Faculté des Sciences
Aix*Marseille Université

Année universitaire 2022-2023

Site : ☐ Luminy ☒ St-Charles ☐ St-Jérôme ☐ Cht-Gombert ☐ Aix-Montperrin ☐ Aubagne-SATIS
Sujet Session de ☒ 1er semestre - ☐ 2ème semestre - ☒ Session unique Durée de l'épreuve : 3h
Examen de ☐ L1/☐ L2/☐ L3 - ☒ M1/☐ M2 - ☐ LP - ☐ DU Nom Diplôme : MAS
Code Apogée du module : SMSAU02C Libellé du module : **Statistique.**
Document autorisé : ☒ OUI-☐ NON Calculette autorisée ☐ OUI-☒ NON

Correction de l'examen du Lundi 12 Décembre 2022 (3 heures)

Les documents, notes de cours, calculatrices et téléphones portables sont interdits, à l'exception d'une feuille A4 recto-verso. Le soin apporté à la rédaction sera un élément d'appréciation.

Exercice 1. Risque quadratique.
Soit $X = (X_1, \cdots; X_n)$ un échantillon de taille n telle que $E(X_1) = \mu$ et $E(X_1^2) = \sigma^2 < \infty$. On pose $V_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $T_n = \frac{2}{n(n+1)}\sum_{i=1}^n iX_i$.

1. Montrer que V_n et T_n sont deux estimateurs sans biais de μ.
   V_n et T_n ne dépendent que de l'observation X et sont à valeurs réelles. Ce sont donc deux estimateurs de μ. Par linéarité de l'espérance,

$$E(V_n) = E(X_1) = \mu\,;\ E(T_n) = \frac{2}{n(n+1)}\sum_{i=1}^n iE(X_i) = \frac{2\mu}{n(n+1)}\sum_{i=1}^n i = \mu\,.$$

   V_n et T_n sont deux estimateurs sans biais de μ.

2. V_n est-il un meilleur estimateur que T_n au sens du risque quadratique ? Justifiez votre réponse.
   V_n et T_n étant sans biais, leur risque quadratique est égal à leur variance. De plus, les variables X_i étant i.i.d.,

$$\mathrm{var}(V_n) = \frac{\mathrm{var}(X_1)}{n}\,;$$

$$\mathrm{var}(T_n) = \frac{4}{n^2(n+1)^2}\sum_{i=1}^n i^2\,\mathrm{var}(X_i) = \frac{4\,\mathrm{var}(X_1)}{n^2(n+1)^2}\sum_{i=1}^n i^2 \quad (1)$$

$$= \frac{4\,\mathrm{var}(X_1)}{n^2(n+1)^2}\frac{n(n+1)(2n+1)}{6} = \frac{2\,\mathrm{var}(X_1)}{3}\frac{2n+1}{n(n+1)}\,. \quad (2)$$

   Comme pour tout $n \geq 1$, $\frac{1}{n} \leq \frac{2(2n+1)}{3n(n+1)}$, V_n est meilleur que T_n.

.

Exercice 2. Statistique bayésienne.
Soit $(X_1, \cdots, X_n)$ un n-échantillon de loi géométrique $\mathcal{G}(p)$ de paramètre $p \in [0,1]$, et on considère la loi a priori sur p : $P \sim \mathrm{Beta}(\alpha, \beta)$.
On rappelle que si $X \sim \mathcal{G}(p)$, pour tout entier $k \geq 1$, $\mathbb{P}(X = k) = p(1-p)^{k-1}$, et $\mathbb{E}(X) = 1/p$. Si $P \sim \mathrm{Beta}(\alpha, \beta)$, la densité de P est donnée par $f_P(t) = \frac{1}{B(\alpha,\beta)}t^{\alpha-1}(1-t)^{\beta-1}\,\mathbb{1}_{]0,1[}(t)$, et $\mathbb{E}(P) = \frac{\alpha}{\alpha+\beta}$.

1. Montrez que l'estimateur du maximum de vraisemblance de p est $\hat{p}_n = n/S_n$, où $S_n = \sum_{i=1}^n X_i$.
La vraisemblance du modèle est donnée par

$$L(p;X) = \prod_{i=1}^n L(p;X_i) = \prod_{i=1}^n p(1-p)^{X_i-1} = p^n(1-p)^{S_n-n}$$

Ainsi, $\frac{d}{dp}\log L(p;X) = \frac{n}{p} - \frac{S_n-n}{1-p}$. L'EMV $\hat{p}_n$ est donc solution de

$$\frac{n}{\hat{p}_n} = \frac{S_n - n}{1-\hat{p}_n} \iff \hat{p}_n = \frac{n}{S_n}\,.$$

