EXAMEN STATISTIQUE - 2ÈME ANNÉE APPRENTISSAGE

Mardi 21 Novembre 2023 (14h-15h30)

Partiel avec documents autorisés

Exercice 1 : Estimation (10 points)

On considère n observations $x_1, ..., x_n$ issues d'un vecteur de n variables aléatoires X_i indépendantes de lois de Poisson de paramètre $i\lambda$, i.e., telles que

$$P[X_i = x_i; \lambda] = \frac{(i\lambda)^{x_i}}{x_i!} e^{-i\lambda}, \; x_i \in \mathbb{N},$$

avec $\lambda > 0$ un paramètre inconnu.

1. On cherche à estimer le paramètre λ à l'aide de l'estimateur

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{i}.$$

Déterminer le biais de cet estimateur. En utilisant la propriété $\sum_{i=1}^{n} \frac{1}{i} \leq n$, donner un majorant de var$(\hat{\lambda})$. En déduire que $\hat{\lambda}$ est un estimateur convergent de λ.
Le biais de l'estimateur $\hat{\lambda}$ est défini par $b(\lambda) = E[\hat{\lambda}] - \lambda$. On a

$$E[\hat{\lambda}] = \frac{1}{n} \sum_{i=1}^{n} E\left[\frac{X_i}{i}\right] = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{i} E(X_i)\right] = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{i} i\lambda\right] = \lambda.$$

L'estimateur $\hat{\lambda}$ est donc un estimateur non biaisé du paramètre λ. La variance de l'estimateur $\hat{\lambda}$ est

$$\text{var}(\hat{\lambda}) = \text{var}\left(\frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{i}\right).$$

En utilisant le fait que les variables X_i sont indépendantes, on obtient

$$\text{var}(\hat{\lambda}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}\left(\frac{X_i}{i}\right) = \frac{1}{n^2} \sum_{i=1}^{n} \left(\frac{1}{i^2} \times i\lambda\right) = \frac{\lambda}{n^2} \sum_{i=1}^{n} \left(\frac{1}{i}\right).$$

En utilisant l'indication, on obtient

$$0 \leq \text{var}(\hat{\lambda}) \leq \frac{\lambda}{n}.$$

Comme $\hat{\lambda}$ est un estimateur non biaisé de λ et que sa variance tend vers 0 lorsque $n \to \infty$, $\hat{\lambda}$ est un estimateur convergent de λ.

2. En utilisant la relation $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, montrer que l'estimateur du maximum de vraisemblance du paramètre λ noté $\widehat{\lambda}_{\text{MV}}$ est défini par

$$\widehat{\lambda}_{\text{MV}} = \frac{2}{n(n+1)} \sum_{i=1}^{n} X_i.$$

Cet estimateur est-il sans biais et convergent ?
On effectue les traitements habituels

- Calcul de la vraisemblance.

$$L(x_1, ..., x_n; \lambda) = \prod_{i=1}^{n} P[X_i = x_i; \lambda] = \prod_{i=1}^{n} \frac{(i\lambda)^{x_i}}{x_i!} e^{-i\lambda},$$

- Calcul de la log vraisemblance

$$\ln L(x_1, ..., x_n; \lambda) = \sum_{i=1}^{n} [x_i \ln(i) + x_i \ln(\lambda) - \ln(x_i!) - i\lambda]$$

- Calcul de la dérivée de la log-vraisemblance et de l'estimateur

$$\frac{\partial \ln L(x_1, ..., x_n; \lambda)}{\partial \lambda} = \sum_{i=1}^{n} \left[\frac{x_i}{\lambda} - i\right] = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - \frac{n(n+1)}{2}$$

qui s'annule pour

$$\frac{1}{\lambda} \sum_{i=1}^{n} x_i - \frac{n(n+1)}{2} = 0 \Leftrightarrow \lambda = \frac{2}{n(n+1)} \sum_{i=1}^{n} x_i.$$

