TD n° 3 : Estimation et tests

Exercice 1. (ADN et test d'adéquation à une loi uniforme)

Le tableau ci-dessous représente le nombre observé de palindromes pour 10 segments de l'ADN du CMV. Ces données suivent-elles une loi uniforme ? Décrivez votre démarche pour conclure.

Segment	1	2	3	4	5	6	7	8	9	10	Total
Effectifs	29	21	32	30	32	31	28	32	34	27	296

Remarque : une sortie numérique donne comme statistique associée $T_{obs} \approx 4.14$ et comme p-valeur approximative 0.90.

Correction:

Déjà rappeler pourquoi on s'intéresse à la loi uniforme — *cf.* cours sur le processus de Poisson. Pour le test d'adéquation on refait le tableau. La statistique observée est alors

Segment	Eff. obs	Eff. théo.
1	29	29.6
2	21	29.6
3	32	29.6
4	30	29.6
5	32	29.6
6	31	29.6
7	28	29.6
8	32	29.6
9	34	29.6
10	27	29.6

Table 1 – Tableau nécessaire pour le test d'adéquation du χ^2.

$$T_{obs} = \frac{(29-29.6)^2}{29.6} + \ldots + \frac{(27-29.6)^2}{29.6} \approx 4.14,$$

or sous H_0 (la distribution est uniforme dans chaque case) la statistique suit une loi du χ^2_{10-1} (on n'a pas à estimer de paramètres). On trouve numériquement que la p-valeur est approximativement 0.90. On accepte donc H_0 au profit de H_1, ce qui veut dire concrètement pour notre exemple que le processus de Poisson semble bien modéliser globalement la position des palindromes dans l'ADN.

Exercice 2. (Jeux vidéos)

Sur 91 étudiants ayant participé à un sondage sur les jeux vidéos, on a relevé les résultats suivants :

Note	A	B	C	D	F	Total
Effectifs	31	52	8	0	0	91

Ces observations sont-elles en adéquation avec la distribution 20% de A, 30% B, 40% C et 10% de D et F ?

Remarque numérique : la p-value associée est $1.6 \cdot 10^{-13}$.

Correction:

Clairement on nous demande de faire un test d'adéquation du χ^2. Notons déjà que nous n'avons pas de paramètres à estimer puisque tout nous est donné. On construit donc le tableau "effectifs observées et effectifs théoriques". On vérifie que les effectifs théoriques sont raisonnable en terme de taille (*e.g.*, > 5)

Note	Eff. Obs	Eff. théo.
A	31	18.2
B	52	27.3
C	8	36.4
D-F	0	9.1

TABLE 2 – Tableau nécessaire au test d'adéquation du χ^2.

et la statistique vaut alors

$$T_{obs} = \frac{(31-18.2)^2}{18.2} + \ldots + \frac{(0-9.1)^2}{9.1} \approx 62.60.$$

Sous H_0 la statistique de test suit un χ^2_{4-1}, et la p-value associée est $1.6 \cdot 10^{-13}$. On rejette clairement H_0 au profit de H_1. Les observations ne semblent pas être en adéquation avec ce qui avait été annoncé.

EXERCICE 3. (Déterministe vs. aléatoire)

Pour un processus de Poisson homogène de taux λ par heure, montrez que le nombre d'occurrences sur deux intervalles disjoints de 1 heure chacun suit une loi de Poisson de paramètre 2λ.

Aide :

$$\mathbb{P}[n \text{ occurrences en deux heures}] = \sum_{k=0}^{n} \mathbb{P}[k \text{ occurrences la } 1^{\text{re}} \text{ heure}, n-k \text{ occurrences la } 2^{\text{nde}}],$$

et

$$\sum_{k=0}^{n} \frac{n!}{k!(n-k)!} = 2^n.$$

Correction:

