
Sujet blanc - Estimation et test statistique Correction

Exercice 1 Estimation d'un paramètre pour une loi discrète

Soit X une variable aléatoire discrète dont la loi de probabilité est donnée par pour tout $x \in \mathbb{N}^*$

$$f(x;\theta) = \mathbb{P}_\theta(X = x) = \frac{\theta^{x-1}}{(\theta+1)^x}$$

où $\theta > 0$ est un paramètre inconnu. On dispose d'un échantillon $X_1, ..., X_n$ *i.i.d* de loi parente X et on souhaite déterminer un estimateur de θ.

1. Écrire le modèle statistique de l'échantillon considéré.

Nos observations sont indépendantes, donc le modèle statistique s'écrit $(\mathcal{H}^n, (\mathbb{P}_\theta^{\otimes n})_{\theta>0})$ avec

$$\mathcal{H} = \mathbb{N}^* \text{ et } \mathbb{P}_\theta \text{ donnée par l'énoncée.}$$

2. Vérifier que X suit une loi géométrique de paramètre $p = \frac{1}{1+\theta}$. En déduire l'espérance et la variance de X.

On sait que l'espérance d'une loi géométrique de paramètre p est $1/p$ et sa variance est $(1-p)/p^2$. Donc en remplaçant p par $1/(1+\theta)$, on a

$$\mathbb{E}_\theta[X] = \theta + 1 \text{ et } \mathrm{Var}_\theta(X) = \theta(\theta+1).$$

3. Déterminer $\hat{\theta}_{MM}$ un estimateur de θ par la méthode des moments en considérant le moment d'ordre 1.

On doit résoudre l'équation

$$\bar{X}_n = \hat{\theta}_{MM} + 1,$$

d'où $\hat{\theta}_{MM} = \bar{X}_n - 1$.

4. Étudier le biais de $\hat{\theta}_{MM}$.

L'estimateur n'est pas biaisé (linéarité de l'espérance).

5. Vérifier que $\hat{\theta}$ est consistant.

La loi X admet bien un moment d'ordre 1 et nos observations sont indépendantes, donc d'après la LGN, on a

$$\bar{X}_n \overset{p.s.}{\to} \mathbb{E}_\theta[X] = \theta + 1.$$

Comme la soustraction est une opération continue, on arrive bien à la consistance de $\hat{\theta}_{MM}$.

6. Montrer la normalité asymptotique de l'estimateur $\hat{\theta}_{MM}$ de θ.

La loi X vérifie bien les hypothèses du TCL, donc

$$\sqrt{n}\frac{\bar{X}_n - (\theta+1)}{\sqrt{\theta(\theta+1)}} \rightsquigarrow \mathcal{N}(0,1),$$

et comme $\hat{\theta}_{MM} = \bar{X}_n - 1$, on a bien

$$\sqrt{n}\frac{\hat{\theta}_{MM} - \theta}{\sqrt{\theta(\theta+1)}} \rightsquigarrow \mathcal{N}(0,1).$$

7. Déterminer un intervalle de confiance asymptotique de θ de niveau de confiance $1-\alpha$.
Pour $\alpha > 0$, on a d'après la définition de la convergence en loi

$$Q_\theta\left(\sqrt{n}\frac{\hat{\theta}_{MM} - \theta}{\sqrt{\theta(\theta+1)}} \in [-q_{1-\alpha/2}, +q_{1-\alpha/2}]\right) \to 1-\alpha.$$

Avant d'inverser l'intervalle, on peut remplacer le terme de variance $\theta(\theta+1)$ par la variance empirique ou par $\hat{\theta}_{MM}(1+\hat{\theta}_{MM})$ d'après le lemme de Slutsky, donc

$$\underbrace{\frac{\sqrt{\theta(\theta+1)}}{\sqrt{\hat{\theta}_{MM}(\hat{\theta}_{MM}+1)}}}_{\to 1} \times \sqrt{n}\frac{\hat{\theta}_{MM} - \theta}{\sqrt{\theta(\theta+1)}} \rightsquigarrow \mathcal{N}(0,1).$$

Nous avons donc par la définition de la convergence en loi

$$Q_\theta\left(\sqrt{n}\frac{\hat{\theta}_{MM} - \theta}{\sqrt{\hat{\theta}_{MM}(\hat{\theta}_{MM}+1)}} \in [-q_{1-\alpha/2}, +q_{1-\alpha/2}]\right) \to 1-\alpha.$$

En inversant l'intervalle on trouve que

$$Q_\theta\left(\left[\hat{\theta}_{MM} \pm \frac{\sqrt{\hat{\theta}_{MM}(\hat{\theta}_{MM}+1)}}{\sqrt{n}}\right] \ni \theta\right) \to 1-\alpha.$$

8. Écrire la fonction vraisemblance L_n associée à l'échantillon considéré.
Comme les observations sont indépendantes, on a pour $x_1, \dots, x_n \in \mathbb{N}^*$ et $\theta > 0$,

$$L_n(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i;\theta) = \frac{\theta^{\sum_{i=1}^n x_i - n}}{(\theta+1)^{\sum_{i=1}^n x_i}}.$$