On en déduit que l'estimateur du maximum de vraisemblance du paramètre λ noté $\widehat{\lambda}_{\text{MV}}$ est défini par

$$\widehat{\lambda}_{\text{MV}} = \frac{2}{n(n+1)} \sum_{i=1}^{n} X_i.$$

On a

$$E\left(\widehat{\lambda}_{\text{MV}}\right) = \frac{2}{n(n+1)} \sum_{i=1}^{n} E(X_i) = \frac{2}{n(n+1)} \sum_{i=1}^{n} (i\lambda) = \lambda.$$

$\widehat{\lambda}_{\text{MV}}$ est donc un estimateur non biaisé de λ. La variance de cet estimateur est (en utilisant à nouveau l'indépendance entre les variables X_i)

$$\text{var}\left(\widehat{\lambda}_{\text{MV}}\right) = \frac{4}{n^2(n+1)^2} \sum_{i=1}^{n} \text{var}(X_i) = \frac{4}{n^2(n+1)^2} \sum_{i=1}^{n} (i\lambda) = \frac{2}{n(n+1)} \lambda.$$

Comme $\widehat{\lambda}_{\text{MV}}$ est un estimateur non biaisé de λ et que sa variance tend vers 0 lorsque $n \to \infty$, $\widehat{\lambda}_{\text{MV}}$ est un estimateur convergent de λ.

3. Déterminer la borne de Cramer-Rao pour un estimateur non biaisé du paramètre λ. L'estimateur $\widehat{\lambda}_{\text{MV}}$ est-il l'estimateur efficace du paramètre λ ?
La dérivée seconde de la log-vraisemblance est

$$\frac{\partial^2 \ln L(x_1, ..., x_n; \lambda)}{\partial \lambda^2} = \frac{-1}{\lambda^2} \sum_{i=1}^{n} x_i.$$

d'où

$$E\left[-\frac{\partial^2 \ln L(X_1, ..., X_n; \lambda)}{\partial \lambda^2}\right] = \frac{1}{\lambda^2} \sum_{i=1}^{n} (i\lambda) = \frac{n(n+1)}{2\lambda}.$$

On en déduit que la borne de Cramér-Rao pour un estimateur non-biaisé de λ est

$$\text{BCR} = \frac{-1}{E\left[\frac{\partial^2 \ln L(X_1,...,X_n;\lambda)}{\partial \lambda^2}\right]} = \frac{2\lambda}{n(n+1)}.$$

Comme $\text{var}\left[\widehat{\lambda}_{\text{MV}}\right] = \text{BCR}$, $\widehat{\lambda}_{\text{MV}}$ est l'estimateur efficace du paramètre λ.

4. En justifiant avec soin votre réponse, lequel des deux estimateurs $\hat{\lambda}$ et $\widehat{\lambda}_{\text{MV}}$ choisiriez vous ?
On choisira l'estimateur efficace $\widehat{\lambda}_{\text{MV}}$ car il est sans biais et de variance minimale.

Exercice 2 : Tests Statistiques (10 **points)**

On considère n observations $x_1, ..., x_n$ issues d'un vecteur $(X_1, ..., X_n)$ de n variables aléatoires indépendantes de même loi de densités

$$f(x_i; a) = \begin{cases} a x_i^{a-1} & \text{si} \ \ x_i \in]0, 1[\\ 0 \text{ sinon} \end{cases}$$

avec $a > 0$. On désire utiliser les observations $x_1, ..., x_n$ pour déterminer si $a = a_0 > 0$ ou si $a = a_1 > a_0$. On considère donc le test d'hypothèses

$$H_0 : a = a_0, \quad H_1 : a = a_1 \qquad \text{avec } a_1 > a_0 > 0.$$

1. Montrer que la variable aléatoire $U_i = -\ln(X_i)$ possède la loi gamma $\mathcal{G}(1, a)$ et en déduire à l'aide des tables sa moyenne $E[U_i]$ et sa variance var$[U_i]$. *Remarque* : on prendra soin de déterminer le domaine de définition de U_i.
La variable aléatoire U_i est à valeurs dans $\mathbb{R}^+$. En faisant un changement de variables (et en faisant attention de ne pas oublier le Jacobien ;-)), on obtient la densité de U_i

$$g(u_i; a) = \begin{cases} a e^{-a u_i} \text{ si } u_i > 0 \\ 0 \quad \text{sinon} \end{cases}$$