L'aide permet de voir que

$$\begin{aligned}
\mathbb{P}[n \text{ occ. en 2h}] &= \sum_{k=0}^{n} \mathbb{P}[k \text{ occ. } 1^{\text{re}} \text{ heure}, n-k \text{ occ. la } 2^{\text{nde}}] \\
&\overset{\text{ind}}{=} \sum_{k=0}^{n} \mathbb{P}[k \text{ occ. la } 1^{\text{re}} \text{ heure}] \cdot \mathbb{P}[n-k \text{ occ. la } 2^{\text{nde}} \text{ heure}] \\
&= \sum_{k=0}^{n} \frac{\lambda^k}{k!} e^{-\lambda} \frac{\lambda^{n-k}}{(n-k)!} e^{-\lambda} \\
&= e^{-2\lambda} \sum_{k=0}^{n} \frac{\lambda^n}{k!(n-k)!} \\
&= e^{-2\lambda} \sum_{k=0}^{n} \frac{\lambda^n n!}{k!(n-k)!n!} \\
&= e^{-2\lambda} \sum_{k=0}^{n} \frac{\lambda^n}{n!} \frac{n!}{k!(n-k)!} \\
&= e^{-2\lambda} \frac{\lambda^n}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \\
&= e^{-2\lambda} \frac{\lambda^n}{n!} 2^n \\
&= \frac{(2\lambda)^n}{n!} e^{-2\lambda}.
\end{aligned}$$

Bien insister sur cette propriété car le processus de Poisson est largement utilisé.

EXERCICE 4. (Méthode des moments - Loi uniforme)

Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}(0,\theta)$ de loi commune la loi uniforme sur l'intervalle $[0,\theta]$ (avec $\theta > 0$).

a) Donner f, la densité de la loi de X_1, et calculer son espérance.

b) Proposez un estimateur de θ par la méthode des moments.

c) Trouvez l'estimateur du maximum de vraisemblance de θ.

d) Calculez l'erreur quadratique moyenne pour ces deux estimateurs. Discutez.

Correction:

a)

$$f(x) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x) = \begin{cases} \frac{1}{\theta} & \text{si} x \in [0,\theta], \\ 0 & \text{sinon.} \end{cases}$$

De plus is $X \sim \mathcal{U}(0,\theta)$:

$$\mathbb{E}[X] = \frac{1}{\theta}\int_0^\theta x dx = \frac{1}{\theta}\left[\frac{x^2}{2}\right]_{x=0}^{x=\theta} = \frac{\theta}{2}$$

b) On a un paramètre donc il faut une équation des moments : pour cela noter que

$$\bar{X}_n = \mathbb{E}[X] = \frac{\theta}{2} \Longrightarrow \hat{\theta}^{\text{moment}} = 2\bar{X}_n.$$

c) La vraisemblance s'écrit

$$L(\theta; X_1, \dots, X_n) = \prod_{i=1}^n \theta^{-1}\mathbb{1}_{\{0 \leqslant X_i \leqslant \theta\}},$$

et l'estimateur du maximum de vraisemblance noté $\hat{\theta}_{\text{MLE}}$ maximise cette quantité. Puisque $\theta > 0$, θ^{-1} l'est tout autant. Or s'il existe au moins un $X_i > \theta$, la vraisemblance s'annule. De même si elle un des X_i est négatif. Mais si $X_i \in [0,\theta]$ pour tout $i = 1, \dots, n$, alors la vraisemblance vaut $\theta^{-n} > 0$. On conclut en notant que l'application $x \mapsto x^{-n}$ définie sur $\mathbb{R}_*^+$ est décroissante. Et donc que

$$\hat{\theta}_{\text{MLE}} = \max\{X_1, \dots, X_n\} \ .$$

Rem: dessiner au besoin la vraisemblance pour un cas avec $n = 3$ ou 4 observations. Montrer qu'avant $\max\{X_1, \dots, X_n\}$ la vraisemblance vaut zéro, et qu'après c'est $\frac{1}{\theta^n}$.