2. Calculer la loi a posteriori de P, et montrez que l'espérance a posteriori est $\tilde{p}_n = \frac{\alpha+n}{\beta+\alpha+S_n}$.
La loi a posteriori de P est la loi conditionnelle de P sachant les observations X. Sa densité est donc proportionnelle à

$$p^{\alpha-1}(1-p)^{\beta-1}\ \mathbb{1}_{[0;1]}(p)L(p;X) = p^{n+\alpha-1}(1-p)^{\beta+S_n-n-1}\ \mathbb{1}_{[0;1]}(p)\,.$$

Il s'agit donc de la loi Beta$(n+\alpha, \beta+S_n-n)$. On en déduit que

$$\tilde{p}_n = E(P|X) = \frac{n+\alpha}{\beta+\alpha+S_n}\,.$$

3. $\tilde{p}_n$ est-il un estimateur consistant de p ?

$$\tilde{p}_n = \frac{1+\frac{\alpha}{n}}{\frac{\beta+\alpha}{n}+\frac{S_n}{n}}\,.$$

Par la loi forte des grands nombres, $\frac{S_n}{n}$ converge p.s. vers $E(X_1) = 1/p$ quand n tend vers $+\infty$. Par conséquent, $\tilde{p}_n$ converge p.s. vers p quand n tend vers $+\infty$, et est donc consistant.

4. Donner la définition du risque quadratique bayésien $R(\tilde{p}_n)$ de $\tilde{p}_n$. Montrez sans calcul que $R(\tilde{p}_n) \leq R(\hat{p}_n)$.
$R(\tilde{p}_n) = \mathbb{E}\left[(\tilde{p}_n - P)^2\right]$, où $\mathbb{E}$ désigne l'espérance sous la loi jointe de (P,X). On a donc

$$R(\tilde{p}_n) = \mathbb{E}\left[(P - \mathbb{E}(P|X))^2\right] \leq \mathbb{E}\left[(P-T(X))^2\right]\,,$$

pour toute fonction des observations $T(X)$ de carré intégrable, par définition de l'espérance conditionnelle. En particulier, $R(\tilde{p}_n) \leq \mathbb{E}\left[(P-\hat{p}_n)^2\right] = R(\hat{p}_n)$.

5. Pour $n = 10$, on a observé $S_{n,obs} = 28$. On a choisi par ailleurs $\alpha = \beta = 2$. Le tableau ci-dessous donne les quantiles de la loi Beta$(12, 20)$. Construire un intervalle de crédibilité pour P de couverture 95%.

α	0.025	0.05	0.95	0.975
Quantile	0.218	0.241	0.518	0.546

La loi a posteriori est la loi Beta$(n+\alpha, \beta+S_{n,obs}-n) =$ Beta$(12,20)$. D'après le tableau des quantiles, on a donc

$$\mathbb{P}\left[0.218 \leq P \leq 0.546 | X = X_{obs}\right] = 0.95\,.$$

Un intervalle de crédibilité pour P de couverture 95% est donc l'intervalle $[0.218; 0.546]$.

Exercice 3. Test du rapport de vraisemblance.
On observe $x = (x_1, \cdots, x_n)$ et $y = (y_1, \cdots, y_m)$ les durées de vie de deux types de composants. On suppose que x (respectivement y) est une réalisation d'un n-échantillon de la loi exponentielle de paramètre $1/\lambda$ notée $\mathcal{E}(1/\lambda)$, (respectivement d'un m-échantillon de la loi exponentielle de paramètre $1/\mu$). On veut tester $(H_0) : \lambda = \mu$ contre $(H_1) : \lambda \neq \mu$.

On rappelle que si $X \sim \mathcal{E}(1/\lambda)$, $\mathbb{E}(X) = \lambda$, $\text{var}(X) = \lambda^2$, la densité de X est donnée par $f_X(t) = \frac{1}{\lambda}\exp(-\frac{t}{\lambda})\ \mathbb{1}_{R^+}(t)$, et sa fonction de répartition $F_X(t) = (1 - \exp(-\frac{t}{\lambda}))\ \mathbb{1}_{R^+}(t)$.

