STATISTIQUE MATHÉMATIQUE - MS ENSAI 2023-2024
EXERCICE 2.10. CORRIGÉ

On considère un n-échantillon $\mathrm{X} = (\mathrm{X}_1, ..., \mathrm{X}_n)$ suivant une loi de Poisson $\mathscr{P}(\lambda)$ de paramètre λ. On note $\mathrm{S} = \sum_{i=1}^n \mathrm{X}_i$ et $s = \sum_{i=1}^n x_i$.

1°. Préciser le modèle statistique et calculer la vraisemblance de l'échantillon.

Il s'agit d'un modèle statistique homogène et paramétrique. L'espace des obserations est $\mathbb{N}^n$ puisque chaque observation de l'échantillon est constituée d'un vecteur de valeurs entières (la loi de Poisson prend ses valeurs dans $\mathbb{N}$). Comme il s'agit d'une loi discrète, on peut choisir comme tribu sur $\mathbb{N}^n$ l'ensemble des parties de cet ensemble. Enfin, la famille de loi de probabilités est l'ensemble des lois produits de n variables aléatoires de poisson de paramètre λ. La vraisemblance est :

$$\mathrm{L}(x_1, ..., x_n, \lambda) = \prod_{i=1}^n \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) = e^{-ns} \frac{\lambda^s}{\prod_{i=1}^n x_i!} \tag{1}$$

si tous les x_i sont entiers, et 0 sinon (en notant $s = \sum_{i=1}^n x_i$).

2°. Montrer que S est une statistique exhaustive de l'échantillon pour le paramètre λ. Montrer qu'elle est complète pour λ.

Par soucis pédagogique, nous donnons trois techniques pour démontrer l'exhaustivité. La première est la plus longue et la plus calculatoire : on se ramène à la définition et il faut montrer que la loi conditionnelle de X sachant $[\mathrm{S} = s]$ ne dépend pas du paramètre. Pour toute observation $x = (x_1, ..., x_n)$, on a :

$$\mathbb{P}[\mathrm{X} = x | \mathrm{S} = s] = \frac{\mathbb{P}[\mathrm{S} = s | \mathrm{X} = x] \mathbb{P}[\mathrm{X} = x]}{\mathbb{P}[\mathrm{S} = s]} \tag{2}$$

d'après la formule de Bayes. $\mathbb{P}[\mathrm{S} = s | \mathrm{X} = x]$ vaut soit 1 soit 0 selon que $s = \sum x_i$ ou pas. Si $s \neq \sum x_i$, la probabilité est nulle et dans le cas contraire, elle vaut 1. Finalement, cette probabilité peut s'écrire sous la forme d'une indicatrice déterministe :

$$\mathbb{P}[\mathrm{S} = s | \mathrm{X} = x] = \mathbb{1}_{[s = \sum x_i]} \tag{3}$$

On a alors :

$$\mathbb{P}[\mathrm{X} = x | \mathrm{S} = s] = \frac{\mathbb{P}[\mathrm{X} = x]}{\mathbb{P}[\mathrm{S} = s]} \mathbb{1}_{[s = \sum x_i]} \tag{4}$$

$$= e^{-n\lambda} \frac{\lambda^s}{\prod_i (x_i!)} \frac{s!}{e^{-n\lambda} (n\lambda)^s} \mathbb{1}_{[s = \sum x_i]} \tag{5}$$

car S suit une loi de Poisson de paramètre $n\lambda$ (la somme de deux lois de Poisson indépendantes suit une loi de Poisson dont le paramètre est la somme des paramètres des deux lois initiales). En simplifiant, il vient :

$$\mathbb{P}[\mathrm{X} = x | \mathrm{S} = s] = \frac{s! n^{-s}}{x_1! ... x_n!} \mathbb{1}_{[s = \sum x_i]} \tag{6}$$

qui ne dépend pas de λ. Ceci prouve que S est exhaustive pour λ.

La seconde méthode utilise le théorème de factorisation de Neyman-Fisher :

$$\mathrm{L}(x, \lambda) = \left(\prod_{i=1}^n x_i! \right)^{-1} \times e^{s \ln \lambda - n\lambda} \tag{7}$$

Le premier terme du produit ne dépend que des observations x tandis que le second, à l'intérieur de l'exponentielle, ne dépend de x qu'au travers de la somme s.

