DST2 – Statistique

Polycopié et calculatrice autorisés. Tout autre document interdit.
Durée 2h
Date : 6 décembre 2021

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

Exercice 1 (questions d'application directe du cours (/9))

1. (3 points) a) Calculer, en justifiant bien toutes les étapes, l'estimateur du maximum de vraisemblance de la moyenne μ de la loi normale $\mathcal{N}(\mu, 2)$. On rappelle que la densité d'une variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right).$$

Solution: La vraisemblance des données s'écrit

$$L(\mu; \mathbf{x}) = \prod_{i=1}^{n} f(x_i), \quad \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Nous calculons la log-vraisemblance

$$\ell(\mu, \mathbf{x}) = \sum_{i=1}^{n} \left(-\frac{(x_i-\mu)^2}{8} - \ln(\sqrt{8\pi})\right).$$

La condition nécessaire (CN) s'écrit

$$\frac{\partial \ell}{\partial \mu}(\mu^*; \mathbf{x}) = \sum_{i=1}^{n} \frac{2(x_i - \mu^*)}{8} = 0$$

i.e.

$$\sum_{i=1}^{n} x_i = n\mu^* \Leftrightarrow \mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

La condition suffisante (CS) s'écrit

$$\frac{\partial^2 \ell}{\partial \mu^2}(\mu^*; \mathbf{x}) = -\frac{n}{4} < 0,$$

qui est vérifiée. Donc

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

b) Quel est le score des observations $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 2)$?

Solution:

$$\frac{\partial \ell}{\partial \mu}(\mu; X_1, \ldots, X_n) = \frac{1}{4}\sum_{i=1}^{n}(X_i - \mu).$$

2. (2 points) Nous considérons un test de région critique

$$\mathcal{R}_\alpha = \{\mathbf{x}; T(\mathbf{x}) < q_\alpha\}.$$

Nous savons que la statistique du test $T(= T(X_1, \ldots, X_n)) \sim \mathcal{N}(0, 1)$ sous H_0 et $T \sim \mathcal{N}(c, 1)$ (avec c une constante non nulle) sous H_1. Exprimer en fonction de q_α, c et de la fonction de répartition F_Z de la loi $\mathcal{N}(0, 1)$:

a) le risque de seconde espère du test.

Solution: Le risque de seconde espèce est la quantité

$$\beta = \mathbb{P}_{H_1}(\mathcal{R}_\alpha^c) = \mathbb{P}_{H_1}(T \geq q_\alpha) = \mathbb{P}_{H_1}(T - c \geq q_\alpha - c) = 1 - F_Z(q_\alpha - c).$$

b) la p-valeur de ce test.

Solution: D'après le cours,

$$p - valeur = F_Z(T(\mathbf{x})).$$

3. On observe $X_1, \ldots, X_n \sim_{i.i.d} \mathcal{N}(\theta, 0.5)$. On pose $\widehat{\theta} = 1$. Est-ce que $\widehat{\theta}$ est un estimateur de θ?

Solution: Oui car c'est une fonction des observations, même si ce n'est pas un estimateur convergent de θ dès que $\theta \neq 1$.

4. On souhaite déterminer si, dans une population de personnes actives données, le groupe socio-professionnel et le sexe sont des variables dépendantes. La variable groupe socio-professionnel a 6 modalités possibles (1. agriculteur exploitant; 2. artisan, commerçant et chef d'entreprise; 3. cadre et profession intellectuelle supérieure; 4. profession intermédiaire; 5. employé et 6. ouvrier). Quelle est la loi asymptotique de la statistique du test d'indépendance du χ^2 (on précisera le nombre de degrés de liberté) ?

Solution: C'est la loi du χ^2 à $(6-1) \times (2-1) = 5$ degrés de liberté.

5. Soit $\Omega = \{1, 2, 3\}$, la fonction Card qui à $A \subset \Omega$ associe le nombre d'éléments de A est-elle une mesure de probabilité sur A ?

Solution: Non, car $\text{Card}(\{1, 2\}) = 2 > 1$.

