Statistique Mathématique **Université de Rennes – M1**

CC1 2023/2024 – Durée 1h30

Les documents et appareils électroniques (calculatrice, téléphone, ordinateur, ...) sont interdits. ***Toutes les réponses doivent être justifiées.***

Exercice 1 - Coins à champignons

Deux mathématiciens amateurs de champignons veulent comparer leurs coins favoris sur la base des récoltes effectuées sur n années, $(X_i, Y_i)_{i=1,\dots,n}$. La récolte du premier chercheur l'année i est notée X_i, celle du second Y_i. On peut modéliser X_i (respectivement Y_i) par une loi de Poisson de paramètre λ_1 (respectivement λ_2). On supposera par ailleurs X_i indépendant de Y_i, et les récoltes de chaque année indépendantes.

1. Proposer un modèle pour cette expérience.
2. Proposer un estimateur $\hat{\lambda}_j$ pour λ_j ($j \in \{1, 2\}$), et déterminer son risque quadratique.
3. Soit $(U_i, V_i)_{i\in\mathbb{N}}$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^2$, avec $U_i \perp\!\!\!\perp V_i$, et (U, V) un couple aléatoire tel que $U_i \rightsquigarrow U$, $V_j \rightsquigarrow V$, et $U \perp\!\!\!\perp V$. Montrer que $(U_i, V_i) \rightsquigarrow (U, V)$.
4. Déterminer le comportement asymptotique de $\sqrt{\hat{\lambda}_1} - \sqrt{\hat{\lambda}_2}$. En déduire un intervalle de niveau de confiance **asymptotique** $1 - \alpha$ sur $\sqrt{\lambda_1} - \sqrt{\lambda_2}$ du type $[\hat{T}_-, +\infty[$ (où $\alpha \in]0, 1[$ et $\hat{T}_-$ est une statistique).
5. En déduire un test d'erreur de première espèce **asymptotique** α pour les hypothèses
$$\begin{cases} H_0 & : \quad \lambda_1 = \lambda_2, \\ H_1 & : \quad \lambda_1 > \lambda_2. \end{cases}$$
6. On note $S_1 = \sum_{i=1}^n X_i$, $S_2 = \sum_{i=1}^n Y_i$. Donner la loi de $S_1 + S_2$. Montrer que, pour $k \in \mathbb{N} \setminus \{0\}$,
$$S_1 \mid \{S_1 + S_2 = k\} \sim \mathcal{B}(k, \rho),$$
où $\rho = \lambda_1/(\lambda_1 + \lambda_2)$.
7. Soient $Z_1 \sim \mathcal{B}(p_1)$, $Z_2 \sim \mathcal{B}(p_2)$, avec $p_1 \leq p_2$. Montrer que $Z_1 \preccurlyeq Z_2$. En déduire que, si $\lambda_1 \leq \lambda_2$, $S_1 \mid \{S_1 + S_2 = k\} \preccurlyeq \mathcal{B}(k, 1/2)$, pour tout $k \geq 1$.
8. Pour $\beta \in]0, 1[$ et $k \geq 1$, on note $q_\beta(k)$ le quantile d'ordre β d'une loi $\mathcal{B}(k, 1/2)$. Montrer que
$$T = \begin{cases} \mathbb{1}_{S_1 > q_{1-\alpha}(S_1+S_2)} & \text{si} \quad S_1 + S_2 \geq 1, \\ U & \text{si} \quad S_1 = S_2 = 0, \end{cases}$$

où $U \sim \mathcal{B}(\alpha)$ est indépendante de (S_1, S_2), est un test d'erreur de première espèce α pour les hypothèses

$$\begin{cases} H_0 & : \quad \lambda_1 \leq \lambda_2, \\ H_1 & : \quad \lambda_1 > \lambda_2. \end{cases}$$

9. Application : donner une p-valeur pour les observations $S_1 = 30$, $S_2 = 20$. En notant F la fonction d'une répartition d'une $\mathcal{B}(50, 0.5)$, on donne les valeurs suivantes de F :

x	28	29	30	31	32	33
$F(x)$	0.839	0.899	0.941	0.968	0.984	0.992

Solution 1 -

1. Modèle : $\left(\mathbb{N}^{2n}, \mathcal{P}(\mathbb{N}^{2n}), (\mathcal{P}(\lambda_1) \otimes \mathcal{P}(\lambda_2))^{\otimes n}_{\lambda_1, \lambda_2 > 0}\right)$.
2. On prend $\hat{\lambda}_1 = \bar{X}_n$, $\hat{\lambda}_2 = \bar{Y}_n$. Ce sont deux estimateurs non biaisés, et

$$E_{\lambda_j}(\hat{\lambda}_j - \lambda_j)^2 = \mathrm{Var}_{\lambda_j}(\hat{\lambda}_j) = \frac{\lambda_j}{n}.$$

