

On considèrera dans les exercices un risque de première espèce de 5%.

Exercice 1**4 points**

Soit X une variable aléatoire normale d'espérance 0 et de variance σ^2 inconnue. On dispose d'un échantillon i.i.d. (X_1, \dots, X_n) de v.a. parente X . On notera T la statistique $\sum_{i=1}^n X_i^2$ et F_{χ^2} la fonction de répartition de la loi du χ^2 à ν degrés de liberté. On veut effectuer le test suivant :

$$\begin{aligned} H_0 : \sigma^2 &= \sigma_0^2 \\ H_1 : \sigma^2 &= \sigma_1^2 (> \sigma_0^2). \end{aligned}$$

1. Montrer que la région critique optimale s'exprime en fonction de la statistique T .

$$\begin{aligned} \frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} > k &\Leftrightarrow -\frac{\sum_{i=1}^n X_i^2}{2\sigma_1^2} + \frac{\sum_{i=1}^n X_i^2}{2\sigma_0^2} > \log k + \log 2\pi\sigma_1 - \log 2\pi\sigma_0 \\ &\Leftrightarrow \sum_{i=1}^n X_i^2 > k' \end{aligned}$$

2. Sous H_0 , quelle est la loi de la variable aléatoire

$$Z = \frac{T}{\sigma_0^2}$$

Z suit une loi du chi2 à $\nu = n$ degrés de liberté

3. Déterminer cette région critique en fonction du quantile de la loi convenable.

$$P(\sum_{i=1}^n X_i^2 > k') = \alpha \Leftrightarrow P(Z > \frac{k'}{\sigma_0^2}) = \alpha$$

on a donc $\frac{k'}{\sigma_0^2} = 1 - \text{icd}(\chi_\nu^2, 1 - \alpha)$ et donc $k' = \sigma_0^2 (1 - \text{icd}(\chi_\nu^2, 1 - \alpha))$

4. Calculer la puissance du test en fonction du quantile de la loi convenable.

$$1 - \beta = P(\sum_{i=1}^n X_i^2 \leq k' \mid \sigma^2 = \sigma_1^2) = P(Z \leq \frac{k'}{\sigma_1^2})$$

5. Déterminer la région critique et la puissance du test pour $n = 10$, $\sigma_0^2 = 1$ et $\sigma_1^2 = 2$.

$$k' = 1 - \text{icd}(\chi_{10}^2, 0.95) = 18,3$$

et

$$1 - \beta = P(Z \leq \frac{18,3}{2}) = P(Z \leq 4.57) = 0,982$$

Exercice 2**4 points**

Soit X_1, \dots, X_n un échantillon i.i.d. dont la v.a. parente X est une v.a. discrète de loi de probabilité

$$p(x) = P(X = x) = \frac{\theta^x}{(1 + \theta)^{x+1}} \text{ pour } x \in \mathbb{N}$$

où θ est un paramètre réel strictement supérieur à 0.

1. Montrer que

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

où $L(x_1, \dots, x_n; \theta)$ est la fonction de vraisemblance de l'échantillon et $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ la moyenne de l'échantillon.

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

et

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

d'où

$$\frac{\partial \log L(x_1, \dots, x_n; \theta)}{\partial \theta} = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{\sum_{i=1}^n x_i + n}{1+\theta} = \frac{(1+\theta) \sum_{i=1}^n x_i - \theta \sum_{i=1}^n x_i - n}{\theta(1+\theta)}$$

2. En déduire l'estimateur du maximum de vraisemblance de θ , que l'on notera $\hat{\theta}$.

$$\hat{\theta} = \bar{x}$$

3. Cet estimateur est-il efficace? Donner son espérance et sa variance. En déduire l'espérance et la variance de X .

On a

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

$\hat{\theta}$ est donc efficace et sa variance est $\frac{\theta(\theta+1)}{n}$.