1. Quelle est l'expression de la vraisemblance $V_1(\lambda, \mu)$ de $(X_1, \cdots, X_n, Y_1, \cdots Y_m)$ sous l'hypothèse (H_1). En déduire que l'estimateur du maximum de vraisemblance de (λ, μ) sous (H_1) est
$$\hat{\lambda} = \frac{S_n}{n}, \hat{\mu} = \frac{T_m}{m}, \text{ où } S_n = \sum_{i=1}^{n} X_i \text{ et } T_m = \sum_{j=1}^{m} Y_j\,.$$
$$\begin{aligned} V_1(\lambda,\mu) &= \prod_{i=1}^{n} \frac{1}{\lambda}\exp\left(-\frac{X_i}{\lambda}\right)\ \mathbb{1}_{X_i\geq 0} \prod_{j=1}^{m} \frac{1}{\mu}\exp\left(-\frac{Y_j}{\mu}\right)\ \mathbb{1}_{Y_j\geq 0} \\ &= \frac{1}{\lambda^n}\exp\left(-\frac{S_n}{\lambda}\right)\frac{1}{\mu^m}\exp\left(-\frac{T_m}{\mu}\right)\ \mathbb{1}_{\min(X_i)\geq 0}\ \mathbb{1}_{\min(Y_j)\geq 0}\,. \end{aligned}$$
On a donc
$$\frac{\partial}{\partial\lambda}\log V_1(\lambda,\mu) = -\frac{n}{\lambda} + \frac{S_n}{\lambda^2}\,,\ \frac{\partial}{\partial\mu}\log V_1(\lambda,\mu) = -\frac{m}{\mu} + \frac{T_m}{\mu^2}\,,$$
ce qui donne le résultat.

2. Quelle est l'expression de la vraisemblance $V_0(\lambda)$ de $(X_1, \cdots, X_n, Y_1, \cdots Y_m)$ sous l'hypothèse (H_0). En déduire que l'estimateur du maximum de vraisemblance de λ $(= \mu)$ sous (H_0) est $\tilde{\lambda} = \frac{S_n+T_m}{n+m}$.
$$V_0(\lambda) = V_1(\lambda,\lambda) = \frac{1}{\lambda^{n+m}}\exp\left(-\frac{S_n+T_m}{\lambda}\right)\ \mathbb{1}_{\min(X_i)\geq 0}\ \mathbb{1}_{\min(Y_j)\geq 0}\,.$$
On a donc
$$\frac{\partial}{\partial\lambda}\log V_0(\lambda) = -\frac{n+m}{\lambda} + \frac{S_n+T_m}{\lambda^2}\,,$$
ce qui donne le résultat.

3. On note $R_{n,m}$ la statistique du test du rapport de vraisemblance. Montrer que
$$\ln(R_{n,m}) = r_{n,m}\left(\frac{T_m}{S_n}\right) - r_{n,m}\left(\frac{m}{n}\right),\ \text{où } r_{n,m}(x) = n\ln(1+x) + m\ln(1+1/x)..$$
$$\begin{aligned} R_{n,m} &= \frac{\sup_{\lambda>0,\mu>0} V_1(\lambda,\mu)}{\sup_{\lambda>0} V_0(\lambda)} = \frac{V_1(\hat{\lambda},\hat{\mu})}{V_0(\tilde{\lambda})} \\ &= \frac{n^n}{S_n^n}e^{-n}\frac{m^m}{T_m^m}e^{-m}\frac{(S_n+T_m)^{n+m}}{(n+m)^{n+m}}e^{n+m} = \frac{(S_n+T_m)^{n+m}}{S_n^n T_m^m}\frac{n^n m^m}{(n+m)^{n=m}} \\ &= \left(1+\frac{T_m}{S_n}\right)^n\left(1+\frac{S_n}{T_m}\right)^m\frac{1}{\left(1+\frac{m}{n}\right)^m}\frac{1}{\left(1+\frac{n}{m}\right)^n} \\ &= \exp\left(r_{n,m}(T_m/S_n)\right)\exp(-r_{n,m}(m/n)) \end{aligned}$$

4. Montrer que $r_{n,m}$ est décroissante sur $]0, m/n]$, et croissante sur $[m/n, +\infty[$. En déduire que la région de rejet du test du rapport de vraisemblance de (H_0) contre (H_1) a la forme

$$\mathcal{R} = \left\{ \frac{T_m}{S_n} \leq u_1 \right\} \cup \left\{ \frac{T_m}{S_n} \geq u_2 \right\}, 0 < u_1 \leq \frac{m}{n} \leq u_2 .$$

$$r'_{n,m}(x) = \frac{n}{1+x} - \frac{m}{x^2(1+1/x)} = \frac{nx - m}{x(1+x)} .$$

$r_{n,m}$ est donc croissante sur $[m/n, +\infty[$, et décroissante sur $]0, m/n]$. La région de rejet du test du rapport de vraisemblance est de la forme

$$\mathcal{R} = \{\log(R_{n,m}) \geq t\} = \{r_{n,m}(T_m/S_n) \geq t + r_{n,m}(m/n)\} ,$$

pour un choix de $t > 0$ dépendant du niveau choisi. Comme $\lim_{x\to+\infty} r_{n,m}(x) = \lim_{x\to+0} r_{n,m}(x) = +\infty$, on a donc $\mathcal{R} = \left\{ \frac{T_m}{S_n} \leq u_1 \right\} \cup \left\{ \frac{T_m}{S_n} \geq u_2 \right\}$, où $0 < u_1 \leq m/n \leq u_2$ et $r_{n,m}(u_1) = r_{n,m}(u_2) = t + r_{n,m}(m/n)$.