6. Nous observons $x_1, \ldots, x_{100}$ une réalisation d'un échantillon de loi $\mathcal{N}(\mu, 1/2)$. La moyenne des observations

$$\widehat{\mu}(\mathbf{x}) = 1.15.$$

Donner, sans justifier votre réponse, un intervalle de confiance à 90% de μ.

Solution: D'après le cours

$$IC_{90\%} = \left[\widehat{\mu}((x) \pm q_{0.95}\frac{\sigma}{\sqrt{n}}\right] = \left[1.15 \pm 1.64\frac{1}{10\sqrt{2}}\right] = [1.03; 1.27]$$

Exercice 2 (Modélisation de la pyramide des âges (11 points))

Cet exercice est inspiré librement du rapport *L'économie sociale et solidaire, très féminisée et des salariés plus âgés* par Erwan Porte, Simonovici Maxime (Insee) et Jeanne Fulloy (Chambre Régionale de l'économie Sociale et Solidaire). INSEE ANALYSES CENTRE-VAL DE LOIRE No 81 Paru le : 30/11/2021.

Nous souhaitons modéliser la répartition des âges des salariés du secteur de l'économie sociale et solidaire. Nous nous appuyons sur les données recueillies dans la région du Centre-Val de Loire représentées dans la figure 1.

Nous considérons une approximation de la loi des observations par une loi uniforme sur l'intervalle $[20, 60[$ et une loi de Pareto sur l'intervalle $[60; +\infty[$. Plus précisément, soit X_i l'âge du i-ème salarié, nous supposons que nous observons $X_1, \ldots, X_n$ i.i.d. selon la loi de densité

$$f_X(t) = \frac{p}{b-a}\mathbf{1}_{[a,b[}(t) + (1-p)k\frac{b^k}{t^{k+1}}\mathbf{1}_{t\geq b},$$

où a, b et p sont supposés connus et fixés à $p = 0.93$, $a = 20$ et $b = 60$ et $k > 0$ est un paramètre inconnu.

1. Nous nous intéressons d'abord à un estimateur des moments de k.

   a) Soit X une variable aléatoire de densité f_X. Montrer que le premier moment $\mu_1 = \mathbb{E}[X]$ de X vérifie

$$\mu_1 = \begin{cases} \frac{(a+b)p}{2} + \frac{bk}{k-1}(1-p) & \text{si } k > 1, \\ +\infty & \text{sinon.} \end{cases}$$

Solution: Si $k > 1$,

$$\begin{aligned}\mu_1 &= \int_{\mathbb{R}} x f_X(x)dx = \frac{p}{b-a}\int_a^b x dx + (1-p)kb^k \int_b^{+\infty} \frac{x}{x^{k+1}}dx \\ &= \frac{p}{b-a}\left[\frac{x^2}{2}\right]_a^b + (1-p)kb^k\left[\frac{x^{-k+1}}{-k+1}\right]_b^{+\infty} \\ &= \frac{p}{b-a}\frac{b^2-a^2}{2} + (1-p)kb^k\frac{b^{-k+1}}{k-1}\end{aligned}$$

ce qui donne bien le résultat attendu en observant que $b^2 - a^2 = (b-a)(b+a)$. Si $k \in]0, 1]$, nous observons que la première intégrale est finie mais la seconde est infinie, donc la somme est infinie.

b) Supposons $k > 1$ et notons

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

Montrer que

$$\widehat{k}^{Mom} = \frac{2\widehat{\mu}_1 - (a+b)p}{2\widehat{\mu}_1 + (b-a)p - 2b},$$

est un estimateur des moments de k.