4. En supposant que n est grand, déterminer un intervalle de confiance bilatéral symétrique approché pour θ au niveau de confiance $1 - \alpha$.

si n est grand on peut prendre une approximation normale pour l'estimateur du max de vraisemblance et supposer que

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\theta(\theta+1)}{n}\right)$$

on a donc

$$P\left(u_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(\theta+1)}{n}}} < -u_{\alpha/2}\right) = 1 - \alpha$$

ce qui suggère d'utiliser comme fonction pivotale

$$u_{\alpha/2} = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(\theta+1)}{n}}}$$

les bornes sont alors données par les racines du polynôme

$$p(\theta) = (u^2 - n)\theta^2 + (u^2 + 2n\hat{\theta})\theta - n\hat{\theta}^2$$

on a

$$\Delta = (u^2 + 2n\hat{\theta})^2 + 4n\hat{\theta}^2(u^2 - n)$$

et le bornes

$$\frac{(u^2 + 2n\hat{\theta})}{2(n - u^2)} \pm \frac{\sqrt{\Delta}}{2(u^2 - n)}$$

ou

$$\frac{\hat{\theta} + \frac{u^2}{2n}}{1 - \frac{u^2}{n}} \pm \frac{\sqrt{\Delta}}{2(u^2 - n)}$$

Exercice 3

4 points

Pour étudier l'action d'un produit sur un paramètre biologique, on a mesuré, sur un échantillon supposé gaussien de 10 individus, la valeur du paramètre avant et après le traitement. Les résultats sont les suivants :

Individu	1	2	3	4	5	6	7	8	9	10
Valeur avant traitement	5,33	6,13	5,66	4,50	5,35	6,32	4,24	5,83	6,27	4,86
Valeur après traitement	5,32	6,00	5,64	4,59	5,19	6,17	4,11	5,86	6,13	4,68

A votre avis, le traitement modifie-t-il de façon significative le paramètre biologique ?

On note D la variable aléatoire des différences. L'échantillon étant supposé gaussien, D l'est aussi. On note μ son espérance et donc $D \sim \mathcal{N}(\mu, \sigma^2)$. Le test s'écrit :

$$\begin{aligned} H_0 &: \text{pas de modification } \mu = 0 \\ H_1 &: \text{il y a une modification } \mu \neq 0. \end{aligned}$$

le rapport de vraisemblance s'écrit

$$\begin{aligned} \frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} > k &\Leftrightarrow -\sum_{i=1}^n (D_i - \mu_1)^2 + \sum_{i=1}^n D_i^2 > 2\sigma^2 \log k \\ &\Leftrightarrow \mu_1 \sum_{i=1}^n D_i > k' \end{aligned}$$

on divise le test en deux tests unilatéraux dont la forme de la région critique dépend du signe de μ_1 sous H_1 .

la statistique $\frac{1}{n} \sum_{i=1}^n D_i$ suit une loi normale centré en μ et de variance $\frac{\sigma^2}{n}$. σ^2 étant inconnu on utilise la loi de Student comme fonction pivotale et

sous H_0 , la variable $U = \frac{\bar{D}}{\sigma/\sqrt{n}}$ suit une loi normale $\mathcal{N}(0, 1)$

sous H_0 , la variable $Z = \frac{\sum_{i=1}^n D_i^2}{\sigma^2}$ suit une loi du chi2 à n degrés de liberté et donc toujours sous H_0 , la variable

$$T = \sqrt{n} \frac{U}{\sqrt{Z}} = \frac{\sum_{i=1}^n D_i}{\sqrt{\sum_{i=1}^n D_i^2}}$$

suit une loi de Student à n degrés de liberté.

on a observé :

Individu	1	2	3	4	5	6	7	8	9	10
D	0,01	0,13	0,02	-0,09	0,16	0,15	0,13	-0,03	0,14	-0,02

et donc $\sum_{i=1}^n D_i = 0,064$ et $\sum_{i=1}^n D_i^2 = 0,64$ soit $T = \frac{0,64}{0,105} = 1,91$ or sur les tables, la valeur critique est 1,812, on rejette donc H_0 .

Exercice 4

4 points

On suppose que le nombre X d'absences par semaine dans une entreprise suit une loi de Poisson de paramètre λ avec

$$p(k) = P(X = k) = \frac{\lambda^k}{k!} \exp^{-\lambda}$$

On dispose des données suivantes, où n_x est le nombre de semaines pour lesquelles on a relevé x absences :

x	0	1	2	3
n_x	5	6	2	3

On suppose que ces données peuvent être considérées comme une réalisation d'un échantillon i.i.d. X_1, \dots, X_n de taille $n = 16$, de variable parente X . On notera T la statistique $\sum_{i=1}^n X_i$. On rappelle que, si $X \sim P(\lambda)$, $Y \sim P(\mu)$ et si X et Y sont indépendantes, alors $X + Y \sim P(\lambda + \mu)$. On cherche à tester les deux hypothèses simples :

$$\begin{aligned} H_0 &: \lambda = \lambda_0 \\ H_1 &: \lambda = \lambda_1 (< \lambda_0). \end{aligned}$$

