EXAMEN STATISTIQUE - 1SN

Lundi 16 janvier Novembre 2023 (8h-9h30)

Partiel sans document (Une feuille A4 recto-verso autorisée)

Exercice 1 : Estimation (10 points)

On considère n variables aléatoires indépendantes $X_1, ..., X_n$ suivant la même loi de Poisson de paramètre λ et m variables aléatoires $Y_1, ..., Y_m$ suivant la même loi normale $\mathcal{N}(\lambda, \sigma^2)$, où $\lambda > 0$ est un paramètre inconnu et σ^2 est un paramètre connu. On suppose de plus que les vecteurs $\boldsymbol{X}_n = (X_1, ..., X_n)^T$ et $\boldsymbol{Y}_m = (Y_1, ..., Y_m)^T$ sont indépendants. Une situation pratique dans laquelle on peut avoir ces deux ensembles d'observations $(X_1, ..., X_n)^T$ et $(Y_1, ..., Y_m)^T$ correspond par exemple au cas où $X_1, ..., X_n$ sont obtenues par un premier capteur qui discrétise les observations de manière à avoir une loi de Poisson et où $(Y_1, ..., Y_m)^T$ sont obtenues par un second capteur fournissant des données de loi normale avec une incertitude σ^2. L'objectif de cet exercice est d'étudier des estimateurs de λ basés sur les vecteurs $\boldsymbol{X}_n$ et $\boldsymbol{Y}_m$ et de déterminer s'il est avantageux de considérer ces deux vecteurs par rapport à utiliser un seul ensemble d'observations $\boldsymbol{X}_{n+m}$ ou $\boldsymbol{Y}_{n+m}$.

1. (3pts) Déterminer l'estimateur du maximum de vraisemblance du paramètre λ construit à partir de $\boldsymbol{X}_n$ noté $\hat{\lambda}_{\text{MV}}$ et montrer qu'il est sans biais et efficace.

2. (5pts) Pour profiter des deux ensembles d'observations $\boldsymbol{X}_n$ et $\boldsymbol{Y}_m$, on propose d'utiliser l'estimateur
$$\hat{\lambda}_{n,m}(\alpha) = \frac{\alpha}{n} \sum_{i=1}^{n} X_i + \frac{1-\alpha}{m} \sum_{j=1}^{m} Y_j, \quad \alpha \in]0, 1[.$$

   - (2pts) Montrer que $\hat{\lambda}_{n,m}(\alpha)$ est un estimateur sans biais du paramètre λ pour toute valeur de α. Déterminer la variance de l'estimateur $\hat{\lambda}_{n,m}(\alpha)$.
   - (1pt) Déterminer la valeur de α optimale noté α^* (dépendant de n, m, σ^2 et λ) permettant de minimiser la variance de l'estimateur $\hat{\lambda}_{n,m}(\alpha)$. On notera $\hat{\lambda}_{\text{opt}} = \hat{\lambda}_{n,m}(\alpha^*)$.
     Remarque : la valeur de α^* dépend de λ inconnu et donc n'est pas utilisable en pratique mais elle permet de quantifier le gain maximum de performance obtenu en utilisant les deux ensembles d'observations $(X_1, ..., X_n)^T$ et $(Y_1, ..., Y_m)^T$.
   - (2pts) Montrer que la variance de $\hat{\lambda}_{\text{MV}}$ **obtenue avec** $n+m$ **observations** $x_1, ..., x_{n+m}$ est inférieure à celle de $\hat{\lambda}_{\text{opt}}$ pour $\lambda < \sigma^2$. Comment expliquer ce résultat ?

3. (2pts) On définit la vraisemblance conjointe des échantillons $\boldsymbol{x}_n$ et $\boldsymbol{y}_m$, notée $L(\boldsymbol{x}_n, \boldsymbol{y}_m; \lambda)$, par le produit des vraisemblances des deux échantillons $(X_1, ..., X_n)^T$ et $(Y_1, ..., Y_m)^T$. Montrer que les extrémas de $L(\boldsymbol{x}_n, \boldsymbol{y}_m; \lambda)$ par rapport à λ sont solution d'une équation du second degré que l'on précisera. Expliquer pourquoi cette équation admet une unique racine positive qu'on ne demande pas de déterminer.

