Statistique Mathématique

Correction de l'examen — mai 2023

Exercice

Il s'agissait d'une reformulation de l'exercice 2 de la feuille de TD8. On pourra se reporter à la correction en ligne pour plus de détails.

I.1. (XX point). On pose H_0 : "la couleur n'a pas d'impact" qu'on teste contre H_1 : "la couleur a un impact." Il s'agit d'un test d'adéquation car on connaît la loi de référence.

I.2. (XX point). On prend comme statistique de test la distance du chi deux entre la loi empirique et la loi théorique, c'est-à-dire

$$\begin{aligned} T &= \frac{(X - np_0)^2}{np_0} + \frac{((n - X) - n(1 - p_0))^2}{n(1 - p_0)} \\ &= \frac{(X - np_0)^2}{n}\left(\frac{1}{p_0} + \frac{1}{1 - p_0}\right), \end{aligned}$$

où X est le nombre de personnes produisant l'anticorps dans notre échantillon. Sous H_0, T suit une loi du chi deux à un degré de liberté.

I.3. (XX point). On fixe comme région de rejet $]t_\alpha, +\infty[$, car T prend de grandes valeurs sous l'hypothèse alternative. Pour déterminer t_α, on se réfère à la table et on lit $3,84$ qui correspond au quantile de niveau $0,95$.

I.4. (XX point). On calcule

$$t = \frac{5^2}{50}(10 + 10/9) = \frac{50}{9} \approx 5,55\,.$$

Comme c'est une valeur strictement plus grande que $3,84$, nous rejetons l'hypothèse nulle.

Problème : estimation avec décote

II.1. (XX point). On a

$$\mathbb{E}_\theta\left[(\hat{\theta}_n - \theta)^2\right] = (\mathrm{bias}(\hat{\theta}_n))^2 + \mathrm{Var}_\theta\left(\hat{\theta}_n\right).$$

II.2. (XX point). On calcule

$$\begin{aligned} \mathrm{bias}(\hat{\theta}_n(a)) &= \mathbb{E}\left[\hat{\theta}_n\right] - \mu \\ &= \mathbb{E}\left[a\overline{X}_n\right] - \mu \\ \mathrm{bias}(\hat{\theta}_n(a)) &= (a - 1)\mu \end{aligned}$$

par linearité de l'espérance.

II.3. (XX point). On calcule

$$\begin{aligned} \mathrm{Var}\left(\hat{\theta}_n(a)\right) &= \mathrm{Var}\left(a\overline{X}_n\right) \\ &= a^2\frac{\sigma^2}{n}, \end{aligned}$$

puisque $\mathrm{Var}\left(\overline{X}_n\right) = \frac{1}{n}\mathrm{Var}\left(X_i\right)$ lorsque les X_i sont i.i.d.

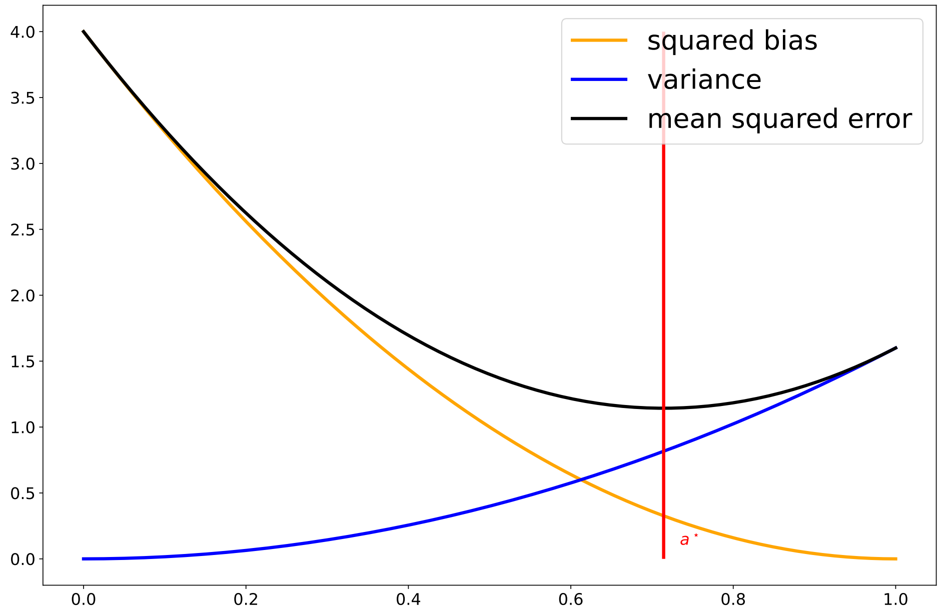

Figure 1: Décomposition biais-variance pour $\hat{\theta}_n(a)$ en fonction de a.