3. Soit $t = (t_1, t_2) \in \mathbb{R}^2$.

$$\begin{aligned} \mathbb{E}(e^{i\langle t, (U_j, V_j)\rangle}) &= \mathbb{E}(e^{it_1 U_j + it_2 V_j}) \\ &= \mathbb{E}(e^{it_1 U_j})\mathbb{E}(e^{it_2 V_j}) \qquad \text{(indépendance)} \\ &\underset{j \to +\infty}{\longrightarrow} \mathbb{E}(e^{it_1 U})\mathbb{E}(e^{it_2 V}) \qquad \text{(cv en loi)} \\ &= \mathbb{E}(e^{i\langle t, (U,V)\rangle}) \qquad \text{(indépendance encore)}. \end{aligned}$$

Donc $(U_j, V_j) \rightsquigarrow (U, V)$.

4. Comme $\mathbb{E}(X_1)^2 < +\infty$, le Théorème Central Limite donne

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \rightsquigarrow \mathcal{N}(0, \lambda_j).$$

Une application de la méthode Δ pour la fonction $\sqrt{}$ (différentiable sur $]0, +\infty[$ donne alors

$$\begin{aligned} \sqrt{n}(\sqrt{\hat{\lambda}_j} - \sqrt{\lambda_j}) &\rightsquigarrow \frac{1}{2\sqrt{\lambda_j}} \mathcal{N}(0, \lambda_j) \\ &\sim \mathcal{N}(0, (1/4)). \end{aligned}$$

En applicant le résulat de la question précédente, on déduit que

$$\sqrt{n}(\sqrt{\hat{\lambda}_1} - \sqrt{\hat{\lambda}_2} - (\sqrt{\lambda_1} - \sqrt{\lambda_2})) \rightsquigarrow N_1 + N_2,$$

où $N_j \sim \mathcal{N}(0,(1/4))$ et $N_1 \perp\!\!\!\perp N_2$. On en déduit

$$\sqrt{n}(\sqrt{\hat{\lambda}_1} - \sqrt{\hat{\lambda}_2} - (\sqrt{\lambda_1} - \sqrt{\lambda_2}) \rightsquigarrow \mathcal{N}(0,(1/2)).$$

On en déduit que

$$\mathbb{P}\left(\sqrt{\hat{\lambda}_1} - \sqrt{\hat{\lambda_2}} \leq \sqrt{\lambda_1} - \sqrt{\lambda_2} + \frac{q_{1-\alpha}(\mathcal{N}(0,1))}{\sqrt{2n}}\right) \underset{n\to+\infty}{\longrightarrow} 1-\alpha,$$

et donc que $I(\alpha) = \left[\sqrt{\hat{\lambda}_1} - \sqrt{\hat{\lambda_2}} - \frac{q_{1-\alpha}(\mathcal{N}(0,1))}{\sqrt{2n}}, +\infty\right[$ est une intervalle de niveau de confiance asymptotique $1-\alpha$ pour $\sqrt{\lambda_1} - \sqrt{\lambda_2}$.

5. En notant $T(\alpha) = \mathbb{1}_{0\notin I(\alpha)}$, on a bien un test d'erreur de première espèce asymptotique α. On peut remarquer que $0 \notin \alpha$ équivaut à $\sqrt{\hat{\lambda}_1} - \sqrt{\hat{\lambda_2}}$ grand, ce qui est cohérent avec la forme de H_1.
6. On a $S_1 \sim \mathcal{P}(n\lambda_1)$, $S_2 \sim \mathcal{P}(n\lambda_2)$, avec $S_1 \perp\!\!\!\perp S_2$, donc $S_1 + S_2 \sim \mathcal{P}(n(\lambda_1+\lambda_2))$. Par ailleurs, pour $j,k \in \mathbb{N}$,

$$\mathbb{P}\left(S_1 = j \cap (S_1+S_2) = k\right)\begin{cases} 0 & \text{si} \quad j > k \\ e^{-n(\lambda_1\lambda_2)}\left(\frac{(n\lambda_1)^j}{j!}\frac{(n\lambda_2)^{k-j}}{(k-j)!}\right) & \text{si} \quad j \leq k. \end{cases}$$

