Exercise 1 — Chose promise chose due — 4 points

1. Qu'est-ce qu'un OPM ?
2. Donnez un exemple de phénomène pouvant être modélisé par une loi exponentielle ?
3. Quelle modèle statistique proposeriez vous pour représenter le nombre de personne entrant dans un bureau de poste dans un intervalle de 15 minutes ?
4. Qu'est-ce qu'une statistique ?
5. Qu'est-ce qu'un échantillon ?
6. En quoi est-ce intéressant qu'il soit i.i.d. ?
7. Comment comparer deux estimateurs ?
8. A quoi sert la borne de Cramer Rao ?

Exercise 2 — Les cars types — 11 points

Soit X une variable aléatoire réelle suivant une loi normale centrée de variance σ^2 inconnue, et donc de densité :

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{x^2}{2\sigma^2}} \quad .$$

Nous allons dans cet exercice étudier une classe d'estimateurs de l'écart type et donc du paramètre σ (et non de la variance σ^2).

1. Montrez que

   a) éventuellement en intégrant par parties

$$\begin{aligned}
\mathbb{E}(|X|) &= \int_{-\infty}^{+\infty} |x| \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{x^2}{2\sigma^2}} dx \\
&= 2\int_{0}^{+\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{x^2}{2\sigma^2}} dx \\
&= \sqrt{\frac{2}{\pi\sigma^2}} \int_{0}^{+\infty} x \exp^{-\frac{x^2}{2\sigma^2}} dx \\
&= \sqrt{\frac{2}{\pi\sigma^2}} \Big[-\sigma^2 \exp^{-\frac{x^2}{2\sigma^2}} \Big]_0^{+\infty} \\
&= \sqrt{\frac{2}{\pi}}\sigma,
\end{aligned}$$

   b) et que, éventuellement en se souvenant que $(a-b)^2 = a^2 - 2ab + b^2$,

$$\begin{aligned}
Var(|X|) &= \mathbb{E}(|X|^2) - \mathbb{E}(|X|)^2 \\
&= \sigma^2 - \frac{2}{\pi}\sigma^2 \\
&= \left(1 - \frac{2}{\pi}\right)\sigma^2.
\end{aligned}$$

2. Quelle est la borne de Cramer Rao de la variance des estimateurs sans biais de σ ?

$$\ell(\mu, X_1, \ldots, X_n) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} X_i^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi$$

$$\frac{\partial \ell(\mu, X_1, \ldots, X_n)}{\partial \sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n} X_i^2 - \frac{n}{\sigma}$$

$$\frac{\partial^2 \ell(\mu, X_1, \ldots, X_n)}{\partial \sigma^2} = -\frac{3}{\sigma^4}\sum_{i=1}^{n} X_i^2 + \frac{n}{\sigma^2}$$

$$\mathbb{E}\left(\frac{\partial^2\ell(\mu,X_1,\ldots,X_n)}{\partial\sigma^2}\right) = -\frac{3n}{\sigma^4}\sigma^2 + \frac{n}{\sigma^2}$$
$$= -\frac{2n}{\sigma^2}$$

de Cramèr Rao de ce paramètre est donc $\dfrac{\sigma^2}{2n}$.

On pose

$$\widehat{\sigma}_a = \frac{a}{n}\sum_{i=1}^{n}|X_i|,$$

où a est un coefficient positif.

3. L'estimateur sans biais
   a) Exprimer le biais de cet estimateur $\widehat{\sigma}_a$ comme une fonction du coefficient a.

$$\begin{aligned} B(a) &= \mathbb{E}(\widehat{\sigma}_a) - \sigma \\ &= \frac{a}{n}n\mathbb{E}(|X|) - \sigma \\ &= a\sqrt{\frac{2}{\pi}}\sigma - \sigma \\ &= \left(a\sqrt{\frac{2}{\pi}} - 1\right)\sigma \end{aligned}$$

   b) Comment choisir a pour que soit sans biais ?

