ASI - Stat - Examen Final 2021/2022 - **3 h**

On considèrera dans les exercices un risque de première espèce de 5%.

Exercise 1 — 4 points

Soit X une variable aléatoire normale d'espérance 0 et de variance σ^2 inconnue. On dispose d'un échantillon i.i.d $(X_1, \ldots, X_n)$ de v.a. parente X. On notera T la statistique $\sum_{i=1}^{n} X_i^2$ et $F_{\chi^2_\nu}$ la fonction de répartition de la loi du χ^2 à ν degrés de liberté. On veut effectuer le test suivant :

$$H0 : \sigma^2 = \sigma_0^2$$
$$H1 : \sigma^2 = \sigma_1^2 \ (> \sigma_0^2).$$

1. Montrer que la région critique optimale s'exprime en fonction de la statistique T .
2. Sous H_0, quelle est la loi de la variable aléatoire
$$Z = \frac{T}{\sigma_0^2}$$
3. Déterminer cette région critique en fonction du quantile de la loi convenable.
4. Calculer la puissance du test en fonction du quantile de la loi convenable.
5. Déterminer la région critique et la puissance du test pour $n = 10$, $\sigma_0^2 = 1$ et $\sigma_1^2 = 2$.

Exercise 2 — 4 points

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. dont la v.a. parente X est une v.a. discrète de loi de probabilité

$$p(x) = P(X = x) = \frac{\theta^x}{(1+\theta)^{x+1}} \text{ pour } x \in N$$

où θ est un paramètre réel strictement supérieur à 0.

1. Montrer que
$$\frac{\partial log L(x_1, \ldots, x_n; \theta)}{\partial \theta} = \frac{n(\overline{x} - \theta)}{\theta(\theta + 1)}.$$
où $L(x_1, \ldots, x_n; \theta)$ est la fonction de vraisemblance de l'échantillon et $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la moyenne de l'échantillon.
2. En déduire l'estimateur du maximum de vraisemblance de θ, que l'on notera $\widehat{\theta}$.
3. Cet estimateur est-il efficace ? Donner son espérance et sa variance. En déduire l'espérance et la variance de X.
4. En supposant que n est grand, déterminer un intervalle de confiance bilatéral symétrique approché pour θ au niveau de confiance $1 - \alpha$.

Exercise 3 — 4 points

Pour étudier l'action d'un produit sur un paramètre biologique, on a mesuré, sur un échantillon supposé gaussien de 10 individus, la valeur du paramètre avant et après le traitement. Les résultats sont les suivants :

Individu	1	2	3	4	5	6	7	8	9	10
Valeur avant traitement	5,33	6.13	5,66	4,50	5,35	6,32	4,24	5,83	6,27	4,86
Valeur après traitement	5,32	6,00	5,64	4,59	5,19	6,17	4,11	5,86	6,13	4.68

A votre avis, le traitement modifie-t-il de façon significative le paramètre biologique ?

Exercise 4 — 4 points

On suppose que le nombre X d'absences par semaine dans une entreprise suit une loi de Poisson de paramètre λ avec

$$p(k) = P(X = k) = \frac{\lambda^k}{k!} \exp^{-\lambda}$$

On dispose des données suivantes, où n_x est le nombre de semaines pour lesquelles on a relevé x absences :

x	0	1	2	3
n_x	5	6	2	3

On suppose que ces données peuvent être considérées comme une réalisation d'un échantillon i.i.d. $X_1, \ldots, X_n$ de taille n = 16, de variable parente X. On notera T la statistique $\sum_{i=1}^{n} X_i$. On rappelle que, si $X \sim P(\lambda)$, $Y \sim P(\mu)$ et si X et Y sont indépendantes, alors $X + Y \sim P(\lambda + \mu)$. On cherche à tester les deux hypothèses simples :

$$H_0 : \lambda = \lambda_0$$
$$H_1 : \lambda = \lambda_1 \ (< \lambda_0).$$

1. Montrer que la région critique optimale s'exprime en fonction de la statistique T.
2. Déterminer la distribution exacte de la statistique T sous l'hypothèse H_0.
3. En déduire la région critique lorsque $\lambda_0 = 1$ et le résultat du test avec les données de l'exercice.
4. Déterminer la puissance du test pour $\lambda_1 = 0.5$.

Exercise 5 — 4 points

On modélise la durée de vie en heures T d'un appareil par une variable aléatoire suivant une loi exponentielle de densité

$$\forall t \geq 0, \quad f_\mu(t) = \frac{1}{\mu} \exp^{-t/\mu}$$

où $\mu > 0$ est un paramètre réel.

On met $n = 255$ appareils en service au même moment et on note T_i la durée de vie de l'appareil numéro i.

