Statistique Mathématique

Corrigé de l'examen — 6 juin 2022

En italiques des remarques qui ne font pas à proprement parler de la correction.

Exercice I

I.1. (1 point) L'hypothèse nulle est

$$H_0 : \quad \text{le patient n'est pas malade}\,,$$

le choix par défaut.

On pouvait justifier plus soigneusement en remarquant que dans ce cas l'erreur de type II est la plus grave.

I.2. (2 points) De l'énoncé, on retire les informations suivantes sur notre test :

- la probabilité d'être malade est $\mathbb{P}\,(\text{malade}) = 1/20000$;
- la probabilité d'être positif en étant malade est $\mathbb{P}\,(\text{positif}\,|\,\text{malade}) = 95/100$;
- la probabilité d'être malade sachant qu'on est positif est supérieure à $1/2$.

En utilisant la formule de Bayes, on a

$$\mathbb{P}\,(\text{positif}\,|\,\text{malade}) = \mathbb{P}\,(\text{malade}\,|\,\text{positif})\,\frac{\mathbb{P}\,(\text{positif})}{\mathbb{P}\,(\text{malade})}\,. \tag{1}$$

La seule inconnue dans l'équation précédente est $\mathbb{P}\,(\text{positif})$, et nous pouvons écrire

$$\mathbb{P}\,(\text{positif}) = \frac{\mathbb{P}\,(\text{positif}\,|\,\text{malade})\,\mathbb{P}\,(\text{malade})}{\mathbb{P}\,(\text{malade}\,|\,\text{positif})}\,.$$

Une application numérique nous donne

$$\mathbb{P}\,(\text{positif}) \leq \frac{\frac{95}{100}\cdot\frac{1}{20000}}{\frac{1}{2}} = \frac{19}{200000} \approx 9.5\times 10^{-5}\,.$$

I.3 (2 points) Le niveau souhaité est

$$\mathbb{P}\,(\text{positif}\,|\,\text{sain}) = \mathbb{P}\,(\text{sain}\,|\,\text{positif})\,\frac{\mathbb{P}\,(\text{positif})}{\mathbb{P}\,(\text{sain})}\,,$$

où nous avons utilisé la formule de Bayes. Tout est connu dans cette équation, à l'exception de

$$\mathbb{P}\,(\text{sain}) = 1 - \mathbb{P}\,(\text{malade}) \quad \text{et} \quad \mathbb{P}\,(\text{sain}\,|\,\text{positif}) = 1 - \mathbb{P}\,(\text{malade}\,|\,\text{positif})\,.$$

Une application numérique donne

$$\mathbb{P}\,(\text{positif}\,|\,\text{sain}) = (1-1/2)\frac{19/200000}{1-1/20000} = \frac{19}{399980} \approx 4.8\times 10^{-5}\,.$$

I.4. (1 point) La puissance souhaitée du test est $\mathbb{P}\,(\text{positif}\,|\,\text{malade})$. En reprenant la question I.1., plus précisément Eq. (1), nous avons

$$\mathbb{P}\,(\text{positif}\,|\,\text{malade}) = \mathbb{P}\,(\text{malade}\,|\,\text{positif})\,\frac{\mathbb{P}\,(\text{positif})}{\mathbb{P}\,(\text{malade})}\,.$$

Une application numérique nous donne

$$\mathbb{P}\,(\text{positif}\,|\,\text{malade}) = \frac{1}{2}\cdot\frac{19/200000}{1/20000} = 0.95\,.$$

Ce résultat était en fait donné par l'énoncé.

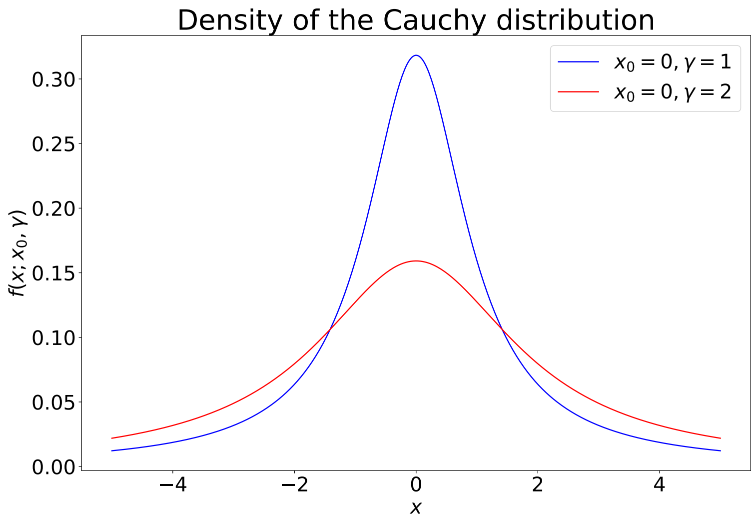

FIGURE 1 – Densité d'une Cauchy de paramètres $x_0 = 0$ et $\gamma \in \{1, 2\}$.