La troisième méthode utilise le fait qu'un échantillon de Poisson fait partie du modèle exponentiel. Sa vraisemblance a déjà été calculée dans la formule précédente et montre que la statistique naturelle est S et le paramètre canonique est $\ln \lambda$. S est donc exhaustive.

Montrons maintenant que la statistique S est complète pour le paramètre λ. Ici encore, nous proposons deux démonstrations différentes. La plus simple et la plus rapide est d'utiliser le fait que le modèle est exponentiel. Nous savons déjà que la statistique naturelle est S. Par ailleurs, l'espace des paramètres est l'ensemble des valeurs prises par $\ln \lambda$ lorsque λ varie dans $]0, \infty[$. Ici, c'est simplement l'image de la fonction logarithme népérien, donc $\mathbb{R}$! Cet ensemble possède des ouverts ($\mathbb{R}$ est un ouvert lui-même) et l'on en déduit donc que la statistique est complète (pour λ).

La seconde méthode pour démontrer la complétude est de revenir à la définition. Il faut prouver que quelque soit la fonction $g(s)$ de $\mathbb{N}$ dans $\mathbb{R}$ (S est la somme des X_i, qui sont des entiers, elle prend donc ses valeurs dans $\mathbb{N}$),

$$\mathbb{E}[g(\mathrm{S})] = 0 \Rightarrow g \equiv 0, \ \forall \lambda > 0. \tag{8}$$

D'après le théorème de transfert,

$$\mathbb{E}_\lambda[g(s)] = \mathrm{F}(\lambda) = \sum_{s \geq 0} g(s) e^{-n\lambda} \frac{(n\lambda)^s}{s!} \tag{9}$$

En tant que fonction de λ, il s'agit d'une somme de série entière dont le rayon de convergence est infini. Elle est donc définie et de classe C^∞ sur $\mathbb{R}$ tout entier et cette fonction est identiquement nulle (pour tout λ) si, et seulement si, tous ses coefficients sont nuls. On doit donc avoir, pour tout $\lambda > 0$ et pour tout s entier,

$$g(s) e^{-n\lambda} \frac{(n\lambda)^s}{s!} = 0 \tag{10}$$

L'exponentielle et la fonction puissance ne sont pas nulles pour tout s, il faut donc que $g(s) = 0$ pour tout s, c'est à dire que g doit être identiquement nulle.

3°. Déduire des questions précédentes un estimateur sans biais de variance minimale (VUMSB) du paramètre λ.

Nous venons de voir que S est une statistique complète et $\overline{\mathrm{X}} = \mathrm{S}/n$ est clairement un estimateur

sans biais de λ. D'après le théorème de Lehmann-Scheffé, l'estimateur

$$S^* = \mathbb{E}[\overline{X}|S] = \overline{X} \tag{11}$$

est sans biais et de variance miminale. L'espérance conditionnelle est égale à $\overline{X}$ car cette v.a. est mesurable par rapport à S.

4°. Le modèle est-il régulier ? Si oui, calculer l'information $\mathbb{I}_X(\lambda)$ au sens de Fisher et en déduire un estimateur efficace.

Le modèle est régulier si, et seulement si, il est dominé, homogène, Θ est ouvert, $L(x,\theta)$ est de classe C^2 en θ pour tout x et deux fois dérivable sous le signe somme. Ici, le modèle est dominé par la mesure de comptage (la loi est discrète), il est homogène car le support est $\mathbb{N}^n$ et ne dépend pas du paramètre. Θ est l'ensemble dans lequel varie le paramètre et dans l'exercice il est ouvert (c'est $]0,\infty[$). En tant que fonction de λ, la vraisemblance est clairement de classe C^2 (c'est une composée et un produit de puissance et d'exponentielle). On admet la propriété de dérivation sous le signe somme. Donc le modèle est régulier et l'on peut calculer l'information de Fisher de l'échantillon :

$$\mathbb{I}_X(\lambda) = \mathbb{V}(S(X,\lambda)) = \mathbb{V}\left(\frac{\partial l}{\partial \lambda}(X,\lambda)\right) \tag{12}$$