1. Calculer l'espérance mathématique de T
2. Donner l'estimateur du maximum de vraisemblance de μ
3. Donner sa variance
4. On suppose que l'espérance du temps de fonctionnement sans panne de ce type d'appareil est inférieur à 750 heures. Tester cette hypothèse quand on a observé

$$\overline{T} = \frac{1}{n} \sum_{i=1}^{n} T_i = 812 \text{ heures}$$

5. Calculer la puissance du test si l'espérance du temps de bon fonctionnement était de 900 heures.

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 : $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \hat{f}_i x_i$

— 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 : $Var(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big)$.

— quantile (ou 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 : $c_{xy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ et corrélation : $\mathrm{cor}(x,y) = \dfrac{c_{xy}}{\sqrt{\widehat{\sigma}_x^2 \widehat{\sigma}_y^2}}$

— probabilité conditionnelle : $\mathbb{P}(X = x_i | Y = y_j) = \dfrac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)}$

— espérance conditionnelle : $\mathbb{E}[Y|X = a] = \displaystyle\sum_{i=1}^{n} y_i \mathbb{P}(Y = y_i | X = a)$

— La loi des grands nombres $\bar{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i \quad \underset{n\to\infty}{\longrightarrow} \quad \mathbb{E}(X)$

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 : $I_n(\theta) = Var\Big(\dfrac{\partial \ell(\theta, X_1, \ldots, X_n)}{\partial \theta}\Big) = -\mathbb{E}\left(\dfrac{\partial^2 \ell(\theta, X_1, \ldots, X_n)}{\partial \theta^2}\right)$

— 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

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

— L'estimateur max de vraisemblance $\widehat{\theta}_{MV}$ d'un paramètre θ est :
  — Asymptotiquement sans biais

$$\widehat{\theta}_{MV} \underset{n\to\infty}{\longrightarrow} \theta^\star$$

  — Asymptotiquement efficace

$$Var(\widehat{\theta}_{MV}) \underset{n\to\infty}{\longrightarrow} I_n^{-1}(\theta^\star)$$

  — Asymptotiquement normal

$$\sqrt{n}(\widehat{\theta}_{MV} - \theta^\star) \underset{n\to\infty}{\longrightarrow} \mathcal{N}\big(0, I_1^{-1}(\theta^\star)\big)$$

  note : $I_n(\theta^\star) = nI_1(\theta^\star)$

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 de n réalisation i.i.d. de cette variable aléatoire.
— La loi du χ^2 : On appelle loi du χ^2 à n degrés de libertés la loi de la variable aléatoire $Z_n = \sum_{i=1}^n Y_i^2$
— La loi de student : On appelle loi de student à n degrés de libertés la loi de la variable aléatoire T_n

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

Tests
— Régions de décision :
  Région d'acceptation : l'ensemble $\overline{W}$ des réalisations $(x_1, \ldots, x_n)$ pour lesquelles on garde H_0
  Région critique : l'ensemble W des réalisations $(x_1, \ldots, x_n)$ pour lesquelles on rejette H_0
— Mesures de performances des tests
  — Risque de première espèce ou niveau de signification du test α : probabilité de rejeter H_0 alors que H_0 est vraie

$$\alpha = \mathbb{P}(D = 1 | H_0) = \int_{(x_1,\ldots,x_n)\in W} L(x_1, \ldots, x_n, H_0)\, dx_1, \ldots, dx_n$$

  — Risque de seconde espèce β : probabilité de garder H_0 alors que H_1 est vraie

$$\beta = \mathbb{P}(D = 0 | H_1) = \int_{(x_1,\ldots,x_n)\in \overline{W}} L(x_1, \ldots, x_n, H_1)\, dx_1, \ldots, dx_n$$

  — puissance d'un test : $1 - \beta$ la probabilité de rejeter hypothèses nulle avec raison
  — p-valeur :
— Tests du rapport de vraisemblance

$$W = \left\{ x_1, \ldots, x_n \mid \frac{L(x_1, \ldots, x_n, H_1)}{L(x_1, \ldots, x_n, H_0)} > k \right\}$$

— Le principe de Neyman–Pearson : pour un α fixé, trouver la fonction de décision qui minimise β (ou qui maximise $1 - \beta$ la puissance du test)
— Le théorème de Neyman–Pearson : pour un test paramétrique de deux hypothèses simples, le test du rapport de vraisemblance

$$W = \left\{ x_1, \ldots, x_n \mid \frac{L(x_1, \ldots, x_n, H_1)}{L(x_1, \ldots, x_n, H_0)} > k \right\}$$

tel que k soit fixé de sorte que

$$\alpha = \int_{(x_1,\ldots,x_n)\in W} L(x_1, \ldots, x_n, H_0)\, dx_1, \ldots, dx_n$$

est optimal au sens du principe de Neyman–Pearson (parmi tous les tests de risque α, c'est celui de puissance maximale).