EXAMEN STATISTIQUE - 2ÈME ANNÉE APPRENTISSAGE

Lundi 18 Novembre 2024 (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 Beta de paramètres $B\left(\frac{1}{\theta}, 1\right)$ de densités

$$p(x_i;\theta) = \begin{cases} \frac{1}{\theta} x_i^{\frac{1}{\theta}-1} \text{ si } x_i \in]0,1[\\ 0 \text{ sinon} \end{cases}$$

avec $\theta > 0$ un paramètre inconnu.

1. Montrer que l'estimateur du maximum de vraisemblance du paramètre θ noté $\widehat{\theta}_{\text{MV}}$ est défini par

$$\widehat{\theta}_{\text{MV}} = -\frac{1}{n}\sum_{i=1}^{n} \ln(X_i).$$

On effectue les traitements habituels

- Calcul de la vraisemblance.

$$L(x_1, ..., x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \left[\frac{1}{\theta} x_i^{\frac{1}{\theta}-1}\right],$$

- Calcul de la log vraisemblance

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

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

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

qui s'annule pour

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

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

$$\widehat{\theta}_{\text{MV}} = -\frac{1}{n}\sum_{i=1}^{n}\ln(X_i).$$

2. Montrer que la variable aléatoire $Y_i = -\ln(X_i)$ suit une loi gamma $\Gamma\left(1, \frac{1}{\theta}\right)$. En déduire la moyenne et la variance de la variable Y_i notées $E[Y_i]$ et var$[Y_i]$.
La variable aléatoire Y_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 Y_i

$$g(y_i;\theta) = \begin{cases} \frac{1}{\theta}\exp\left(-\frac{y_i}{\theta}\right) & \text{si } y_i > 0 \\ 0 & \text{sinon} \end{cases}$$

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

$$E[Y_i] = \theta \text{ et } \text{var}[Y_i] = \theta^2.$$

3. L'estimateur $\widehat{\theta}_{\text{MV}}$ est-il sans biais et convergent ?
On a

$$E\left(\widehat{\theta}_{\text{MV}}\right) = -\frac{1}{n}\sum_{i=1}^{n} E[\ln(X_i)] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \theta.$$

$\widehat{\theta}_{\text{MV}}$ est donc un estimateur non biaisé de θ. La variance de cet estimateur est (en utilisant l'indépendance entre les variables Y_i)

$$\text{var}\left(\widehat{\theta}_{\text{MV}}\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(Y_i) = \frac{\theta^2}{n}.$$

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

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

$$\frac{\partial^2 \ln L(x_1,...,x_n;\theta)}{\partial\theta^2} = \frac{n}{\theta^2} + \frac{2}{\theta^3}\sum_{i=1}^{n}\ln(x_i).$$

d'où

$$E\left[-\frac{\partial^2 \ln L(X_1,...,X_n;\theta)}{\partial\theta^2}\right] = \frac{n}{\theta^2} + \frac{2}{\theta^3}\sum_{i=1}^{n} E[\ln(X_i)] = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} E[Y_i] = -\frac{n}{\theta^2}.$$

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

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

Comme $\text{var}\left[\widehat{\theta}_{\text{MV}}\right] = \text{BCR}$ et que l'estimateur $\widehat{\theta}_{\text{MV}}$ est non biaisé, $\widehat{\theta}_{\text{MV}}$ est l'estimateur efficace du paramètre θ.

5. Montrer que l'estimateur des moments de θ défini à partir de $E[X_i]$ est

$$\widehat{\theta}_{\text{Mo}} = \frac{n}{\sum_{i=1}^{n} X_i} - 1.$$

6. Lequel des deux estimateurs $\widehat{\theta}_{\text{Mo}}$ et $\widehat{\theta}_{\text{MV}}$ choisiriez vous (justifier votre réponse) ?
On choisira l'estimateur efficace $\widehat{\theta}_{\text{MV}}$ 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 lois de Poisson de paramètre $i\lambda$, c'est-à-dire, telles que

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

avec $\lambda > 0$. On notera que le paramètre de la loi de Poisson pour la variable aléatoire X_i dépend de l'indice i. On désire utiliser les observations $x_1, ..., x_n$ pour déterminer si $\lambda = \lambda_0 > 0$ ou si $\lambda = \lambda_1 \in]0, \lambda_0[$. On considère donc le test d'hypothèses

$$H_0 : \lambda = \lambda_0, \quad H_1 : \lambda = \lambda_1 \qquad \text{avec } 0 < \lambda_1 < \lambda_0.$$