I.5. (1 point) Partant de

$$\mathbb{P}\left(\mathcal{N}\left(a, b^2\right) > t\right) = \mathbb{P}\left(\mathcal{N}\left(0,1\right) > \frac{t-a}{b}\right) \leq \mathrm{e}^{-(t-a)^2/(2b^2)} ,$$

on déduit facilement que

$$t > a + b\sqrt{2 \log 1/\alpha} .$$

I.6. (2 points) Au vu del'énoncé et des questions précédentes, on cherche un test de la forme "rejette si valeur plus grande que t." Nous avons deux conditions à respecter : celle sur le niveau et celle sur la puissance. En écrivant $t = \mu_0 + \sigma y$ et en utilisant la question I.5., nous prenons

$$y \leq \sqrt{2 \log 1/\alpha} ,$$

où α est le niveau calculé question I.3. Ensuite, nous partons de l'expression de la puissance :

$$\mathbb{P}\left(\mathcal{N}\left(\mu_0 + \Delta, \sigma^2\right) \geq \mu_0 + \sigma y\right) = 1 - \mathbb{P}\left(\mathcal{N}\left(0,1\right) > \frac{\Delta}{\sigma} - y\right) .$$

Cette dernière expression doit être supérieure à 0.95. En notant $\beta := 0.05$, nous trouvons

$$\sigma < \frac{\Delta}{\sqrt{2 \log 1/\alpha} + \sqrt{2 \log 1/\beta}} .$$

Exercice II

II.1. (1 point) Voir Figure 1.

II.2. (1 point) La distribution est centrée en x_0. Ce paramètre règle donc la position générale de la variable aléatoire. Le paramètre γ quant à lui règle l'échelle de la variable aléatoire : plus γ est grand plus les queues de la distribution sont lourdes.

II.3. (1 point) Une variable aléatoire suivant une loi de Cauchy n'admet pas de moment d'ordre 1 (et *a fortiori* n'admet aucun moment). On ne peut donc pas mettre en œuvre la premi'ere étape de la méthode des moments.

II.4. (1 point) Il s'agit d'une variable aléatoire à densité, et les observations sont indépendantes. On écrit donc

$$\begin{aligned}\mathcal{L}(x_1, \dots, x_n; x_0, \gamma) &= \prod_{i=1}^{n} \frac{\gamma}{\pi} \frac{1}{(x_i - x_0)^2 + \gamma^2} \\ &= \left(\frac{\gamma}{\pi}\right)^n \prod_{i=1}^{n} \frac{1}{(x_i - x_0)^2 + \gamma^2} \,.\end{aligned}$$

II.5. (1 point) De la question II.4. on déduit la log-vraisemblance

$$\ell(x_0, \gamma) := \log \mathcal{L}(x_1, \dots, x_n; x_0, \gamma) = -\sum_{i=1}^{n} \log\left[(x_i - x_0)^2 + \gamma^2\right] + n \log \frac{\gamma}{\pi} \,.$$

On dérive par rapport à x_0 et γ pour obtenir

$$\frac{\partial \ell}{\partial x_0} = \sum_{i=1}^{n} \frac{2(x_i - x_0)}{(x_i - x_0)^2 + \gamma^2} \quad \text{and} \quad \frac{\partial \ell}{\partial \gamma} = -\sum_{i=1}^{n} \frac{2\gamma}{(x_i - x_0)^2 + \gamma^2} + \frac{n}{\gamma} \,.$$

II.6. (1 point) En mettant les dérivées partielles à 0, on obtient des équations polynomiales de degré $2n - 1$. Cela est justifié car la log vraisemblance est une fonction concave.

II.7. (1 point) On peut par exemple faire une descente de gradient. Une bonne initialisation serait de commencer à la médiane des observations pour x_0^0.