$$= \mathbb{V}\left(\frac{\partial l}{\partial \lambda}(\ln\alpha + S\ln\lambda - n\lambda)\right) \tag{13}$$

$$= \mathbb{V}\left(\frac{S}{\lambda} - n\right) = \frac{1}{\lambda^2} \times \lambda \times n = \frac{n}{\lambda} \tag{14}$$

avec $\alpha = (\prod x_i!)^{-1}$. Puisque $V(\overline{X}) = \lambda/n = \mathbb{I}_X(\lambda)^{-1}$, l'estimateur atteint la borne de Cramer-Rao : il est efficace.

5°. On s'intéresse maintenant au paramètre $\theta = e^{-\lambda}$. Quelle est la signification de θ ? Démontrer que $\widehat{\theta}_1 = \exp(\overline{X})$ est l'estimateur du maximum de vraisemblance de θ et qu'il est biaisé.

$\mathbb{P}[X_1 = 0] = e^{-\lambda} = \theta$. La loi de Poisson représente le nombre d'apparitions d'un phénomène aléatoire durant un laps de temps donné. θ représente donc la probabilité de ne pas voir le phénomène apparaître durant ce laps de temps.

On sait que $\overline{X}$ est l'EMV de λ. D'après le théorème de reparamétrisation, $\widehat{\theta}_1 = \exp(\overline{X})$ est l'EMV de $\theta = e^{-\lambda}$. Déterminons son espérance :

$$\mathbb{E}[\widehat{\theta}_1] = \mathbb{E}[e^{S/n}] = \prod_{i=1}^{n} \mathbb{E}[e^{-X_i/n}] = \mathbb{E}[e^{-X_1/n}]^n \tag{15}$$

par indépendance et identique distribution des X_i. Mais d'après le théorème de transfert,

$$\mathbb{E}[e^{-X_1/n}] = \sum_{k=0}^{+\infty} e^{-k/n} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{+\infty} \frac{(e^{-1/n}\lambda)^k}{k!} \tag{16}$$

$$= e^{-\lambda + e^{-1/n}\lambda} \tag{17}$$

Ainsi,

$$\mathbb{E}[\widehat{\theta}_1] = e^{-\lambda} e^{\lambda e^{-1/n}} \neq e^{-\lambda} \tag{18}$$

l'EMV $\widehat{\theta}_1$ est donc un estimateur biaisé.

6°. Soient $Y_i = \mathbb{1}_{[X_i=0]}$. Montrer que Y_1 est un estimateur des moments de θ et qu'il est non biaisé.

Y_1 est une fonction mesurable des seules observations X_i, c'est donc un estimateur de θ. Par ailleurs, $\mathbb{E}[Y_1] = \mathbb{P}[X_1 = 0] = \theta$ et donc Y_1 est sans biais.

7°. Déterminer la loi conditionnelle de Y_1 sachant S. En déduire l'estimateur VUSMB de θ.

$$\mathbb{E}[Y_1|S=k] = \mathbb{P}[X_1 = 0|S=k] = \frac{\mathbb{P}([X_1=0] \cap [S=k])}{P[S=k]} \tag{19}$$

Si $[X_1 = 0]$, on a $S = \sum_{i=1}^n X_i = \sum_{i=2}^n X_i$ qui est alors indépendant de X_1. On peut donc écrire :

$$\mathbb{E}[Y_1|S=k] = \frac{\mathbb{P}[X_1=0] \times \mathbb{P}[\sum_{i=2}^n X_i = k]}{P[S=k]} \tag{20}$$

$\sum_{i=2}^n X_i$ est la somme de $n-1$ v.a. indépendantes de loi de Poisson de paramètre λ. Elle suit donc une loi de Poisson de paramètre $(n-1)\lambda$. De la même façon, au dénominateur, S suit une loi de Poisson de paramètre $n\lambda$. Ainsi,

$$\mathbb{E}[Y_1|S=k] = \frac{e^{-\lambda}e^{-(n-1)\lambda}((n-1)\lambda)^k/k!}{e^{-n\lambda}(n\lambda)^k/k!} \tag{21}$$

$$= \left(\frac{n-1}{n}\right)^k = \left(1 - \frac{1}{n}\right)^k \tag{22}$$

L'espérance conditionnelle de Y_1 sachant S est donc $(1-1/n)^S$.