9. Montrer que pour tout $x_1, \dots, x_n \in \mathbb{N}$ et $\theta > 0$

$$\frac{\partial}{\partial\theta}(\log L_n(x_1, \dots, x_n; \theta)) = \frac{n(\bar{x} - 1 - \theta)}{\theta(\theta+1)}$$

où $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i$.
En passant au logarithme, on a

$$\log L_n(x_1, \dots, x_n; \theta) = \left(\sum_{i=1}^n x_i - n\right)\log\theta - \sum_{i=1}^n x_i \log(1+\theta) = n(\bar{x}-1)\log\theta - n\bar{x}\log(1+\theta).$$

Donc en dérivant par rapport à θ, on obtient

$$\partial_\theta \log L_n(x_1, \dots, x_n; \theta) = \frac{n(\bar{x}-1)}{\theta} - \frac{n\bar{x}}{1+\theta} = \frac{n(\bar{x}-1-\theta)}{\theta(\theta+1)}.$$

10. En déduire que $\hat{\theta}_{MV}$ l'estimateur de maximum de vraisemblance de θ est identique à $\hat{\theta}_{MM}$.
On voit bien que la dérivée s'annule pour $\theta = \bar{x} - 1$. Il reste à vérifier en faisant une étude du signe de la dérivée qu'il s'agit bien d'un maximum.
On définit $\hat{\theta} = \hat{\theta}_{MM} = \hat{\theta}_{MV}$.
11. Calculer la quantité d'information de Fisher du modèle de l'échantillon. $\hat{\theta}$ est-il efficace ?
Pour déterminer l'information de Fisher, il faut calculer la variance de la log −vraisemblance, donc pour $\theta > 0$,

$$\mathrm{Var}_\theta(\partial_\theta \log L_n(X_1, \dots, X_n; \theta)) = \mathrm{Var}_\theta\left(\frac{\sum_{i=1}^n X_i - n - n\theta}{\theta(\theta+1)}\right) = \frac{n\mathrm{Var}_\theta(X)}{\theta^2(\theta+1)^2},$$

car nos observations sont indépendantes. Nous avons déjà calculé la variance de X, d'où $I_n(\theta) = \frac{n}{\theta(\theta+1)}$. Il reste à déterminer le risque quadratique de l'estimateur $\hat{\theta}_{MM}$, on a d'après la décomposition biais/variance

$$\mathcal{R}(\theta, \hat{\theta}_{MM}) = b_n(\theta)^2 + \mathrm{Var}_\theta(\hat{\theta}_{MM}).$$

Nous avons déjà dit que l'estimateur était non biaisé et sa variance se calcule aisément. On trouve que le risque de $\hat{\theta}_{MM}$ est l'inverse de l'information de Fisher, donc le risque de l'estimateur atteint la borne de Cramer-Rao, il est donc efficace.

Exercice 2 Réussite d'un examen

On souhaite s'assurer que les élèves aient bien compris le cours de Statistique, i.e. que leurs notes soient centrées autour d'une note **strictement supérieure à 10**. Pour simplifier la modélisation, on suppose que la note d'un élève suit la loi $\mathcal{N}(m, \sigma^2)$ et que les différentes copies sont indépendantes. On cherche à tester la valeur de m lorsque la variance σ^2 est **inconnue**
1. Choisir une alternative de l'hypothèse nulle

$$(H_0) : "m = 10".$$

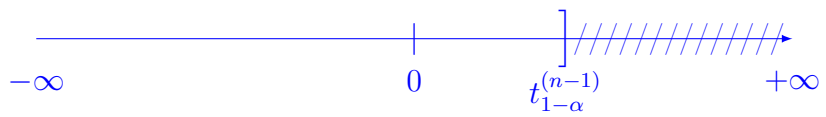
L'hypothèse (H_1) est l'hypothèse que l'on essaye de vérifier, on mettra donc

$$(H_1) : "m > 10".$$

Ainsi, faute de preuve suffisante, on supposera que $m = 10$ (ou $m \leqslant 10$), et donc l'enseignant doit continuer de travailler pour amener ses élèves à la réussite.
2. Proposer une procédure de test pour tester les hypothèses précédentes. Bien justifier le choix de la zone de rejet et le comportement **non asymptotique** sous (H_0) de la statistique

$$T_n = \sqrt{n} \times \frac{\bar{X}_n - 10}{\sqrt{S_n^2}}.$$

D'après le cours, on a $T_n \sim \mathcal{T}(n-1)$. De plus, sous (H_0), la statistique T_n aura tendance à être autour de 0. Sous (H_1), la statistique T_n aura tendance à tendre vers $+\infty$, donc nous allons mettre une zone de rejet plutôt à droite de 0.



Sous l'hypothèse (H_0), on a

$$Q_{10}(T_n \leqslant t_{1-\alpha^{(n-1)}}) = 1 - \alpha,$$

donc la procédure suivante définit bien un test de niveau α

- on rejette (H_0), si l'observation de T_n est strictement supérieur à $t_{1-\alpha}^{(n-1)}$, ceci est équivalent à rejeter si

$$\bar{X}_n > 10 + t_{1-\alpha}^{(n-1)} \frac{\sqrt{S_n^2}}{\sqrt{n}}$$

- on ne rejette pas sinon.

3. Avec un échantillon de taille $n = 23$, la moyenne du CC2 est de 10.2 pour un écart-type de 2.7. On observe ici $T_n = 0.36$, quelle est la conclusion de votre test avec un seuil de 5% ?

Le quantile d'ordre $1 - \alpha$ de $\mathcal{T}(23 - 1)$ est 1.717, donc T_n ne se situe pas dans la zone de rejet. Ainsi, on ne rejette pas l'hypothèse null.