On en déduit que $\mathbb{P}(S_1 = j \mid S_1 + S_2 = k) = 0$ si $j > k$, et

$$\begin{aligned}\mathbb{P}(S_1 = j \mid S_1+S_2 = k) &= e^{-n(\lambda_1\lambda_2)}\left(\frac{(n\lambda_1)^j}{j!}\frac{(n\lambda_2)^{k-j}}{(k-j)!}\right) / \left(e^{-n(\lambda_1+\lambda_2)}\frac{(n(\lambda_1+\lambda_2))^k}{k!}\right) \\ &= \binom{k}{j}\rho^j(1-\rho)^{k-j},\end{aligned}$$

pour $j \leq k$, avec $\rho = \frac{\lambda_1}{\lambda_1+\lambda_2}$. On en déduit $S_1 \mid \{S_1+S_2 = k\} \sim \mathcal{B}(k,\rho)$.
7. Soit $U \sim \mathcal{U}(]0,1[)$, et $V_j = \mathbb{1}_{U\leq p_j}$. On a $V_j \sim Z_j$, et $V_1 \leq V_2$. On en déduit que $Z_1 \preccurlyeq Z_2$. On en déduit aussi que si $Z_j \sim \mathcal{B}(\ell, p_j)$, pour $\ell \geq 1$, $Z_1 \preccurlyeq Z_2$ (regarder $\sum_{r=1}^{\ell} \mathbb{1}_{U_r\leq p_j}$, où $U_1,\ldots,U_r$ i.i.d. $\mathcal{U}(0,1[)$). Si $\lambda_1 \leq \lambda_2$, $\rho \leq 1/2$, et donc $S_1 \mid \{S_1 + S_2 = k\} \sim \mathcal{B}(k,\rho) \preccurlyeq \mathcal{B}(k,(1/2))$.
8. Supposons $\lambda_1 \leq \lambda_2$ (et donc $\rho \leq 1/2$). On a

$$\begin{aligned}P_{\lambda_1,\lambda_2}(T=1) &= \alpha P_{\lambda_1,\lambda_2}(S_1+S_2=0) + \sum_{k=1}^{+\infty} P_{\lambda_1,\lambda_2}(S_1+S_2=k)P_{\lambda_1,\lambda_2}(T=1 \mid (S_1+S_2=k)) \\ &= \alpha P_{\lambda_1,\lambda_2}(S_1+S_2=0) + \sum_{k=1}^{+\infty} P_{\lambda_1,\lambda_2}(S_1+S_2=k)P_{\lambda_1,\lambda_2}(S_1 > q_{1-\alpha}(k) \mid (S_1+S_2=k)).\end{aligned}$$

Or, pour tout $k \geq 1$,

$$\begin{aligned}P_{\lambda_1,\lambda_2}(S_1 > q_{1-\alpha}(k) \mid (S_1+S_2=k)) &= \mathbb{P}(\mathcal{B}(k,\rho) > q_{1-\alpha}(k)) \\ &\leq \mathbb{P}(\mathcal{B}(k,\rho) > q_{1-\alpha}(k)) \qquad (\mathcal{B}(k,\rho) \preccurlyeq \mathcal{B}(k,1/2)) \\ &\leq 1-(1-\alpha).\end{aligned}$$

On en déduit

$$P_{\lambda_1,\lambda_2}(T=1) \leq \alpha \sum_{k=0}^{+\infty} P_{\lambda_1,\lambda_2}(S_1+S_2=k) = \alpha.$$

9. La p-valeur basée sur ces observations est définie par

$$p_{val} = \inf\{\alpha \mid 30 > q_{1-\alpha}(50)\},$$

ce qui correspond au plus petit niveau de test menant au rejet de H_0 en se basant sur ces observations. Or

$$\begin{aligned} 30 > q_{1-\alpha}(50) &\Leftrightarrow 29 \geq q_{1-\alpha}(50) \\ &\Leftrightarrow F(29) \geq 1-\alpha \\ &\Leftrightarrow 0.899 \geq 1-\alpha. \end{aligned}$$

On en déduit $p_{val} = 0.101$, on a donc heuristiquement 10% de chances de rejeter à tort H_0 en se basant sur ces observations (ça fait beaucoup alors on s'abstient en général).

Exercice 2 - Un modèle Gaussien

On note ε_i des variables i.i.d. $\mathcal{N}(0,\sigma^2)$, pour un σ^2 inconnu, et on suppose que l'on observe $Y_1 = 3\theta_1 - 2\theta_2 + \varepsilon_1$, $Y_2 = 5\theta_1 + 2\theta_2 + \varepsilon_2$, $Y_3 = \theta_1 - 2\theta_2 + \varepsilon_3$, $Y_4 = \theta_1 - 2\theta_2 + \varepsilon_4$, où $\theta = (\theta_1,\theta_2)^T$ est lui aussi inconnu.