On reconnaît une loi gamma $\mathcal{G}(1, a)$ (ou loi exponentielle) dont la moyenne et la variance (donnée dans la table) sont

$$E[U_i] = \frac{1}{a} \text{ et var}[U_i] = \frac{1}{a^2}.$$

2. Montrer que la statistique du test de Neyman Pearson est $T_n = \sum_{i=1}^n U_i$ et déterminer la région critique associée.
Le test de Neyman Pearson est défini par

$$\text{Rejet de } H_0 \text{ si } \frac{L(x_1, ..., x_n; \lambda_1)}{L(x_1, ..., x_n; a_\lambda)} > S_{1,\alpha}$$

où $S_{1,\alpha}$ est un seuil dépendant du risque de première espèce α. Mais

$$\frac{L(x_1, ..., x_n; a_1)}{L(x_1, ..., x_n; a_0)} > S_{1,\alpha} \Leftrightarrow \ln \left[\frac{L(x_1, ..., x_n; a_1)}{L(x_1, ..., x_n; a_0)} \right] > S_{2,\alpha}$$
$$\Leftrightarrow (a_1 - a_0) \sum_{i=1}^n \ln(x_i) > S_{3,\alpha}.$$

Comme $a_1 > a_0$, on a la règle de décision

$$T_n = \sum_{i=1}^n U_i \leq S_\alpha.$$

La région critique du test est l'ensemble des vecteurs $(u_1, ..., u_n) \in]0, +\infty[^n$ tels que $T_n \leq S_\alpha$ et la statistique de test est $T_n = \sum_{i=1}^n U_i$.

3. Montrer que la loi approchée de T_n issue du théorème central limite est la loi normale $\mathcal{N}\left(\frac{n}{a}, \frac{n}{a^2}\right)$.
   Comme $T_n = \sum_{i=1}^n U_i$, on a

$$E[T_n] = n\sum_{i=1}^n E[U_i] = \frac{n}{a} \text{ et } \mathrm{var}[T_n] = \sum_{i=1}^n \mathrm{var}[U_i] = \frac{n}{a^2}.$$

   La loi approchée de T_n issue du théorème central limite est donc la loi normale $\mathcal{N}\left(\frac{n}{a}, \frac{n}{a^2}\right)$.

4. On note $F_{\mathcal{N}(0,1)}$ la fonction de répartition d'une loi du normale $\mathcal{N}(0,1)$. Exprimer le risque de première espèce α en fonction du seuil du test de Neyman Pearson noté S_α, de $F_{\mathcal{N}(0,1)}(\alpha)$, n et de a_0. En déduire la valeur du seuil S_α en fonction de $F^{-1}_{\mathcal{N}(0,1)}(\alpha)$ et de n et a_0.
   Le risque α est défini par

$$\alpha = P[\text{Rejeter } H_0|H_0 \text{ vraie}] = P\left[T_n \leq S_\alpha | T_n \sim \mathcal{N}\left(\frac{n}{a_0}, \frac{n}{a_0^2}\right)\right],$$

   soit

$$\alpha = F_{\mathcal{N}(0,1)}\left[\frac{S_\alpha - \frac{n}{a_0}}{\sqrt{\frac{n}{a_0^2}}}\right] = F_{\mathcal{N}(0,1)}\left(\frac{a_0}{\sqrt{n}}S_\alpha - \sqrt{n}\right).$$

   d'où

$$S_\alpha = \frac{\sqrt{n}}{a_0}\left[F^{-1}_{\mathcal{N}(0,1)}(\alpha) + \sqrt{n}\right]$$