Exercice 2 : Tests Statistiques (10 points)

On considère n variables aléatoires $X_1, ..., X_n$ indépendantes de densités

$$p(x_i; \theta) = \frac{3}{\theta} x_i^2 \exp\left(-\frac{x_i^3}{\theta}\right) \mathbb{I}_{\mathbb{R}^+}(x)$$

avec $\theta > 0$ et où $\mathbb{I}_{\mathbb{R}^+}(x)$ indique la fonction indicatrice sur $\mathbb{R}^+$. On veut à l'aide d'observations $x_1, ..., x_n$, de ces variables aléatoires tester les deux hypothèses suivantes

$$H_0 : \theta = \theta_0 \text{ contre } H_1 : \theta = \theta_1 \qquad \text{avec } \theta_1 < \theta_0$$

1. (2pts) À l'aide du théorème de Neyman-Pearson, calculer la statistique T_n du test le plus puissant et indiquer la région critique de ce test. On retiendra pour T_n la fonction seule des observations.

2. (2pts) Montrer que la loi de $Y_i = \frac{2}{\theta} X_i^3$ est une loi du chi-deux à 2 degrés de liberté. Montrer que si Z et T sont deux variables aléatoires indépendantes suivant des lois du chi-deux à n_1 et n_2 degrés de liberté (i.e., $Z \sim \chi^2_{n_1}$ et $T \sim \chi^2_{n_2}$), alors on a $Z + T \sim \chi^2_{n_1+n_2}$. En déduire la loi de $\frac{2}{\theta} T_n$.

3. (1pt) En utilisant les résultats de la question précédente, exprimer le seuil du test de Neyman-Pearson en fonction du risque de première espèce α et de l'inverse de la fonction de répartition d'une loi du chi-deux dont on prendra soin de déterminer le nombre de degrés de liberté. On notera F_ν la fonction de répartition d'une loi du chi-deux à ν degrés de liberté et F_ν^{-1} son inverse.

4. (1pt) Déterminer la puissance du test en fonction du seuil du test de Neyman-Pearson, de θ_1 et de F_ν, où ν est le nombre de degrés de liberté trouvé à la question précédente.

5. (2pts) Déterminer les courbes COR associées à ce test et analyser leur comportement en fonction de θ_0 et θ_1.

6. (2pts) On désire vérifier que les observations $x_1, ..., x_n$ suivent la loi de densité $p(x_i; \theta)$ avec $\theta = 1$ à l'aide d'un test d'un Kolmogorov. Rappeler l'expression de la statistique de ce test (en fonction des valeurs de la fonction de répartition de la loi de densité $p(x; \theta = 1)$ aux points $x_{(i)}$ notées $F_\theta(x_{(i)})$, où $(x_{(1)}, x_{(2)}, ..., x_{(n)})$ est la statistique d'ordre de l'échantillon $(x_1, ..., x_n)$). Quelle est la région critique de ce test ? Qu'appelle-t-on statistique d'ordre de de l'échantillon $(x_1, ..., x_n)$? Comment calculer la valeur du seuil à partir de l'inverse de la fonction de répartition de la loi de Kolmogorov ?