1. Donner (sous forme matricielle) le modèle de cette expérience.
2. Déterminer l'estimateur des moindres carrés pour θ, noté $\hat{\theta}_{LS}$. Développer les calculs jusqu'au bout (ne pas se contenter de la forme matricielle) et donner sa loi.
3. Donner un estimateur $\hat{\sigma}^2$ de σ^2 et déterminer sa loi (on pourra se contenter ici d'une formule matricielle).
4. Donner une ellipsoïde de niveau de confiance 95% pour θ. Quels en sont les grands axes ?
5. Donner un intervalle de niveau de confiance 98% pour $u = 3\theta_1 - 2\theta_2$.

Solution 2 -

1. En notant $Y = (Y_1,Y_2,Y_3,Y_4)^T$, $\theta = (\theta_1,\theta_2)^T$, $\varepsilon = (\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4)^T$, et

$$X = \begin{pmatrix} 3 & -2 \\ 5 & 2 \\ 1 & -2 \\ 1 & -2 \end{pmatrix},$$

le modèle se met sous la forme

$$Y = X\theta + \varepsilon,$$

ou encore $Y \sim \mathcal{N}(X\theta, \sigma^2 I_4)$.

2. En remarquant que

$$X^T X = \begin{pmatrix} 36 & 0 \\ 0 & 16 \end{pmatrix},$$

on a

$$\begin{aligned} \hat{\theta}_{LS} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} \begin{pmatrix} 3Y_1 + 5Y_2 + Y_3 + Y_4 \\ -2Y_1 + 2Y_2 - 2Y_3 - 2Y_4 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{36} (3Y_1 + 5Y_2 + Y_3 + Y_4) \\ \frac{1}{8} (-Y_1 + Y_2 - Y_3 - Y_4) \end{pmatrix}. \end{aligned}$$

Par ailleurs $\hat{\theta}_{LS} \sim \mathcal{N}(\theta, \sigma^2 (X^T X)^{-1} \sim \mathcal{N}\left(\theta, \sigma^2 \begin{pmatrix} 1/36 & 0 \\ 0 & 1/16 \end{pmatrix}\right)$.

3. En posant

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\theta}\|^2}{2} = \frac{\|(I_4 - X(X^T X)^{-1} X^T) Y\|^2}{2},$$

on a $\hat{\sigma}^2 \sim \sigma^2 \frac{\chi^2(2)}{2}$ (d'après le cours ou en utilisant le Théorème de Cochran).

4. D'après les questions précédentes, on a que

$$\frac{\|X(\hat{\theta}_{LS} - \theta)\|^2 / 2}{\hat{\sigma}^2} \sim \mathcal{F}(2, 2).$$

En notant q_1 le quantile d'ordre 95% d'une telle loi, on a

$$\mathbb{P}\left(\theta \in \{u \left| \frac{\|X(\hat{\theta}_{LS} - u)\|^2}{\hat{\sigma}^2} \leq q_1 \right.\right) = 95\%.$$

On en déduit que

$$\left\{ u \left| \frac{18}{\hat{\sigma}^2} (u_1 - (\hat{\theta}_{LS})_1)^2 + \frac{8}{\hat{\sigma}^2} (u_2 - (\hat{\theta}_{LS})_2)^2 \leq q_1 \right. \right\}$$

est une ellipsoïde de niveau de confiance 95% pour θ. Ses grands axes sont la base canonique de $\mathbb{R}^2$.

5. En notant $v = (3, -2)^T$, on a que

$$\begin{aligned} v^T \hat{\theta}_{LS} &\sim \mathcal{N}(u, \sigma^2 v^T (X^T X)^{-1} v) \\ &\sim \mathcal{N}(u, (\sigma^2/2)), \end{aligned}$$

et donc que

$$\frac{\sqrt{2}(v^T \hat{\theta}_{LS} - u)}{\hat{\sigma}} \sim \mathcal{T}(2).$$

En notant q_2 le quantile d'ordre 1% d'une $\mathcal{T}(2)$ (loi de Student à deux degrés de libertés), on en déduit que

$$\left[v^T \hat{\theta}_{LS} - \frac{\hat{\sigma}}{\sqrt{2}} q_2, v^T \hat{\theta}_{LS} + \frac{\hat{\sigma}}{\sqrt{2}} q_2 \right]$$

est un intervalle de niveau de confiance 98% sur u.