$$B(a) = 0 \quad \Leftrightarrow \quad a_{sb}\sqrt{\frac{2}{\pi}} - 1 = 0 \quad \Leftrightarrow \quad a_{sb} = \sqrt{\frac{\pi}{2}}$$

   c) Quel est le risque de cet estimateur sans biais de a ?

$$\begin{aligned} R(\widehat{\sigma}_{a_{sb}}) &= Var(\widehat{\sigma}_{a_{sb}}) \\ &= Var\left(\sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n}|X_i|\right) \\ &= \frac{\pi}{2n}Var\left(|X|\right) \\ &= \frac{\pi}{2n}\left(1 - \frac{2}{\pi}\right)\sigma^2 \\ &= \frac{1}{n}\left(\frac{\pi}{2} - 1\right)\sigma^2 \end{aligned}$$

   d) Est-il efficace ? Pour le savoir il nous faut comparer sa variance à la borne de Cramèr Rao de ce paramètre
   Comme $\left(\frac{\pi}{2} - 1\right) > .5$, cet estimateur n'est pas efficace.

4. Le choix de a
   a) Exprimer la variance de $\widehat{\sigma}_a$ comme une fonction du coefficient a.

$$\begin{aligned} Var(\widehat{\sigma}_{a_{sb}}) &= Var\left(\frac{a}{n}\sum_{i=1}^{n}|X_i|\right) \\ &= \frac{a^2}{n^2}Var\left(|X|\right) \\ &= \frac{a^2}{n}\left(1 - \frac{2}{\pi}\right)\sigma^2 \end{aligned}$$

   b) En déduire le risque de cet estimateur $\widehat{\sigma}_a$ comme une fonction du coefficient a.

$$\begin{aligned} R(\widehat{\sigma}_a) &= B(\widehat{\sigma}_a)^2 + Var(\widehat{\sigma}_a) \\ &= \left(a\sqrt{\frac{2}{\pi}} - 1\right)^2\sigma^2 + \frac{a^2}{n}\left(1 - \frac{2}{\pi}\right)\sigma^2 \\ &= \left(\left(\frac{2}{\pi} + \frac{1}{n} - \frac{2}{n\pi}\right)a^2 - 2\sqrt{\frac{2}{\pi}}a + 1\right)\sigma^2 \end{aligned}$$

c) Montrez que, pour $n = 1$, a_{opt} la valeur de a qui minimise ce risque est

$$a_{opt} = \sqrt{\frac{2}{\pi}}.$$

$$R'(a) = \left(2\Big(\frac{2}{\pi} + \frac{1}{n} - \frac{2}{n\pi}\Big)a - 2\sqrt{\frac{2}{\pi}}\right) \sigma^2$$

soit pour $n = 1$

$$R'(a) = 2\left(a - \sqrt{\frac{2}{\pi}}\right) \sigma^2$$

et donc

$$R'(a_{opt}) = 0 \quad \Leftrightarrow \quad a_{opt} = \sqrt{\frac{2}{\pi}}.$$

5. Conclusion
   a) A la vue de ces résultats, quel est à votre sens le meilleur estimateur du paramètre σ et pourquoi ? $\widehat{\sigma}_{a_{opt}}$ car il minimise le risque,
   b) Quel est ici l'intérêt de la borne de Crame Rao ? Savoir si on peut faire mieux (ce qui est peu être le cas ici)

Exercise 3 — Un genre de trinomiale — 5 points

Soit X une variable aléatoire ne pouvant prendre que les trois valeurs 0, 1 ou 2 associée, pour un paramètre $\theta \in [0,1]$ inconnu, à la loi P_θ suivante :

$$P_\theta(X = 0) = \frac{1}{4}, \qquad P_\theta(X = 1) = \frac{3\theta}{4}, \qquad P_\theta(X = 2) = \frac{3(1-\theta)}{4}$$