1. Montrer que la région critique optimale s'exprime en fonction de la statistique T .

$$\begin{aligned} \frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} > k &\Leftrightarrow \log(\lambda_1) \sum_{i=1}^n X_i - n\lambda - \log(\prod X_i) - \log(\lambda_0) \sum_{i=1}^n X_i + n\lambda + \log(\prod X_i) > \log k \\ &\Leftrightarrow (\log(\lambda_1) - \log(\lambda_0)) \sum_{i=1}^n X_i > \log k \\ &\Leftrightarrow T = \sum_{i=1}^n X_i \leq k' \end{aligned}$$

2. Déterminer la distribution exacte de la statistique T sous l'hypothèse H_0 .

$$T = \sum_{i=1}^n X_i \sim P(n\lambda_0) \sim \mathcal{N}(n\lambda_0, n\lambda_0)$$

3. En déduire la région critique lorsque $\lambda_0 = 1$ sous H_0 on a

$$\begin{aligned} T &\sim \mathcal{N}(16, 16) \\ P(T \leq k') &= P\left(U \leq \frac{k' - 16}{\sqrt{16}}\right) = \alpha \end{aligned}$$

on a donc

$$\frac{k' - 16}{\sqrt{16}} = -1,645$$

soit

$$k' = 16 - 1,645\sqrt{16} = 9.42$$

et le résultat du test avec les données de l'exercice.

or on observe $T = 19$, on garde donc H_0 .

4. Déterminer la puissance du test pour $\lambda_1 = 0.5$.

$$\begin{aligned} 1 - \beta &= 1 - P(T > 9.42 \mid \lambda = 0.5) \\ &= 1 - P\left(U > \frac{9.42 - 8}{8}\right) = 1 - P(U > 0.177) = 0.57 \end{aligned}$$

Exercice 5

4 points

On modélise la durée de vie en heures T d'un appareil par une variable aléatoire suivant une loi exponentielle de densité

$$\forall t \geq 0, \quad f_\mu(t) = \frac{1}{\mu} \exp^{-t/\mu}$$

où $\mu > 0$ est un paramètre réel.

On met $n = 255$ appareils en service au même moment et on note T_i la durée de vie de l'appareil numéro i .

1. Calculer l'espérance mathématique de T

$$\mathbb{E}(T) = \int_0^\infty \frac{t}{\mu} \exp^{-t/\mu} dt$$

en intégrant par parties avec $u = t, u' = 1, v' = \frac{1}{\mu} \exp^{-t/\mu}$ et $v = -\exp^{-t/\mu}$ on a

$$\mathbb{E}(T) = [-t \exp^{-t/\mu}]_0^\infty + \int_0^\infty \exp^{-t/\mu} dt = \mu$$

2. Donner l'estimateur du maximum de vraisemblance de μ

$$L(t_1, \dots, t_n; \theta) = \prod_{i=1}^n \frac{1}{\mu} \exp^{-t_i/\mu} = \frac{1}{\mu^n} \exp^{-\sum_{i=1}^n t_i/\mu}$$

$$\log L(t_1, \dots, t_n; \theta) = -n \log \mu - \frac{\sum_{i=1}^n t_i}{\mu}$$

et donc

$$\frac{\partial \log L(t_1, \dots, t_n; \mu)}{\partial \mu} = -\frac{n}{\mu} + \frac{\sum_{i=1}^n t_i}{\mu^2} = -\frac{n}{\mu^2} \left(\frac{1}{n} \sum_{i=1}^n t_i - \mu \right)$$

soit

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n t_i$$

3. Donner sa variance

$$\text{Var}(\hat{\mu}) = \frac{\mu^2}{n}$$

et donc on peut approcher la loi de $\hat{\mu}$ par une loi normale

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\mu^2}{n}\right)$$

4. On suppose que l'espérance du temps de fonctionnement sans panne de ce type d'appareil est inférieur à 750 heures. Tester cette hypothèse quand on a observé