LOIS DE PROBABILITÉ CONTINUES

$\mathbf{m}$: moyenne σ^2 : variance   **F. C. :** fonction caractéristique

LOI	**Densité de probabilité**	**m**	σ^2	**F. C.**
Uniforme	$f(x) = \frac{1}{b-a}$ $x \in \left]a, b\right[$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\dfrac{e^{itb} - e^{ita}}{it(b-a)}$
Gamma $\mathcal{G}(\nu, \theta)$	$f(x) = \frac{\theta^\nu}{\Gamma(\nu)} e^{-\theta x} x^{\nu-1}$ $\theta > 0,\ \nu > 0$ $x \geq 0$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\dfrac{\nu}{\theta}$	$\dfrac{\nu}{\theta^2}$	$\dfrac{1}{\left(1 - i\frac{t}{\theta}\right)^\nu}$
Inverse gamma $\mathcal{IG}(\nu, \theta)$	$f(x) = \frac{\theta^\nu}{\Gamma(\nu)} e^{-\frac{\theta}{x}} \frac{1}{x^{\nu+1}}$ $\theta > 0,\ \nu > 0$ $x \geq 0$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{\theta}{\nu-1}$ si $\nu > 1$	$\frac{\theta^2}{(\nu-1)^2(\nu-2)}$ si $\nu > 2$	(*)
Première loi de Laplace	$f(x) = \frac{1}{2} e^{-\lvert x \rvert},\quad x \in \mathbb{R}$	0	2	$\dfrac{1}{1+t^2}$
Normale univariée $\mathcal{N}(m, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},\quad x \in \mathbb{R}$	m	σ^2	$e^{imt - \frac{\sigma^2 t^2}{2}}$
Normale multivariée $\mathcal{N}_p(\boldsymbol{m}, \boldsymbol{\Sigma})$	$f(x) = K e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{m})}$ $K = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}}$ $x \in \mathbb{R}^p$	$\boldsymbol{m}$	$\boldsymbol{\Sigma}$	$e^{i\boldsymbol{u}^T\boldsymbol{m} - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}}$
Khi_2 χ^2_ν $\Gamma\left(\frac{1}{2}, \frac{\nu}{2}\right)$	$f(x) = k e^{-\frac{x}{2}} x^{\frac{\nu}{2}-1}$ $k = \dfrac{1}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)}$ $\nu \in \mathbb{N}^*,\ x \geq 0$	ν	2ν	$\dfrac{1}{(1-2it)^{\frac{\nu}{2}}}$
Cauchy $c_{\lambda,\alpha}$	$f(x) = \dfrac{1}{\pi\lambda\left(1 + \left(\frac{x-\alpha}{\lambda}\right)^2\right)}$ $\lambda > 0,\ \alpha \in \mathbb{R}$	(-)	(-)	$e^{i\alpha t - \lambda\lvert t \rvert}$
Beta $B(a, b)$	$f(x) = k x^{a-1} (1-x)^{b-1}$ $k = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ $a > 0,\ b > 0$ $x \in \left]0, 1\right[$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	(*)

LOIS DE PROBABILITÉ DISCRÈTES

$\mathbf{m}$: moyenne σ^2 : variance **F. C. :** fonction caractéristique

$p_k = P[X = k]$ $p_{1,...,m} = P[X_1 = k_1, ..., X_m = k_m]$

LOI	**Probabilités**	**m**	σ^2	**F. C.**
Uniforme	$p_k = \frac{1}{n}$ $k \in \{1, ..., n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\dfrac{e^{it}(1-e^{itn})}{n(1-e^{it})}$
Bernoulli	$p_1 = P[X = 1] = p$ $p_0 = P[X = 0] = q$ $p \in [0,1]$ $q = 1 - p$	p	pq	$pe^{it} + q$
Binomiale $B(n,p)$	$p_k = C_n^k p^k q^{n-k}$ $p \in [0,1]$ $q = 1 - p$ $k \in \{0, 1, ..., n\}$	np	npq	$(pe^{it} + q)^n$
Binomiale négative	$p_k = C_{n+k-1}^{n-1} p^n q^k$ $p \in [0,1]$ $q = 1 - p$ $k \in \mathbb{N}$	$n\dfrac{q}{p}$	$n\dfrac{q}{p^2}$	$\left(\dfrac{p}{1-qe^{it}}\right)^n$
Multinomiale	$p_{1,...,m} = \frac{n!}{k_1!...k_m!} p_1^{k_1} ... p_m^{k_m}$ $p_j \in [0,1]$ $q_j = 1 - p_j$ $k_j \in \{0, 1, \ldots, n\}$ $\sum_{j=1}^m k_j = n$ $\sum_{j=1}^m p_j = 1$	np_j	Variance : np_jq_j Covariance : $-np_jp_k$	$\left(\sum_{j=1}^m p_j e^{it}\right)^n$
Poisson $P(\lambda)$	$p_k = e^{-\lambda}\dfrac{\lambda^k}{k!}$ $\lambda > 0$ $k \in \mathbb{N}$	λ	λ	$\exp[\lambda(e^{it} - 1)]$
Géométrique	$p_k = pq^{k-1}$ $p \in [0,1]$ $q = 1 - p$ $k \in \mathbb{N}^*$	$\dfrac{1}{p}$	$\dfrac{q}{p^2}$	$\dfrac{pe^{it}}{1-qe^{it}}$