5. Quelle est la loi de X_1/λ ? Montrer que sous (H_0), la loi de $\frac{T_m}{S_n}$ est indépendante de $\lambda (= \mu)$.
Comme $X_i \sim \mathcal{E}(1/\lambda)$, $X'_i := X_i/\lambda \sim \mathcal{E}(1)$. Sous (H_0), $Y'_j := Y_j/\lambda$ suit aussi la loi $\mathcal{E}(1)$. Par ailleurs $\frac{T_m}{S_n} = \frac{T'_m}{S'_n}$, où $T'_m = \sum_{j=1}^m Y'_j$ et $S'_n = \sum_{i=1}^n X'_i$. Les variables Y'_j, X'_i étant indépendantes et de loi $\mathcal{E}(1)$ sous (H_0), la loi de $\frac{T_m}{S_n}$ sous (H_0) ne dépend pas de λ.

6. Quelle quantité est estimée par la valeur `q` en sortie du script `R` suivant ?

```
n <- 10
m <- 15
Nsim <- 10000
X <- rexp(n*Nsim, rate = 1)
X <- matrix(X,nrow = Nsim)
S <- apply(X,1,sum)
Y <- rexp(m*Nsim, rate = 1)
Y <- matrix(Y,nrow = Nsim)
T <- apply(Y, 1, sum)
R = T/S
alpha <- 0.05
q <- quantile(R,alpha)
```

Dans ce script,
— `S` est un vecteur de taille `Nsim` contenant des réalisations indépendantes de S_n/λ ;
— `T` est un vecteur de taille `Nsim` contenant des réalisations indépendantes de T_m/μ ;
— `R` est un vecteur de taille `Nsim` contenant des réalisations indépendantes de T_m/S_n sous (H_0) ;
— `q` est le quantile empirique d'ordre α de l'échantillon constitué par les `Nsim` coordonnées de `R`.
`q` est donc une estimation du quantile d'ordre α de la loi de la statistique T_m/S_n sous (H_0).

7. On note $Q_{n,m}$ la fonction quantile de la loi de la variable T_m/S_n sous l'hypothèse (H_0). Pour $n = 10$ et $m = 15$, on donne

α	0.025	0.05	0.95	0.975
$Q_{n,m}(\alpha)$	0.686	0.773	3.064	3.573

Construire un test de niveau 5% de (H_0) contre (H_1). On a observé $T_{m,obs}/S_{n,obs} = 1.178$. Que concluez-vous au niveau 5% ?

On a déja vu que la région de rejet du test du rapport de vraisemblance était du type $\mathcal{R} = \left\{\frac{T_m}{S_n} \leq u_1\right\} \cup \left\{\frac{T_m}{S_n} \geq u_2\right\}$. Pour obtenir un test de niveau 5%, on peut choisir u_1 et u_2 de telle sorte que

$$\mathbb{P}_{(H_0)}\left[\frac{T_m}{S_n} \leq u_1\right] = \mathbb{P}_{(H_0)}\left[\frac{T_m}{S_n} \geq u_2\right] = 0.025\,.$$

On a donc $u_1 = Q_{n,m}(0.025) = 0.686$, et $u_2 = Q_{n,m}(0.975) = 3.573$. On observe $T_{m,obs}/S_{n,obs} = 1.178 \in]0.686; 3.573[$. On ne peut donc pas rejeter (H_0). On accepte l'hypothèse selon laquelle $\lambda = \mu$, sans connaître la probabilité de se tromper.

8. Montrez que $\frac{\lambda T_m}{\mu S_n}$ est un pivot pour l'estimation du paramètre $\frac{\lambda}{\mu}$. Pour $n = 10$ et $m = 15$, construire un intervalle de confiance pour $\frac{\lambda}{\mu}$ de coefficient de sécurité 95% ?

$$\frac{\lambda T_m}{\mu S_n} = \frac{\sum_{j=1}^m Y_j/\mu}{\sum_{i=1}^n X_i/\lambda}$$

Les variables Y_j/μ et X_i/λ sont indépendantes et de loi $\mathcal{E}(1)$. $\frac{\lambda T_m}{\mu S_n}$ a donc la même loi que la loi de $\frac{T_m}{S_n}$ sous (H_0), et sa fonction quantile est donc $Q_{n,m}$. En particulier,

$$\mathbb{P}_{\lambda,\mu}\left[0.686 \leq \frac{\lambda T_m}{\mu S_n} \leq 3.573\right] = 0.95\,.$$

L'intervalle $[0.686\frac{S_n}{T_m}; 3.573\frac{S_n}{T_m}]$ est donc un intervalle de confiance pour $\frac{\lambda}{\mu}$ de coefficient de sécurité 95%.