II.4. (XX point). En injectant les réponses des questions II.2. et II.3. dans l'égalité rappelée en II.1., on obtient

$$\mathbb{E}_\theta\left[(\hat{\theta}_n - \theta)^2\right] = (a-1)^2\mu^2 + \frac{a^2\sigma^2}{n} = \left(\mu^2 + \frac{\sigma^2}{n}\right)a^2 - 2\mu^2 a + \mu^2\,,$$

un polynôme du second degré en a.

II.5. (XX point). Voir Figure 1.

II.6. (XX point). On étudie la fonction erreur quadratique moyenne en fonction de a, on trouve un minimum global en

$$a^\star = \frac{1}{1+\frac{\sigma^2}{n\mu^2}}\,.$$

Voir Figure 1. Il existe donc un choix e paramètre $a = a^\star$ tel que l'erreur quadratique moyenne de l'estimateur est minimale. En particulier, elle est inférieure à l'erreur quadratique moyenne de l'estimateur "naturel" donné par $\overline{X}_n$.

Remarque : $a^\star$ est inférieur à 1, d'où le nom de l'estimateur.

II.7. (XX point). Lorsque σ est connu, on peut remplacer μ par $\overline{X}_n$ et considérer l'estimateur suivant :

$$\hat{\theta}_n^{(0)} := \frac{\overline{X}_n}{1+\frac{\sigma^2}{n\overline{X}_n^2}}\,.$$

Pour de grandes valeurs de n, $\frac{\sigma^2}{n\overline{X}_n^2}$ est petit, et on peut utiliser l'approximation $\frac{1}{1+x} \approx 1 - x$ pour obtenir

$$\hat{\theta}_n^{(1)} := \overline{X}_n\left(1 - \frac{\sigma^2}{n\overline{X}_n^2}\right)\,.$$

II.8. (XX point). On calcule

$$\begin{aligned}
\mathrm{bias}(\hat{\theta}_n^{(1)}) &= \mathbb{E}\left[\hat{\theta}_n^{(1)}\right] - \mu \\
&= \mathbb{E}\left[\overline{X}_n - \frac{\sigma^2}{n\overline{X}_n}\right] - \mu \\
\mathrm{bias}(\hat{\theta}_n^{(1)}) &= -\sigma^2 \mathbb{E}\left[\frac{1}{X_1 + \cdots + X_n}\right] .
\end{aligned}$$

II.9. (XX point). Ici l'astuce est de remarquer que, pour tout $x > 0$,

$$\int_0^{+\infty} \exp(-tx)\,\mathrm{d}t = \left[\frac{\exp(-tx)}{x}\right]_{t=0}^{t=+\infty} = \frac{1}{x} .$$

Donc, presque sûrement,

$$\frac{1}{X_1 + \cdots + X_n} = \int_0^{+\infty} \exp(-t(X_1 + \cdots + X_n))\,\mathrm{d}t ,$$

puisque les X_i sont des variables aléatoires positives. En intégrant des deux côtés puis en utilisant Fubini,

$$\mathbb{E}\left[\frac{1}{X_1 + \cdots + X_n}\right] = \int_0^{+\infty} \mathbb{E}\left[\exp(-t(X_1 + \cdots + X_n))\right] \mathrm{d}t .$$

On conclut en utilisant le fait que les X_i sont i.i.d., et ainsi

$$\mathbb{E}\left[\exp(-t(X_1 + \cdots + X_n))\right] = \mathbb{E}\left[\exp(-tX)\right]^n .$$

II.10. (XX point). Classiquement, pour $X \sim \mathcal{E}(\lambda)$, on obtient

$$\mathbb{E}\left[\exp(-tX)\right] = \frac{\lambda}{\lambda + t} .$$

En utilisant la question II.9. (ce qui est justifié car une variable aléatoire exponentielle est positive presque sûrement), on écrit