d) L'erreur quadratique d'un estimateur $\hat{\theta}$ de θ est donnée par $\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{B}(\hat{\theta})^2 + \mathbb{V}\text{ar}(\hat{\theta})$. On a donc

$$\mathbb{B}(\hat{\theta}^{\text{moment}}) = \mathbb{E}[\hat{\theta}^{\text{moment}}] - \theta = 2\mathbb{E}[\bar{X}_n] - \theta = \theta - \theta = 0$$

$$\mathbb{V}\text{ar}[2\bar{X}_n] = 4\,\mathbb{V}\text{ar}[\bar{X}_n] = \frac{4}{n}\,\mathbb{V}\text{ar}[X_1] = \frac{4}{n}\frac{(\theta-0)^2}{12} = \frac{\theta^2}{3n} \ ,$$

car $\mathbb{V}\text{ar}(X_1) = \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = \frac{1}{\theta}\int_0^\theta x^2 dx - \theta^2/4 = \frac{1}{\theta}\left[x^3/3\right]_{x=0}^{x=\theta} - \theta^2/4 = \theta^2(\frac{1}{3} - \frac{1}{4}) = \theta^2\frac{1}{12}$ Ainsi

$$\text{MSE}(\hat{\theta}^{\text{moment}}) = \frac{\theta^2}{3n} \ .$$

Pour le maximum de vraisemblance c'est un peu plus dur. Posons $M_n = \max\{X_1, \dots, X_n\}$. On a alors clairement

$$\mathbb{P}[M_n \leqslant x] = \left(\frac{x}{\theta}\right)^n, \qquad 0 \leqslant x \leqslant \theta \ ,$$

d'où découle la densité de M_n qui vaut

$$f(x) = \frac{n}{\theta^n}x^{n-1}, \qquad 0 \leqslant x \leqslant \theta \ .$$

On peut alors calculer les deux premiers moments

$$\mathbb{E}[M_n] = \frac{n}{\theta^n}\int_0^\theta x^n\,dx = \frac{n}{n+1}\theta$$

$$\mathbb{E}[M_n^2] = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,dx = \frac{n}{n+2}\theta^2,$$

et donc le biais et la variance valent

$$\mathbb{B}[\hat{\theta}_{\text{MLE}}] = \mathbb{E}[M_n] - \theta = -\frac{\theta}{n+1}$$
$$\mathbb{V}\text{ar}[\hat{\theta}_{\text{MLE}}] = \frac{n\theta^2}{(n+1)^2(n+2)} \ ,$$

et donc

$$\begin{aligned}
\text{MSE}(\hat{\theta}_{\text{MLE}}) &= \mathbb{B}(\hat{\theta}_{\text{MLE}})^2 + \mathbb{V}\text{ar}[\hat{\theta}_{\text{MLE}}] \\
&= \frac{\theta^2}{(n+1)^2} + \frac{n\theta^2}{(n+1)^2(n+2)} \\
&= \frac{2(n+1)\theta^2}{(n+1)^2(n+2)} \\
&= \frac{2\theta^2}{(n+1)(n+2)} \ .
\end{aligned}$$

On voit donc que l'erreur quadratique pour le MLE sera plus petite dès lors que n est suffisamment grand. C'est donc un meilleur estimateur !

Exercice 5. (Maximum de vraisemblance - loi de Pareto)

Trouvez l'estimateur du maximum de vraisemblance pour $\theta > 0$ à partir de n observations $X_1, \ldots, X_n$ suivant une loi de Pareto

$$f(x;\theta) = \theta \mu^\theta x^{-\theta-1}, \qquad x \geqslant \mu,$$

où $\mu > 0$ est inconnu.

Correction:

La log vraisemblance s'écrit

$$\begin{aligned}
\ell(\theta; X_1, \ldots, X_n) &= \sum_{i=1}^{n} \{\log(\theta) + \theta \log(\mu) - (\theta+1)\log(X_i)\} \\
&= n\log(\theta) + n\theta\log(\mu) - (\theta+1)\sum_{i=1}^{n}\log(X_i).
\end{aligned}$$

L'estimateur du maximum de vraisemblance est solution de

$$\begin{aligned}
\frac{\partial \ell(\theta; X_1, \ldots, X_n)}{\partial \theta} = 0 &\iff \frac{n}{\theta} + n\log(\mu) - \sum_{i=1}^{n}\log(X_i) = 0 \\
&\iff \hat{\theta}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n}\log(X_i) - n\log(\mu)}
\end{aligned}$$

Note : il est bon de vérifier que le dénominateur précédent est différent de zéro.