Solution: Nous avons, d'après a)

$$\frac{b}{1-1/k}(1-p) = \frac{bk}{k-1}(1-p) = \mu_1 - \frac{(a+b)p}{2}.$$

donc

$$1 - \frac{1}{k} = \frac{b(1-p)}{\mu_1 - \frac{(a+b)p}{2}} = \frac{2b(1-p)}{2\mu_1 - (a+b)p}$$

et

$$\frac{1}{k} = 1 - \frac{2b(1-p)}{2\mu_1 - (a+b)p} = \frac{2\mu_1 - (a+b)p - 2b(1-p)}{2\mu_1 - (a+b)p} = \frac{2\mu_1 + (b-a)p - 2b}{2\mu_1 - (a+b)p}$$

donc

$$k = \frac{2\mu_1 - (a+b)p}{2\mu_1 + (b-a)p - 2b}$$

d'où le résultat.

c) Montrer que $\mathrm{Var}(X) < +\infty$ si et seulement si $k > 2$ (sans calculer la valeur précise de $\mathrm{Var}(X)$).

Solution: Il suffit de vérifier que $\mathbb{E}[X^2] < +\infty$. Nous avons

$$\mathbb{E}[X^2] = \frac{p}{b-1}\int_a^b x^2 dx + (1-p)kb^k \int_b^{+\infty} x^{-k+1} dx.$$

La première intégrale est celle d'une fonction bornée sur un ensemble compact, elle est donc finie, la seconde est finie si et seulement si $-k+1 < -1$ i.e. $k > 2$.

e) En déduire que, si $k > 2$, $\widehat{k}^{Mom}$ est asymptotiquement normal et écrire sa variance asymptotique en fonction de $\mathrm{Var}(X)$, μ_1, a, b et p. On supposera ici que $\mu_1 \neq (a-b)p/2 + b$ et on ne calculera pas explicitement $\mathrm{Var}(X)$.

Solution: On applique la méthode Δ. On remarque que $\widehat{k}^{Mom} = g(\widehat{\mu}_1)$ avec

$$g(t) = \frac{2t - (a+b)p}{2t + (b-a)p - 2b}.$$

La fonction g est de classe $\mathcal{C}^1$ sur l'ensemble

$$\{t, 2t + (b-a)p - 2b \neq 0\} = \mathbb{R} \setminus \left\{ \frac{a-b}{2}p + b \right\}$$

auquel μ_1 appartient. Nous avons

$$\begin{aligned} g'(t) &= \frac{2(2t + (b-a)p - 2b) - 2(2t - (a+b)p)}{(2t + (b-a)p - 2b)^2} \\ &= \frac{2((b-a)p - 2b) + 2(a+b)p}{(2t + (b-a)p - 2b)^2} \\ &= \frac{4b(p-1)}{(2t + (b-a)p - 2b)^2} \neq 0. \end{aligned}$$

Donc, d'après 4. et la méthode Δ, la variance asymptotique de $\widehat{k}^{Mom}$ est

$$\mathrm{Var}^{(n)}(\widehat{k}) = \frac{\mathrm{Var}(X) g'(\mu_1)^2}{n}.$$

2. Nous nous tournons maintenant vers l'estimation par maximum de vraisemblance.

a) Montrer que la log-vraisemblance des données peut s'écrire, pour tout $\mathbf{x} = (x_1, \ldots, x_n) \in]a, +\infty[$,

$$\ell(k; \mathbf{x}) = \sum_{i=1}^n \mathbf{1}_{\{x_i \geq b\}} (\ln(k) + k \ln b - (k+1)\ln(x_i) + \ln(1-p)) + \sum_{i=1}^n \mathbf{1}_{[a,b[}(x_i) \ln\left(\frac{b}{b-a}\right)$$

Solution: Nous avons

$$L(k;\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i).$$

Donc

$$\begin{aligned}
\ell(k;\mathbf{x}) &= \ln(L(k;\mathbf{x})) = \sum_{i=1}^{n} \ln f_X(x_i) \\
&= \sum_{i=1}^{n} \ln\left(\frac{p}{b-a}\mathbf{1}_{[a,b[}(x_i) + (1-p)k\frac{b^k}{x_i^{k+1}}\mathbf{1}_{x_i\geq b}\right) \\
&= \sum_{i=1}^{n}\left(\mathbf{1}_{[a,b[}(x_i)\ln\left(\frac{p}{b-a}\right) + \mathbf{1}_{x_i\geq b}\ln\left((1-p)k\frac{b^k}{x_i^{k+1}}\right)\right),
\end{aligned}$$

on obtient le résultat en utilisant les propriétés de la fonction logarithme et en réarrangeant les termes de la somme.