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{X_1 + \cdots + X_n}\right] &= \int_0^{+\infty} \left(\frac{\lambda}{\lambda + t}\right)^n \mathrm{d}t \\
&= \left[\frac{-\lambda^n}{(n-1)(\lambda + t)^{n-1}}\right]_0^{+\infty} \\
\mathbb{E}\left[\frac{1}{X_1 + \cdots + X_n}\right] &= \frac{\lambda}{n-1} .
\end{aligned}$$

Remarque : la question II.9. et ce petit calcul nous évitent de devoir obtenir la loi d'une somme de n exponentielles (c'est une loi Gamma).

D'aprs̀ la question II.8., le biais de notre estimateur est donné par

$$\mathrm{bias}(\hat{\theta}_n^{(1)}) = \frac{-\sigma^2\lambda}{n-1} = \frac{-1}{\lambda(n-1)} ,$$

puisque la variance d'une exponentielle de paramètre λ est donnée par $1/\lambda^2$. On en déduit que $\hat{\theta}_n^{(1)}$ est *asymptotiquement sans biais.*

II.11. (XX point). Calculons tout d'abord

$$\begin{aligned}
\mathrm{Var}\left(\hat{\theta}_n^{(1)}\right) &= \mathbb{E}\left[\left(\overline{X}_n - \frac{\sigma^2}{n\overline{X}_n}\right)^2\right] - \left(\mathbb{E}\left[\overline{X}_n - \frac{\sigma^2}{n\overline{X}_n}\right]\right)^2 \\
&= \mathbb{E}\left[\overline{X}_n^2 - 2\frac{\sigma^2}{n} + \frac{\sigma^4}{n^2\overline{X}_n^2}\right] - \left(\mu - \sigma^2\mathbb{E}\left[\frac{1}{X_1 + \cdots + X_n}\right]\right)^2 \\
&= \sigma^4 \mathrm{Var}\left(\frac{1}{X_1 + \cdots + X_n}\right) + 2\mu\sigma^2\mathbb{E}\left[\frac{1}{X_1 + \cdots + X_n}\right] - \frac{\sigma^2}{n} .
\end{aligned}$$

On remarque que

$$\int_0^{+\infty} \int_0^{+\infty} \exp\left(-(s+t)x\right) \mathrm{d}s\mathrm{d}t = \left(\int_0^{+\infty} \exp\left(-sx\right) \mathrm{d}s \right)^2 = \frac{1}{x^2}\,.$$

Ainsi, de la même manière qu'à la question II.9.,

$$\mathbb{E}\left[\frac{1}{(X_1 + \cdots + X_n)^2} \right] = \int_0^{+\infty} \int_0^{+\infty} \mathbb{E}\left[\exp\left(-(s+t)X\right)\right]^n \mathrm{d}s\mathrm{d}t\,.$$

En travaillant toujours sous l'hypothèse que $X \sim \mathcal{E}(\lambda)$, on en déduit que

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{(X_1 + \cdots + X_n)^2} \right] &= \int_0^{+\infty} \int_0^{+\infty} \left(\frac{\lambda}{\lambda + t + s} \right)^n \mathrm{d}s\mathrm{d}t \\
&= \int_0^{+\infty} \frac{\lambda^n}{(n-1)(\lambda+t)^{n-1}} \mathrm{d}t \\
\mathbb{E}\left[\frac{1}{(X_1 + \cdots + X_n)^2} \right] &= \frac{\lambda^2}{(n-1)(n-2)}\,.
\end{aligned}$$

On en déduit que

$$\mathrm{Var}\left(\frac{1}{X_1 + \cdots + X_n} \right) = \frac{\lambda^2}{(n-1)^2(n-2)}\,.$$

En injectant ce résultat dans le calcul général de la variance, on obtient

$$\mathrm{Var}\left(\hat{\theta}_n^{(1)}\right) = \frac{1}{\lambda^2} \cdot \frac{n^3 - 2n^2 + 2}{n(n-1)^2(n-2)}\,,$$

qui tend bien vers 0 lorsque $n \to +\infty$.