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = 812 \text{ heures}$$

on a

$$\begin{aligned} H_0 : \mu &\leq \mu_0 = 750 \\ H_1 : \mu &> \mu_0. \end{aligned}$$

on écrit le rapport de vraisemblance

$$\begin{aligned} \frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} > k &\Leftrightarrow -n \log \mu_1 - \frac{\sum_{i=1}^n t_i}{\mu_1} + n \log \mu_0 + \frac{\sum_{i=1}^n t_i}{\mu_0} > \log k \\ &\Leftrightarrow \sum_{i=1}^n t_i \left(\frac{1}{\mu_0} - \frac{1}{\mu_1} \right) > \log k - n \log \mu_0 + n \log \mu_1 \\ &\Leftrightarrow \bar{T} = \frac{1}{n} \sum_{i=1}^n t_i > k' \end{aligned}$$

on a donc

$$P(\bar{T} > k' \mid \mu_0 = 750) = P\left(\frac{\bar{T} - \mu_0}{\mu_0/\sqrt{n}} > \frac{k' - \mu_0}{\mu_0/\sqrt{n}}\right) = 0.05$$

on lit sur la table

$$\frac{k' - \mu_0}{\mu_0/\sqrt{n}} = 1,645$$

et donc

$$k' = \mu_0 + 1,645 \frac{\mu_0}{\sqrt{n}} = 750 + 1,645 \times \frac{750}{\sqrt{255}} = 827.26$$

comme on a observé 812 heures, on garde H_0 .

5. Calculer la puissance du test si l'espérance du temps de bon fonctionnement était de 900 heures.

$$\begin{aligned} 1 - \beta &= 1 - P(\bar{T} \leq 827.26 \mid \mu_0 = 900) = 1 - P\left(\frac{\bar{T} - 900}{900/\sqrt{255}} \leq \frac{827.26 - 900}{900/\sqrt{255}}\right) \\ &= 1 - P(U \leq -1.29) = 1 - 0.0985 = 0.9015 \end{aligned}$$

Formulaire

- fréquences $\hat{f}_i = \frac{n_i}{n}$ où n est le nombre total d'observations et n_i le nombre d'observant de la modalité i
- moyenne : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^n \hat{f}_i x_i$
- variance empirique : $\hat{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
- espérance d'une variable aléatoire discrète : $\mathbb{E}(X) = \sum_i x_i \mathbb{P}(x_i)$.
- espérance d'une variable aléatoire continue de densité $f(x)$: $\mathbb{E}(X) = \int x f(x) dx$.
- variance d'une variable aléatoire X : $Var(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$.
- quantile (ou fractiles) à l'ordre p , $\forall p \in [0, 1]$, $\hat{\Phi}_p$ telle que $\hat{\mathbb{P}}(X \leq \hat{\Phi}_p) = p$
- les quartiles :
 - $\hat{\Phi}_{\frac{1}{4}} = \hat{Q}_1$, telle que $\hat{F}(\hat{Q}_1) = \frac{1}{4}$,
 - $\hat{\Phi}_{\frac{1}{2}} = \hat{Q}_2 = \hat{M}$, telle que $\hat{F}(\hat{M}) = \frac{1}{2}$,
 - $\hat{\Phi}_{\frac{3}{4}} = \hat{Q}_3$, telle que $\hat{F}(\hat{Q}_2) = \frac{3}{4}$.
- covariance : $c_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ et corrélation : $\text{cor}(x, y) = \frac{c_{xy}}{\sqrt{\hat{\sigma}_x^2 \hat{\sigma}_y^2}}$
- probabilité conditionnelle : $\mathbb{P}(X = x_i | Y = y_j) = \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)}$
- espérance conditionnelle : $\mathbb{E}[Y | X = a] = \sum_{i=1}^n y_i \mathbb{P}(Y = y_i | X = a)$
- La loi des grands nombres $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{} \mathbb{E}(X)$

Estimateurs :

- Le risque : $R_{\hat{\theta}}(\theta) = \mathbb{E}((\hat{\theta}(X_1, \dots, X_n) - \theta)^2)$
 - Le biais d'un estimateur : $\mathbb{E}(\hat{\theta}) - \theta$
 - La variance d'un estimateur : $\mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2)$
- Soit X une variable aléatoire de distribution $f_{\theta}(x)$ et (X_1, \dots, X_n) une famille de n variables aléatoires i.i.d. ayant pour loi parente la loi de X .
- Vraisemblance :

$$L(\theta, X_1, \dots, X_n) = \prod_{i=1}^n f_{\theta}(X_i)$$

Log vraisemblance :

$$\ell(\theta, X_1, \dots, X_n) = \log L(\theta, X_1, \dots, X_n) = \sum_{i=1}^n \log f_{\theta}(X_i)$$