5. Déterminer les caractéristiques opérationnelles du récepteur (courbes COR) pour ce test et montrer qu'elles ne dépendent que de n, $F^{-1}_{\mathcal{N}(0,1)}(\alpha)$ et de $\frac{a_1}{a_0}$. Analyser les performances du test en fonction de $\frac{a_1}{a_0}$ et de n et représenter l'allure de ces courbes pour diverses valeurs de $\frac{a_1}{a_0}$ et de n.
   Le risque β est défini par

$$\beta = P[\text{Rejeter } H_1|H_1 \text{ vraie}] = P\left[T_n \geq S_\alpha | T_n \sim \mathcal{N}\left(\frac{n}{a_1}, \frac{n}{a_1^2}\right)\right],$$

   soit (en remplaçant a_0 par a_1 dans l'expression de α :

$$\pi = 1 - \beta = F_{\mathcal{N}(0,1)}\left(\frac{a_1}{\sqrt{n}}S_\alpha - \sqrt{n}\right).$$

   Les courbes COR sont donc définie par

$$\pi = F_{\mathcal{N}(0,1)}\left[\frac{a_1}{a_0}\left(F^{-1}_{\mathcal{N}(0,1)}(\alpha) + \sqrt{n}\right) - \sqrt{n}\right].$$

   Les courbes COR sont des fonctions croissantes de n et de $\frac{a_1}{a_0}$. L'allure approximative de ces courbes pour différentes valeurs de $\frac{a_1}{a_0}$ et de n est représentée ci-dessous :

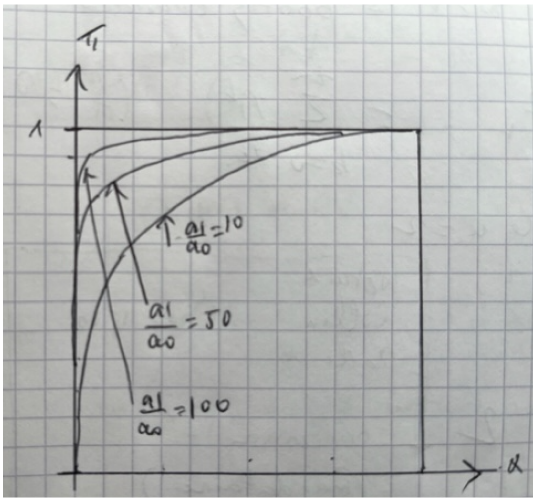

Figure 1: Allure des courbes COR pour différentes valeurs de $\frac{a_1}{a_0}$.

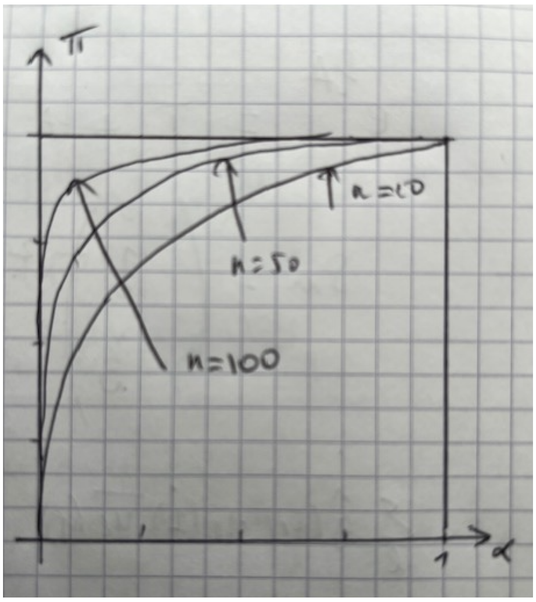

Figure 2: Allure des courbes COR pour différentes valeurs de n.