Y_1 est un estimateur sans biais de θ et S est une statistique complète. On est donc tenté de conclure que, d'après le théorème de Lehmann-Scheffé, $\mathbb{E}[Y_1|S]$ est l'estimateur VUMSB de θ. Mais on ne peut pas faire cela. En effet, S est une statistique complète pour λ et pour conclure, nous avons besoin d'une statistique complète pour θ ! Il faut donc démontrer que S est aussi complète pour θ. Pour faire cela, il faut reécrire la vraisemblance en fonction de θ, puis redémontrer la complétude à l'aide de la définition. En passant sur les calculs, on obtient :

$$L(x_1,...,x_n,\theta) = \frac{(\ln\theta)^s \theta^n}{\prod x_i!} \tag{23}$$

et le théorème de factorisation de Neyman-Fischer permet de conclure. De la même façon que précédemment, si g est une fonction mesurable de S telle que

$$\mathbb{E}[g(S)] = 0, \ \forall \theta \in]0,1[, \tag{24}$$

alors $g \equiv 0$ (la somme est une série entière nulle, donc tous ses coefs sont nuls).

S est donc exhaustive et complète pour θ et d'après le théorème de Lehmann-Scheffé, l'estimateur VUMSB de θ est donc :

$$\widehat{\theta}_2 = \left(1 - \frac{1}{n}\right)^S \tag{25}$$

8°. L'estimateur $\widehat{\theta}_2$ est-il efficace ?

Il faut calculer la borne de Cramer-Rao pour $\theta = e^{-\lambda} = \psi(\lambda)$. On connaît déjà $\mathbb{I}(\lambda)$ et les propriétés de reparamétrisation de l'information au sens de Fischer permettent d'écrire :

$$\mathbb{I}(\theta) = \psi'(\lambda)^{-1}\mathbb{I}(\lambda)\psi'(\lambda)^{-1} = e^{2\lambda}\frac{n}{\lambda} \tag{26}$$

Il faut comparer ce résultat à la variance de l'estimateur. Pour conduire le calcul, nous aurons besoin de la formule suivante :

$$\mathbb{E}[e^{tS}] = e^{-\lambda(1-e^t)} \tag{27}$$

On a :

$$\mathbb{E}[\widehat{\theta}_2^2] = \mathbb{E}\left[\left(\frac{n-1}{n}\right)^{2S}\right] = \mathbb{E}\left[e^{2S\ln(\frac{n-1}{n})}\right] \tag{28}$$

$$= e^{n\lambda(e^{2\ln((n-1)/n)}-1)} = e^{n\lambda(-2n+1)/n^2} = e^{-2\lambda+\lambda/n} \tag{29}$$

Par ailleurs,

$$\mathbb{E}[\widehat{\theta}_2] = e^{-\lambda} \tag{30}$$

On en déduit que

$$\mathbb{V}(\widehat{\theta}_2) = e^{-2\lambda}(e^{\lambda/n} - 1) \tag{31}$$

Cette expression est différente de $\mathbb{I}(\theta)^{-1} = e^{-2\lambda}\frac{\lambda}{n}$ et l'estimateur n'est donc pas efficace. Il est par contre asymptotiquement efficace car les deux expressions sont équivalentes en l'infini.

9°. On considère maintenant l'estimateur $\widehat{\theta}_3 = \overline{Y}$, avec $\overline{Y}$ la moyenne arithmétique des Y_i. Démontrer qu'il est VUMSB et efficace pour θ.

$\overline{Y}$ est un estimateur sans biais de θ et puisque Y_i suit une loi de Benoulli de paramètre θ, le modèle est exponentiel et la moyenne empirique est une statistique exhaustive. On sait également qu'elle est complète pour θ. On peut donc appliquer le théorème de Lehmann-Scheffé et conclure que :

$$\overline{Y} = \mathbb{E}[\overline{Y}|\overline{Y}] \tag{32}$$

est VUMSB. C'est un estimateur efficace car l'inverse de l'information de Fisher est égale à la variance θ/n de $\overline{Y}$.