b) Calculer la dérivée partielle $\frac{\partial}{\partial k}\ell(k;\mathbf{x})$ et vérifier que

$$\frac{\partial^2}{\partial k^2}\ell(k;\mathbf{x}) = -\frac{\sum_{i=1}^{n}\mathbf{1}_{\{x_i\geq b\}}}{k^2}.$$

Solution:

$$\frac{\partial}{\partial k}\ell(k;\mathbf{x}) = \sum_{i=1}^{n}\mathbf{1}_{\{x_i\geq b\}}\left(\frac{1}{k} + \ln b - \ln(x_i)\right),$$

on dérive une nouvelle fois pour obtenir la dérivée seconde.

d) Calculer l'estimateur du maximum de vraisemblance de k.

Solution: Nous résolvons d'abord l'équation

$$\frac{\partial}{\partial k}\ell(k^*;\mathbf{x}) = 0.$$

Cela donne, d'après b),

$$\frac{\sum_{i=1}^n \mathbf{1}_{\{x_i\geq b\}}}{k^*} = \sum_{i=1}^n \mathbf{1}_{\{x_i\geq b\}}(\ln(x_i)-\ln(b)) = \sum_{i=1}^n \mathbf{1}_{\{x_i\geq b\}}\ln\left(\frac{x_i}{b}\right).$$

D'où

$$k^* = \frac{\sum_{i=1}^n \mathbf{1}_{\{x_i\geq b\}}}{\sum_{i=1}^n \mathbf{1}_{\{x_i\geq b\}}\ln\left(\frac{x_i}{b}\right)}.$$

D'après c), la dérivée seconde est négative, le point critique k^* est donc bien un maximum de la log-vraisemblance. Nous avons donc

$$\widehat{k}^{EMV} = \frac{\sum_{i=1}^n \mathbf{1}_{\{X_i\geq b\}}}{\sum_{i=1}^n \mathbf{1}_{\{X_i\geq b\}}\ln\left(\frac{X_i}{b}\right)}.$$

e) Calculer l'information de Fisher associée à l'échantillon $(X_1,\ldots,X_n)$ pour le paramètre k.

Solution: D'après le cours,

$$\mathcal{I}(k) = -\mathbb{E}\left[\frac{\partial^2}{\partial k^2}\ell(k;X_1,\ldots,X_n)\right] = -\frac{1}{k^2}\mathbb{E}\left[\sum_{i=1}^n \mathbf{1}_{\{X_i\geq b\}}\right] = -\frac{n}{k^2}\mathbb{P}(X\geq b).$$

Or

$$\mathbb{P}(X\geq b) = \int_b^{+\infty} f_X(t)dt = \int_b^{+\infty}(1-p)k\frac{b^k}{t^{k+1}}dt = 1-p.$$

Donc

$$\mathcal{I}(k) = \frac{n(1-p)}{k^2}.$$

f) Nous supposons vérifiées les hypothèses H_{reg} décrites dans le polycopié. Montrer que l'estimateur du maximum de vraisemblance est convergent et asymptotiquement normal, de variance asymptotique

$$\mathrm{Var}^{(n)}(\widehat{k}^{EMV}) = \frac{k^2}{n(1-p)}.$$

Solution: On applique le théorème du cours et la question e).

g) En déduire un intervalle de confiance asymptotique au niveau de risque α pour k.

