normal et sa variance asymptotique est

$$\mathrm{Var}^{(n)}(\widehat{\lambda}^{Mom}) = \frac{(g'(\mu_1))^2 \mathrm{Var}(X)}{n} = \frac{1}{\mu_1^4(\lambda+\lambda_0)^2 n}.$$

Comme $\mu_1 = \frac{1}{\lambda+\lambda_0}$, cela donne

$$\mathrm{Var}^{(n)}(\widehat{\lambda}^{Mom}) = \frac{(\lambda+\lambda_0)^2}{n}.$$

3. Passons maintenant à l'estimation par maximum de vraisemblance.

(a) Calculer l'estimateur du maximum de vraisemblance de $\lambda$.

**Solution:** D'après la question 1.(d), la log-vraissemblance du problème est, pour tout $\mathbf{x} = (x_1, \ldots, x_n)$ tel que $x_i > 0$ pour tout $i$,

$$\begin{aligned}\ell(\lambda; \mathbf{x}) &= \sum_{i=1}^{n} \log(f_X(x_i)) = \sum_{i=1}^{n} \log((\lambda + \lambda_0) e^{-(\lambda+\lambda_0)x_i}) \\ &= \sum_{i=1}^{n} \log(\lambda + \lambda_0) - (\lambda + \lambda_0) \sum_{i=1}^{n} x_i \\ &= n \log(\lambda + \lambda_0) - (\lambda + \lambda_0) \sum_{i=1}^{n} x_i.\end{aligned}$$

Calculons la dérivée de la log-vraisemblance

$$\frac{\partial \ell}{\partial \lambda}(\lambda; \mathbf{x}) = \frac{n}{\lambda + \lambda_0} - \sum_{i=1}^{n} x_i.$$

La condition nécessaire nous donne

$$0 = \frac{\partial \ell}{\partial \lambda}(\lambda^*; \mathbf{x}) = \frac{n}{\lambda^* + \lambda_0} - \sum_{i=1}^{n} x_i$$

c'est-à-dire

$$\lambda^* = \frac{n}{\sum_{i=1}^{n} x_i} - \lambda_0.$$

Pour vérifier la condition suffisante, calculons la dérivée seconde

$$\frac{\partial^2 \ell}{\partial \lambda^2}(\lambda^*; \mathbf{x}) = -\frac{n}{(\lambda^* + \lambda_0)^2} < 0.$$

Donc l'estimateur du maximum de vraisemblance est

$$\widehat{\lambda}^{EMV} = \frac{n}{\sum_{i=1}^{n} X_i} - \lambda_0.$$

(b) Calculer l'information de Fisher associée au problème d'estimation.

**Solution:**

$$I(\lambda) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \lambda^2}(\lambda; \mathbf{x})\right] = \frac{n}{(\lambda + \lambda_0)^2}.$$

(c) Nous supposons que les conditions $H_{reg}$ du cours sont vérifiées. En déduire que $\widehat{\lambda}^{EMV}$ est asymptotiquement normal et préciser sa variance asymptotique.

**Solution:** D'après le cours, $\widehat{\lambda}^{EMV}$ est asymptotiquement normal et sa variance asymptotique est

$$\mathrm{Var}^{(n)}(\widehat{\lambda}^{EMV}) = \frac{1}{I(\lambda)} = \frac{(\lambda+\lambda_0)^2}{n}.$$

4. À partir des questions précédentes, donner un intervalle de confiance au niveau $\alpha$ pour $\lambda$.

**Solution:** On remarque que $\widehat{\lambda}^{EMV} = \widehat{\lambda}^{Mom}$, nous noterons donc par la suite $\widehat{\lambda} = \widehat{\lambda}^{EMV} = \widehat{\lambda}^{Mom}$. En utilisant indifféremment la question 3.(c) ou la question 2.(c), nous avons

$$\frac{\sqrt{n}}{\lambda+\lambda_0}\left(\widehat{\lambda}-\lambda\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

$\lambda_0$ est supposé connu mais $\lambda$ est inconnu. D'après, 2.(b), $\widehat{\lambda}$ est un estimateur convergent de $\lambda$, donc, par continuité de la fonction $t \mapsto \frac{t+\lambda_0}{\lambda+\lambda_0}$,

$$\frac{\widehat{\lambda}+\lambda_0}{\lambda+\lambda_0} \xrightarrow[n\to+\infty]{\mathbb{P}} 1.$$

En utilisant le théorème de Slutsky, nous avons finalement,

$$\frac{\sqrt{n}}{\widehat{\lambda}+\lambda_0}\left(\widehat{\lambda}-\lambda\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Nous pouvons en déduire directement l'intervalle de confiance suivant pour $\lambda$.

$$\left]\widehat{\lambda} \pm q_{1-\alpha/2}\frac{\widehat{\lambda}+\lambda_0}{\sqrt{n}}\right[.$$