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

   $$\text{Rejet de } H_0 \text{ si } \frac{L(x_1, ..., x_n; \theta_1)}{L(x_1, ..., x_n; \theta_0)} > S_{1,\alpha}$$

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

   $$\begin{aligned} \frac{L(x_1, ..., x_n; \theta_1)}{L(x_1, ..., x_n; \theta_0)} > S_{1,\alpha} &\Leftrightarrow \ln\left[\frac{L(x_1, ..., x_n; \theta_1)}{L(x_1, ..., x_n; \theta_0)}\right] > S_{2,\alpha} \\ &\Leftrightarrow \ln\left[\frac{\prod_{i=1}^n \frac{(i\lambda_1)^{x_i}}{x_i!} e^{-i\lambda_1}}{\prod_{i=1}^n \frac{(i\lambda_0)^{x_i}}{x_i!} e^{-i\lambda_0}}\right] > S_{2,\alpha} \\ &\Leftrightarrow [\ln(\lambda_1) - \ln(\lambda_0)] \sum_{i=1}^n x_i > S_{3,\alpha}. \end{aligned}$$

   Comme $\lambda_1 < \lambda_0$, on rejette H_0 si

   $$T_n = \sum_{i=1}^n X_i < S_\alpha.$$

   La région critique du test est donc l'ensemble des vecteurs $(x_1, ..., x_n) \in \mathbb{N}^n$ tels que $T_n < S_\alpha$ et la statistique de test est $T_n = \sum_{i=1}^n X_i$.

2. On donne la relation $\sum_{i=1}^n i = \frac{n(n+1)}{2}$. Montrer que la loi approchée de T_n issue du théorème central limite est la loi normale $\mathcal{N}\left(\frac{n(n+1)}{2}\lambda, \frac{n(n+1)}{2}\lambda\right)$. On supposera dans la suite de cet exercice qu'on peut approcher la loi de T_n par cette loi normale.
   Comme $T_n = \sum_{i=1}^n X_i$, on a

   $$E[T_n] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n (i\lambda) = \frac{n(n+1)}{2}\lambda$$

   et (comme les variables X_i sont indépendantes)

   $$\text{var}[T_n] = \sum_{i=1}^n \text{var}[X_i] = \sum_{i=1}^n (i\lambda) = \frac{n(n+1)}{2}\lambda.$$

   La loi approchée de T_n issue du théorème central limite est donc la loi normale $\mathcal{N}\left(\frac{n(n+1)}{2}\lambda, \frac{n(n+1)}{2}\lambda\right)$.

3. On note F 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 $F(\alpha)$, n et de λ_0. En déduire la valeur de S_α en fonction de $F^{-1}(\alpha)$, n et λ_0.
Le risque α est défini par

$$\alpha = P[\text{Rejeter } H_0 | H_0 \text{ vraie}] = P\left[T_n < S_\alpha | T_n \sim \mathcal{N}\left(\frac{n(n+1)}{2}\lambda_0, \frac{n(n+1)}{2}\lambda_0\right)\right],$$

soit

$$\alpha = F\left[\frac{S_\alpha - \frac{n(n+1)}{2}\lambda_0}{\sqrt{\frac{n(n+1)}{2}\lambda_0}}\right],$$

d'où

$$S_\alpha = \frac{n(n+1)}{2}\lambda_0 + F^{-1}(\alpha)\sqrt{\frac{n(n+1)}{2}\lambda_0}$$

4. Déterminer les caractéristiques opérationnelles du récepteur (courbes COR) pour ce test en fonction de n, $F^{-1}(\alpha)$, λ_0 et λ_1. Représenter l'allure de ces courbes COR pour diverses valeurs de n. Les performances du test seront-elles meilleures pour $(\lambda_0, \lambda_1) = (10, 1)$ ou pour $(\lambda_0, \lambda_1) = (1000, 100)$?
La puissance du test est définie par

$$\pi = P[\text{Rejeter } H_0 | H_1 \text{ vraie}] = F\left[\frac{S_\alpha - \frac{n(n+1)}{2}\lambda_1}{\sqrt{\frac{n(n+1)}{2}\lambda_1}}\right].$$