Solution: D'après f),

$$\sqrt{\frac{n(1-p)}{k^2}}\left(\widehat{k}^{(EMV)}-k\right)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}(0,1).$$

Pour obtenir un intervalle de confiance, nous devons remplacer k inconnu par un estimateur convergent qui peut être $\widehat{k}^{EMV}$ (en utilisant le résultat de la question f). En appliquant le théorème de Slutsky, nous avons,

$$\frac{\sqrt{n(1-p)}}{\widehat{k}^{(EMV)}}\left(\widehat{k}^{(EMV)}-k\right)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}(0,1)$$

d'où

$$\lim_{n\to+\infty}\mathbb{P}\left(\left|\frac{\sqrt{n(1-p)}}{\widehat{k}^{(EMV)}}\left(\widehat{k}^{(EMV)}-k\right)\right|\leq q_{1-\alpha/2}\right)=1-\alpha.$$

On en déduit l'intervalle de confiance

$$\left[\widehat{k}^{(EMV)}\pm q_{1-\alpha/2}\frac{\widehat{k}^{(EMV)}}{\sqrt{n(1-p)}}\right].$$

3. Application numérique : nous calculons à partir des observations $x_1,\ldots,x_n$ des âges de $n = 53\,392$ salariés :

   - un âge moyenne de 40.65 ans;
   - que 3642 salariés ont plus de 60 ans;
   - une valeur de

   $$\sum_{i=1}^{n}\mathbf{1}_{\{x_i\geq 60\}}\ln(x_i)=15\,097$$

   Calculer les réalisations des estimateurs des moments et du maximum de vraisemblance sur ces données ainsi que l'intervalle de confiance de la question 2.g).

Solution: Nous avons, à deux chiffres après la virgule près,

$$\widehat{k}^{Mom}=-4.63<0!!!$$

$$\widehat{k}^{EMV}=19.69$$

et un intervalle de confiance au niveau 95% de

$$[19.05;20.33].$$

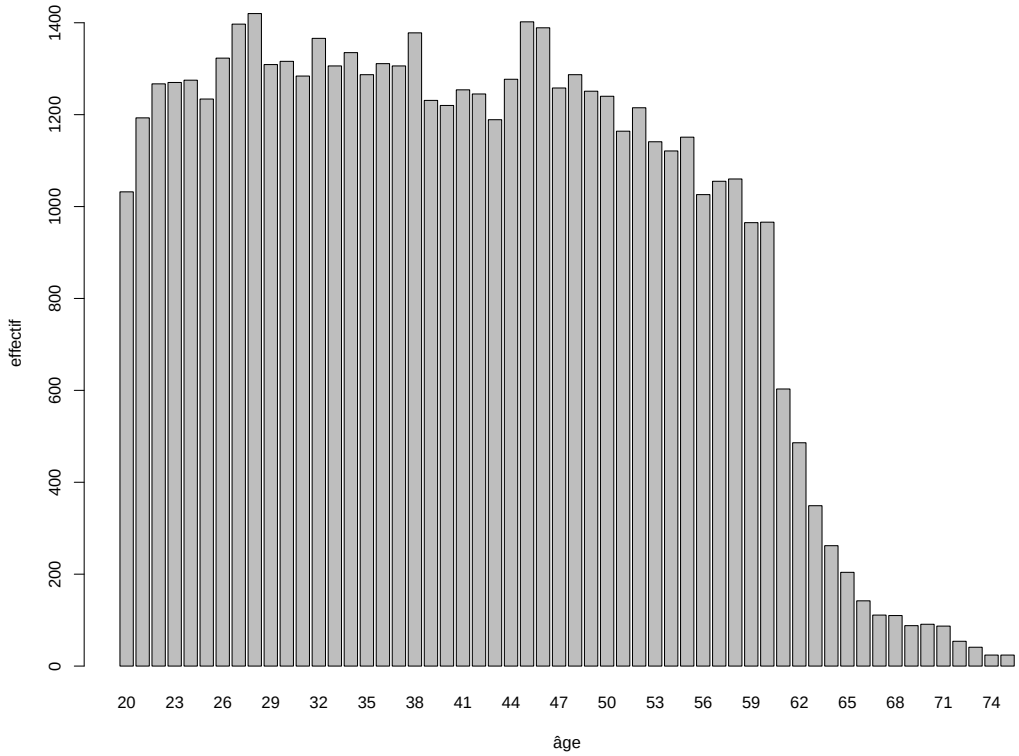

Figure 1: Pyramide des âges des hommes salariés dans le secteur de l'ESS en Centre-Val de Loire