1. Vérifier que, quel que soit $\theta \in [0,1]$, P_θ défini bien une distribution de probabilité.

$$P_\theta(X = 0) + P_\theta(X = 1) + P_\theta(X = 2) = \frac{1}{4}, + \frac{3\theta}{4} + \frac{3(1-\theta)}{4} = 1$$

2. Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. de n observations de loi parente P_θ avec N_0, N_1 et N_2 les variables aléatoire comptant respectivement le nombre d'observations égales à 0,1 et 2 dans l'échantillon, telles que $N_0 + N_1 + N_2 = n$ et $N_k = \sum_{i=1}^n \mathbb{I}_{\{X_i = k\}}$ pour $k = 0, 1$ et 2, où $\mathbb{I}_A$ est la fonction indicatrice de l'ensemble A, c'est-à-dire telle que $\mathbb{I}_A(x) = 1$ si $x \in A$ et 0 sinon.
   a) Montrez que $\mathbb{E}(N_0) = \frac{n}{4}$.

$$\mathbb{E}(N_0) = \mathbb{E}\left(\sum_{i=1}^n \mathbb{I}_{\{X_i=0\}}\right) = n\mathbb{E}((\mathbb{I}_{\{X_i=0\}}) = nP_\theta(X = 0) = n/4$$

   b) Quelles est la valeur de $\mathbb{E}(N_1)$?

$$\mathbb{E}(N_1) = \frac{3n\theta}{4}$$

   c) Montrez que la vraisemblance de l'échantillon s'écrit :

$$\begin{aligned} L(\theta, X_1, \ldots, X_n) &= \prod_{i=1}^n P_\theta(X = X_i) \\ &= \left(\frac{1}{4}\right)^{N_0} \left(\frac{3\theta}{4}\right)^{N_1} \left(\frac{3(1-\theta)}{4}\right)^{N_2} \end{aligned}$$

   d) Quel est l'estimateur maximum de vraisemblance de θ ?

$$\ell(\theta, X_1, \ldots, X_n) = N_0 \log(1/4) + N_1 \log\left(\frac{3\theta}{4}\right) + N_2 \log\left(\frac{3(1-\theta)}{4}\right)$$

$$\frac{\partial \ell(\theta, X_1, \ldots, X_n)}{\partial \theta} = \frac{N_1}{\theta} - \frac{N_2}{1-\theta}$$

$$\frac{N_1}{\widehat{\theta}_{MV}} - \frac{N_2}{1-\widehat{\theta}_{MV}} = 0 \Leftrightarrow \quad \widehat{\theta}_{MV} = \frac{N_1}{N_1+N_2}$$

e) Montrez que, pour ce modèle, l'information de Fisher de θ est $\dfrac{3n}{4\theta(1-\theta)}$.

$$\begin{aligned}
\frac{\partial^2 \ell(\theta, X_1, \ldots, X_n)}{\partial \theta^2} &= -\frac{N_1}{\theta^2} - \frac{N_2}{(1-\theta)^2} \\
\mathbb{E}\left(\frac{\partial^2 \ell(\theta, X_1, \ldots, X_n)}{\partial \theta^2}\right) &= -\frac{3n\theta}{4\theta^2} - \frac{3n(1-\theta)^2}{4(1-\theta)^2} \\
&= -\frac{3n}{4\theta} - \frac{3n}{4(1-\theta)} \\
&= -\frac{3n}{4}\frac{1}{\theta(1-\theta)}
\end{aligned}$$