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

LOI	**Densité de probabilité**	$\mathbf{m}$	σ^2	**F. C.**
Uniforme	$f(x) = \frac{1}{b-a}$ $x \in]a, b[$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb} - e^{ita}}{it(b-a)}$
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 $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{\nu}{\theta}$	$\frac{\nu}{\theta^2}$	$\frac{1}{\left(1 - i\frac{t}{\theta}\right)^\nu}$
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	$f(x) = \frac{1}{2} e^{-\|x\|}, \quad x \in \mathbb{R}$	0	2	$\frac{1}{1+t^2}$
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 $\mathcal{N}_p(\boldsymbol{m}, \boldsymbol{\Sigma})$	$f(x) = K e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})}$ $K = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}}$ $\boldsymbol{x} = (x_1, ..., x_p)^T \in \mathbb{R}^p$	$\boldsymbol{m}$	$\boldsymbol{\Sigma}$	$e^{i\boldsymbol{u}^T \boldsymbol{m} - \frac{1}{2}\boldsymbol{u}^T \boldsymbol{\Sigma} \boldsymbol{u}}$
Khi$_2$ χ^2_ν $\Gamma\left(\frac{1}{2}, \frac{\nu}{2}\right)$	$f(x) = k e^{-\frac{x}{2}} x^{\frac{\nu}{2}-1}$ $k = \frac{1}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)}$ $\nu \in \mathbb{N}^*,\ x \geq 0$	ν	2ν	$\frac{1}{(1-2it)^{\frac{\nu}{2}}}$
Cauchy $c_{\lambda,\alpha}$	$f(x) = \frac{1}{\pi\lambda\left(1 + \left(\frac{x-\alpha}{\lambda}\right)^2\right)}$ $\lambda > 0,\ \alpha \in \mathbb{R}$	(-)	(-)	$e^{i\alpha t - \lambda\|t\|}$
Beta $B(a, b)$	$f(x) = k x^{a-1} (1-x)^{b-1}$ $k = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ $a > 0,\ b > 0$ $x \in]0, 1[$ 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	$\mathbf{m}$	σ^2	F. C.
Uniforme	$p_k = \frac{1}{n}$ $k \in \{1, ..., n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\dfrac{e^{it}\left(1-e^{itn}\right)}{n\left(1-e^{it}\right)}$
Bernoulli	$p_1 = P[X = 1] = p$ $p_0 = P[X = 0] = q$ $p \in [0,1] \quad q = 1-p$	p	pq	$pe^{it} + q$
Binomiale $B(n,p)$	$p_k = C_n^k p^k q^{n-k}$ $p \in [0,1] \quad q = 1-p$ $k \in \{0, 1, ..., n\}$	np	npq	$\left(pe^{it} + q\right)^n$
Binomiale négative	$p_k = C_{n+k-1}^{n-1} p^n q^k$ $p \in [0,1] \quad q = 1-p$ $k \in \mathbb{N}$	$n\dfrac{q}{p}$	$n\dfrac{q}{p^2}$	$\left(\dfrac{p}{1-qe^{it}}\right)^n$
Multinomiale	$p_{1,...,m} = \frac{n!}{k_1!...k_m!} p_1^{k_1} ... p_m^{k_m}$ $p_j \in [0,1] \quad q_j = 1-p_j$ $k_j \in \{0, 1, \ldots, n\}$ $\sum_{j=1}^m k_j = n \quad \sum_{j=1}^m p_j = 1$	np_j	Variance : $np_j q_j$ Covariance : $-np_j p_k$	$\left(\sum_{j=1}^m p_j e^{it}\right)^n$
Poisson $P(\lambda)$	$p_k = e^{-\lambda} \dfrac{\lambda^k}{k!}$ $\lambda > 0 \quad k \in \mathbb{N}$	λ	λ	$\exp\left[\lambda\left(e^{it} - 1\right)\right]$
Géométrique	$p_k = pq^{k-1}$ $p \in [0,1] \quad q = 1-p$ $k \in \mathbb{N}^*$	$\dfrac{1}{p}$	$\dfrac{q}{p^2}$	$\dfrac{pe^{it}}{1-qe^{it}}$