Université Paris-Dauphine L3 CPES

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 $X \sim \mathcal{N}(\mu, \sigma^2)$

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

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

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 espèce du test.

   b) la p-valeur de ce test.

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?

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) ?

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 ?

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.

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 $X_1, \dots, X_n$ 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
   $$\mu_1 = \begin{cases} \frac{(a+b)p}{2} + \frac{bk}{k-1}(1-p) & \text{si } k > 1, \\ +\infty & \text{sinon.} \end{cases}$$

   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.

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

   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 $\mathrm{Var}(X)$.

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 $\mathbf{x} = (x_1, \dots, x_n) \in]a, +\infty[$,
   $$\ell(k; \mathbf{x}) = \sum_{i=1}^{n} \mathbf{1}_{\{x_i \geq b\}}(\ln(k) + k\ln b - (k+1)\ln(x_i) + \ln(1-p)) + \sum_{i=1}^{n} \mathbf{1}_{[a,b[}(x_i)\ln\left(\frac{b}{b-a}\right)$$

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

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

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

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

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

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

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

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

- un âge moyen 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).

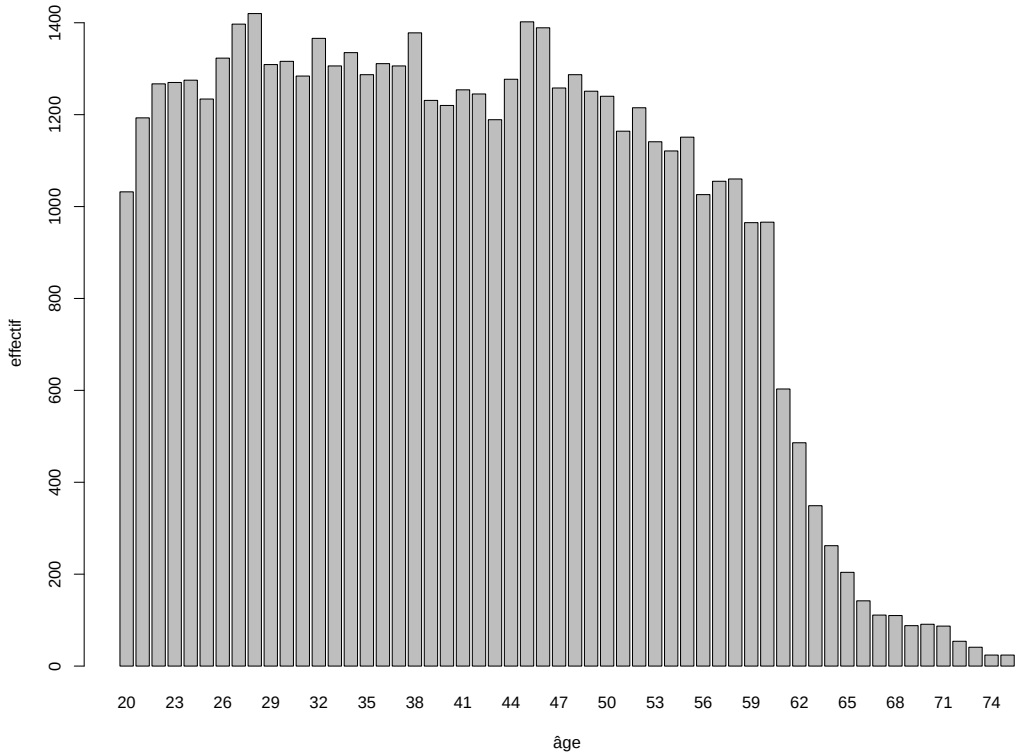

Figure 1: Pyramide des âges des hommes salariés dans le secteur de l'ESS en Centre-Val de Loire