Exercice 1 — Chose promise chose due — 4 points

1. On observe $x_i, i = 1, \dot{;} n$, le département de naissance des n élèves de la classe. Donnez la formule mathématique permettant de calculer la moyenne de ces observations ?
2. Après calcul, on a trouvé un coefficient de corrélation de -1,23. Comment interpréter ce résultat ?
3. Soient $(x_i, y_i), i = 1, \ldots, n$, n observation d'un couple de variables quantitatives. Montrez que
$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$
4. Qu'est-ce qu'une statistique ?
5. Donner un exemple d'OPM
6. En quoi est-ce intéressant qu'un échantillon soit i.i.d. ?
7. Comment comparer deux estimateurs ?
8. A quoi sert la borne de Cramer Rao ?

Exercice 2 — La danse si Ted joue int — 3 points

On considère la densité de la loi jointe

$$f_{X,Y}(x, y) = \lambda^2 \exp^{-\lambda y} \text{ pour } 0 \le x \le y \text{ et } 0 \text{ sinon}$$

1. Montrez que la densité de la loi marginale en y est
$$f_Y(y) = \lambda^2 y \exp^{-\lambda y} \text{ pour } 0 \le y \text{ et } 0 \text{ sinon}$$
2. Calculez la densité conditionnelle $f_{X|Y}(x)$ de x sachant y. Vérifiez qu'il s'agit bien d'une densité de probabilité.
3. Calculez l'espérance conditionnelle $\mathbb{E}_{X|Y}(x)$ de x sachant y.

Exercice 3 — L'équilibre de Hardy-Weinberg — 3 points

En faisant l'hypothèse que les fréquences d'apparition des gènes sont en équilibre, les génotypes MM, MN et NN apparaissent dans une population avec des probabilité $(1-\theta)^2$, $2\theta(1-\theta)$ et θ^2 (et ce conformément au principe de Hardy-Weinberg) où θ est un paramètre inconnu à déterminer (qui modélise la probabilité de trouver l'allèle M dans le population). Dans un échantillon de la population chinoise de Hong Kong en 1937, les groupes sanguins se présentaient avec les fréquences suivantes, où M et N sont des antigènes érythrocytaires

Groupe sanguin	MM	MN	NN	Total
Nombre d'individu	342	500	187	1029

Donner l'estimateur max de vraisemblance de θ et l'estimation associée. Cet estimateur est-il efficace ?

Exercice 4 **L'avarie en ce début de médian** **10 points**

Soit X une variable aléatoire réelle suivant une loi normale de paramètres μ et σ^2 tous deux inconnus, et donc de densité :

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

Nous allons dans cet exercice étudier une classe d'estimateurs du paramètre σ^2.

1. Dans cette première parie, on suppose μ connu
   a) Donner l'estimateur max de vraisemblance de σ^2
   b) Cet estimateur est-il efficace ?
   c) Quelle est sa variance ?
2. Dans cette deuxième partie, on suppose μ inconnu On pose

$$\widehat{\sigma}_a^2 = \frac{a}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

où a est un coefficient positif. On rappelle que $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.
   a) Montrer que

$$\widehat{\sigma}_a^2 = \frac{a}{n} \sum_{i=1}^{n} (X_i - \mu)^2 - a(\overline{X} - \mu)^2,$$

   b) Exprimer le biais de cet estimateur $\widehat{\sigma}_a^2$ comme une fonction du coefficient a.
   c) Comment choisir a pour que soit sans biais ?
3. Le choix de a. On admettra que

$$S^2 = \frac{n}{a} \frac{\widehat{\sigma}_a^2}{\sigma^2}$$

   suit une loi du chi2 à $n-1$ degrés de liberté et donc que $\mathbb{E}(S^2) = n-1$ et $Var(S^2) = 2(n-1)$.
   a) Exprimer la variance de $\widehat{\sigma}_a^2$ comme une fonction du coefficient a.
   b) En déduire le risque de cet estimateur $\widehat{\sigma}_a$ comme une fonction du coefficient a.
   c) Montrez que a_{opt} la valeur de a qui minimise ce risque est

$$a_{opt} = \frac{n}{n+1}.$$

4. Conclusion
   a) A la vue de ces résultats, quel est à votre sens le meilleur estimateur du paramètre σ^2 et pourquoi ?
   b) Quel est ici l'intérêt de la borne de Crame Rao ?

Formulaire

- fréquences $\widehat{f_i} = \frac{n_i}{n}$ où n est le nombre total d'observations et n_i le nombre d'observation de la modalité i
- moyenne : $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \hat{f}_i x_i$
- variance empirique : $\widehat{\sigma}_x^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$
- espérance d'une variable aléatoire discrète : $\mathbb{E}(X) = \displaystyle\sum_i x_i \mathbb{P}(x_i)$.
- espérance d'une variable aléatoire continue de densité $f(x)$: $\mathbb{E}(X) = \displaystyle\int x f(x) dx$.
- 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 : $c_{xy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$
- corrélation : $\mathrm{cor}(x,y) = \dfrac{c_{xy}}{\sqrt{\widehat{\sigma}_x^2 \widehat{\sigma}_y^2}}$
- probabilité conditionnelle : $\mathbb{P}(X = x_i | Y = y_j) = \dfrac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)}$
- densité conditionnelle : $f_{X|Y}(x|y) = \dfrac{f_{(X,Y)}(x,y)}{f_Y(y)}$
- espérance conditionnelle (v.a. discrète) : $\mathbb{E}[Y|X = a] = \displaystyle\sum_{i=1}^{n} y_i \mathbb{P}(Y = y_i | X = a)$
- La loi des grands nombres : $\bar{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i \quad \underset{n\to\infty}{\longrightarrow} \quad \mathbb{E}(X)$
- Le théorème central limite : $\sqrt{n}\Big(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i - \mathbb{E}(X)\Big) \quad \underset{n\to\infty}{\overset{(d)}{\longrightarrow}} \quad \mathcal{N}(0,\sigma^2)$

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 dépendant d'un paramètre $\theta \in \mathbb{R}$.
  - Vraisemblance :

$$L(\theta, X_1,\ldots,X_n) = \prod_{i=1}^{n} f_\theta(X_i)$$

Log vraisemblance :

$$\ell(\theta, X_1,\ldots,X_n) = log\, L(\theta, X_1,\ldots,X_n) = \sum_{i=1}^{n} log\, f_\theta(X_i)$$

Le score :

$$s(\theta, X_1, \ldots, X_n) = \frac{\partial \ell(\theta, X_1, \ldots, X_n)}{\partial \theta}$$

— 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

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

Variables aléatoires et lois

— Soit $Y \sim \mathcal{N}(0, 1)$ une variable aléatoire normale centrée réduite.

— Soit $Y_1, Y_2, ..., Y_n$ un échantillon i.i.d. de taille n de loi parente $\mathcal{N}(0, 1)$.

— On appelle loi du χ^2 à n degrés de libertés la loi de la variable aléatoire $Z_n = \sum_{i=1}^n Y_i^2$. L'espérance et la variance de cette loi sont $\mathbb{E}(Z_n) = n$ et $Var(Z_n) = 2n$

— On appelle loi de student à n degrés de libertés la loi de la variable aléatoire T_n

$$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}} \qquad \begin{array}{l} N \sim \mathcal{N}(0, 1) \\ \\ X_n \sim \chi^2_n \end{array}$$