Formulaire

- fréquences $\widehat{f_i} = \frac{n_i}{n}$ où n est le nombre total d'observations et n_i le nombre d'observant de la modalité i
- moyenne : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \hat{f}_i x_i$
- variance empirique : $\widehat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- espérance d'une variable aléatoire discrète : $\mathbb{E}(X) = \sum_i x_i \mathbb{P}(x_i)$.
- espérance d'une variable aléatoire continue de densité $f(x)$: $\mathbb{E}(X) = \int x f(x)dx$.
- variance d'une variable aléatoire X : $Var(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big)$.
- médiane : $\mathbb{P}(X < M) = 0,5$
- mode : $\underset{x\in\Omega}{Argmax}\{\mathbb{P}(x)\}$
- fractiles à l'ordre p, $\forall p \in [0,1], \widehat{\Phi}_p$ telle que $\widehat{\mathbb{P}}(X \leq \widehat{\Phi}_p) = p$
- les quartiles :
  - $\widehat{\Phi}_{\frac{1}{4}} = \widehat{Q}_1,$ telle que $\widehat{F}(\widehat{Q}_1) = \frac{1}{4}$,
  - $\widehat{\Phi}_{\frac{1}{2}} = \widehat{Q}_2 = \widehat{M},$ telle que $\widehat{F}(\widehat{M}) = \frac{1}{2}$,
  - $\widehat{\Phi}_{\frac{3}{4}} = \widehat{Q}_3,$ telle que $\widehat{F}(\widehat{Q}_2) = \frac{3}{4}$.
- covariance : $c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$
- corrélation : $\mathrm{cor}(x,y) = \frac{c_{xy}}{\sqrt{\widehat{\sigma}_x^2\widehat{\sigma}_y^2}}$
- probabilité conditionnelle : $\mathbb{P}(X = x_i | Y = y_j) = \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)}$
- espérance conditionnelle : $\mathbb{E}[Y|X = a] = \sum_{i=1}^{n} y_i \mathbb{P}(Y = y_i | X = a)$
- La loi des grands nombres $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \underset{n\to\infty}{\longrightarrow} \quad \mathbb{E}(X)$

Estimateurs :

- Le risque : $R_{\widehat{\theta}}(\theta) = \mathbb{E}\big((\widehat{\theta}(X_1,\ldots,X_n) - \theta)^2\big)$
  - Le biais d'un estimateur : $\mathbb{E}(\widehat{\theta}) - \theta$
  - La variance d'un estimateur : $\mathbb{E}\Big(\big(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\big)^2\Big)$
- Soit X une variable aléatoire de distribution $f_\theta(x)$ et $(X_1,\ldots,X_n)$ une famille de n variables aléatoires i.i.d. ayant pour loi parente la loi de X.
  - Vraisemblance :
  $$L(\theta, X_1,\ldots,X_n) = \prod_{i=1}^{n} f_\theta(X_i)$$
  Log vraisemblance :
  $$\ell(\theta, X_1,\ldots,X_n) = log\, L(\theta, X_1,\ldots,X_n) = \sum_{i=1}^{n} log\, f_\theta(X_i)$$
- L'information de Fisher : $I_n(\theta) = Var\Big(\frac{\partial\ell(\theta, X_1,\ldots,X_n)}{\partial\theta}\Big) = -\mathbb{E}\left(\frac{\partial^2\ell(\theta, X_1,\ldots,X_n)}{\partial\theta^2}\right)$
- Pour $\theta \in \mathbb{R}^p$, L'information de Fisher est la matrice $p \times p$
$$I_n(\theta) = \mathbb{E}\big(\nabla\ell(\theta, X_1,\ldots,X_n)\nabla\ell(\theta, X_1,\ldots,X_n)^\top\big)$$

- Borne de Cramer Rao
  - pour un estimateur sans biais de θ

$$BCR(\theta) = \frac{1}{I_n(\theta)} \ \leq \ \mathrm{Var}\big(\widehat{\theta}(X_1, \ldots, X_n)\big)$$

  - Pour un estimateur sans biais d'une fonction de θ : $\mathbb{E}(\widehat{u}(\theta)) = u(\theta)$

$$BCR(\theta) = \frac{u'(\theta)}{I_n(\theta)} \ \leq \ \mathrm{Var}\big(\widehat{u}(\theta(X_1, \ldots, X_n)\big)$$

  - pour un estimateur biaisé de biais B

$$BCR(\theta) = \frac{1 + B'(\theta)}{I_n(\theta)} \ \leq \ \mathrm{Var}\big(\widehat{\theta}(X_1, \ldots, X_n)\big)$$