En remplaçant l'expression du seuil S_α déterminée précédemment, on obtient

$$\pi = F\left[F^{-1}(\alpha)\sqrt{\frac{\lambda_0}{\lambda_1}} + \sqrt{\frac{n(n+1)}{2}}\frac{\lambda_0 - \lambda_1}{\sqrt{\lambda_1}}\right].$$

L'allure des courbes COR pour différentes valeurs de n est représentée ci-dessous :

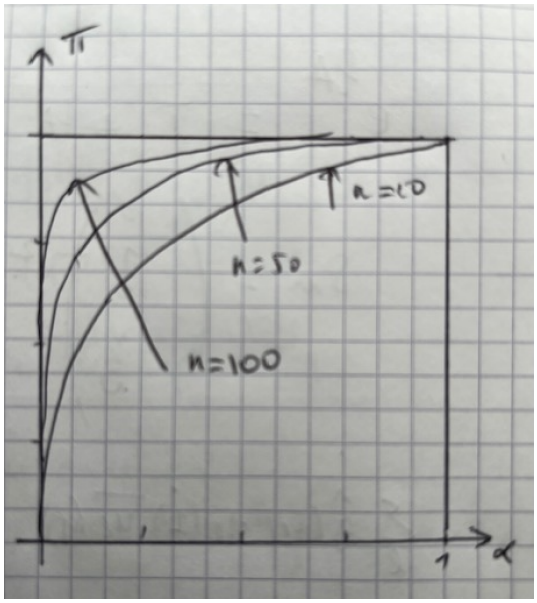

Figure 1: Allure des courbes COR pour différentes valeurs de n.

Pour les deux couples $(\lambda_0, \lambda_1) = (10, 1)$ et $(\lambda_0, \lambda_1) = (1000, 100)$, on a $\frac{\lambda_0}{\lambda_1} = 10$. Comme F est une fonction croissante (c'est une fonction de répartition), les performances du test sont donc d'autant meilleures que $\frac{\lambda_0 - \lambda_1}{\sqrt{\lambda_1}}$ est grand. Pour le premier couple $(\lambda_0, \lambda_1) = (10, 1)$, on a

$$\frac{\lambda_0 - \lambda_1}{\sqrt{\lambda_1}} = \frac{9}{1} = 9$$

tandis que pour $(\lambda_0, \lambda_1) = (1000, 100)$, on a

$$\frac{\lambda_0 - \lambda_1}{\sqrt{\lambda_1}} = \frac{900}{10} = 90.$$

Les performances du test seront donc meilleures pour $(\lambda_0, \lambda_1) = (1000, 100)$.

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	$f(x) = \frac{1}{2} e^{-\lvert x \rvert}, \quad x \in \mathbb{R}$	0	2	$\frac{1}{1+t^2}$
Normale univariée $\mathcal{N}\left(m, \sigma^2\right)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$	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 $c_{\lambda,\alpha}$	$f(x) = \frac{1}{\pi\lambda\left(1 + \left(\frac{x-\alpha}{\lambda}\right)^2\right)}$ $\lambda > 0,\ \alpha \in \mathbb{R}$	(-)	(-)	$e^{i\alpha t - \lambda\lvert t \rvert}$
Beta $B(a,b)$	$f(x) = k x^{a-1} (1-x)^{b-1}$ $k = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ $a > 0,\ b > 0$ $x \in]0,1[$ 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	$p_k = \frac{1}{n}$ $k \in \{1, ..., n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{e^{it}\left(1-e^{itn}\right)}{n\left(1-e^{it}\right)}$
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	$p_k = C_{n+k-1}^{n-1} p^n q^k$ $p \in [0,1] \quad q = 1-p$ $k \in \mathbb{N}$	$n\frac{q}{p}$	$n\frac{q}{p^2}$	$\left(\frac{p}{1-qe^{it}}\right)^n$
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_jq_j Covariance : $-np_jp_k$	$\left(\sum_{j=1}^m p_j e^{it}\right)^n$
Poisson $P(\lambda)$	$p_k = e^{-\lambda}\frac{\lambda^k}{k!}$ $\lambda > 0 \quad k \in \mathbb{N}$	λ	λ	$\exp\left[\lambda\left(e^{it}-1\right)\right]$
Géométrique	$p_k = pq^{k-1}$ $p \in [0,1] \quad q = 1-p$ $k \in \mathbb{N}^*$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{pe^{it}}{1-qe^{it}}$