- L'information de Fisher : $I_n(\theta) = Var\left(\frac{\partial \ell(\theta, X_1, \dots, X_n)}{\partial \theta}\right) = -\mathbb{E}\left(\frac{\partial^2 \ell(\theta, X_1, \dots, X_n)}{\partial \theta^2}\right)$
- Pour $\theta \in \mathbb{R}^p$, L'information de Fisher est la matrice $p \times p$

$$I_n(\theta) = \mathbb{E}(\nabla \ell(\theta, X_1, \dots, X_n) \nabla \ell(\theta, X_1, \dots, X_n)^{\top})$$

- Borne de Cramer Rao
 - pour un estimateur sans biais de θ

$$BCR(\theta) = \frac{1}{I_n(\theta)} \leq \text{Var}(\hat{\theta}(X_1, \dots, X_n))$$

- Pour un estimateur sans biais d'une fonction de θ : $\mathbb{E}(\hat{u}(\theta)) = u(\theta)$

$$BCR(\theta) = \frac{u'(\theta)}{I_n(\theta)} \leq \text{Var}(\hat{u}(\theta(X_1, \dots, X_n)))$$

— pour un estimateur biaisé de biais B

$$BCR(\theta) = \frac{1 + B'(\theta)}{I_n(\theta)} \leq \text{Var}(\widehat{\theta}(X_1, \dots, X_n))$$

— L'estimateur max de vraisemblance $\widehat{\theta}_{MV}$ d'un paramètre θ est :

— Asymptotiquement sans biais

$$\widehat{\theta}_{MV} \xrightarrow[n \rightarrow \infty]{} \theta^*$$

— Asymptotiquement efficace

$$\text{Var}(\widehat{\theta}_{MV}) \xrightarrow[n \rightarrow \infty]{} I_n^{-1}(\theta^*)$$

— Asymptotiquement normal

$$\sqrt{n}(\widehat{\theta}_{MV} - \theta^*) \xrightarrow[n \rightarrow \infty]{} \mathcal{N}(0, I_1^{-1}(\theta^*))$$

note : $I_n(\theta^*) = nI_1(\theta^*)$

Variables aléatoires et lois

— Soit $Y \sim \mathcal{N}(0, 1)$ une variable aléatoire normale centrée réduite.

— Soit Y_1, Y_2, \dots, Y_n un échantillon de n réalisations i.i.d. de cette variable aléatoire.

— La loi du χ^2 : On appelle loi du χ^2 à n degrés de libertés la loi de la variable aléatoire $Z_n = \sum_{i=1}^n Y_i^2$

— La loi de student : On appelle loi de student à n degrés de libertés la loi de la variable aléatoire T_n

$$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}} \quad \begin{array}{l} N \sim \mathcal{N}(0, 1) \\ X_n \sim \chi_n^2 \end{array}$$

Tests

— Régions de décision :

Région d'acceptation : l'ensemble \overline{W} des réalisations (x_1, \dots, x_n) pour lesquelles on garde H_0

Région critique : l'ensemble W des réalisations (x_1, \dots, x_n) pour lesquelles on rejette H_0

— Mesures de performances des tests

— Risque de première espèce ou niveau de signification du test α : probabilité de rejeter H_0 alors que H_0 est vraie

$$\alpha = \mathbb{P}(D = 1 | H_0) = \int_{(x_1, \dots, x_n) \in W} L(x_1, \dots, x_n, H_0) dx_1, \dots, dx_n$$

— Risque de seconde espèce β : probabilité de garder H_0 alors que H_1 est vraie

$$\beta = \mathbb{P}(D = 0 | H_1) = \int_{(x_1, \dots, x_n) \in \overline{W}} L(x_1, \dots, x_n, H_1) dx_1, \dots, dx_n$$

— puissance d'un test : $1 - \beta$ la probabilité de rejeter hypothèses nulle avec raison

— p -valeur :

— Tests du rapport de vraisemblance

$$W = \left\{ x_1, \dots, x_n \mid \frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} > k \right\}$$

— Le principe de Neyman–Pearson : pour un α fixé, trouver la fonction de décision qui minimise β (ou qui maximise $1 - \beta$ la puissance du test)

— Le théorème de Neyman–Pearson : pour un test paramétrique de deux hypothèses simples, le test du rapport de vraisemblance

$$W = \left\{ x_1, \dots, x_n \mid \frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} > k \right\}$$

tel que k soit fixé de sorte que

$$\alpha = \int_{(x_1, \dots, x_n) \in W} L(x_1, \dots, x_n, H_0) dx_1, \dots, dx_n$$

est optimal au sens du principe de Neyman–Pearson (parmi tous les tests de risque α , c'est celui de puissance maximale).