Exercice 3 - Bonus

Soient $X_1, \ldots, X_n$ des variables i.i.d., avec $X_1 \in L_2(\mathbb{P})$. On note $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ et $\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Montrer que

$$\bar{X}_n \perp\!\!\!\perp \hat{\sigma}_n^2 \quad \Leftrightarrow \quad X_1 \text{ est Gaussienne.}$$

Indication : on pourra calculer de deux manières différentes $\mathbb{E}\left(n\hat{\sigma}^2 e^{itn\bar{X}_n}\right)$.

Solution 3 -

Le sens indirect est une propriété de cours.

Regardons le sens direct (supposons donc $\bar{X}_n \perp\!\!\!\perp \hat{\sigma}_n^2$). Sans perte de généralité on suppose que X est centrée et $\sigma^2 := \mathrm{Var}(X_1) = 1$ (sinon on regarde $Y_1 = (X_1 - \mathbb{E}(X_1))/\sigma$, ce qui ne change rien aux propriétés d'indépendance ni au caractère Gaussien). Notons aussi ϕ la fonction caractéristique de X_1. Comme $\mathbb{E}(X_1^2) < +\infty$, on a que ϕ est $\mathcal{C}^2$, et vérifie, pour tout $t \in \mathbb{R}$,

$$\begin{aligned} \phi'(t) &= i\mathbb{E}(X e^{itX}) \\ \phi''(t) &= -\mathbb{E}(X^2 e^{itX}). \end{aligned}$$

Soit $t \in \mathbb{R}$, on peut alors écrire

$$\begin{aligned} \mathbb{E}\left(n\hat{\sigma}^2 e^{itn\bar{X}_n}\right) &= n\mathbb{E}(\hat{\sigma}^2)\mathbb{E}(e^{itn\bar{X}_n}) \qquad \text{(indépendance)} \\ &= n \times \frac{n-1}{n} \phi(t)^n \\ &= (n-1)\phi(t)^n. \end{aligned}$$

D'un autre côté, on peut décomposer

$$\begin{aligned}\mathbb{E}\left(n\hat{\sigma}^2 e^{itn\bar{X}_n}\right) &= \mathbb{E}\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 e^{it\sum_{j=1}^n X_j}\right)\\ &= \sum_{i=1}^n \mathbb{E}(X_i^2 e^{it\sum_{j=1}^n X_j}) - n\mathbb{E}(\bar{X}_n^2 e^{it\sum_{j=1}^n X_j}).\end{aligned}$$

Pour un i donné, on a

$$\begin{aligned}\mathbb{E}(X_i^2 e^{it\sum_{j=1}^n X_j}) &= \mathbb{E}(X_i^2 e^{itX_i})\phi(t)^{n-1} \qquad \text{(indépendance)}\\ &= -\phi''(t)\phi(t)^{n-1}.\end{aligned}$$

Par ailleurs

$$\begin{aligned}-n\mathbb{E}(\bar{X}_n^2 e^{it\sum_{j=1}^n X_j}) &= -\frac{1}{n}\mathbb{E}((\sum_{j=1}^n X_j)^2 e^{it\sum_{j=1}^n X_j})\\ &= \frac{1}{n}\psi''(t),\end{aligned}$$

où $\psi(t)$ est la fonction caractéristique de $n\bar{X}_n$, vérifiant $\psi(t) = \phi(t)^n$. On a alors $\psi''(t) = n(n-1)\phi'(t)^2\phi(t)^{n-2} + n\phi''(t)\phi(t)^{n-1}$. On en déduit

$$-n\mathbb{E}(\bar{X}_n^2 e^{it\sum_{j=1}^n X_j}) = (n-1)\phi'(t)^2\phi(t)^{n-2} + \phi''(t)\phi(t)^{n-1}.$$

On peut maintenant comparer les deux écritures. On a

$$\begin{aligned}(n-1)\phi(t)^n &= \mathbb{E}\left(n\hat{\sigma}^2 e^{itn\bar{X}_n}\right)\\ &= -n\phi''(t)\phi(t)^{n-1} + (n-1)\phi'(t)^2\phi(t)^{n-2} + \phi''(t)\phi(t)^{n-1},\end{aligned}$$

ou bien encore

$$\phi(t)^n + \phi''(t)\phi(t)^{n-1} - \phi'(t)^2\phi(t)^{n-2} = 0. \tag{1}$$

Comme $\phi(0) = 1$ et ϕ est continue, il existe un voisinage non vide $U_0 =]-t_0, t_0[$ de 0 tel que $\phi(t) \neq 0$ pour $t \in U_0$. Sur ce voisinage, on a

$$\frac{\phi''(t)\phi(t) - \phi'(t)^2}{\phi(t)^2} = -1,$$

ou encore

$$f''(t) = -1,$$

avec $f(t) = \log(\phi(t))$. Comme $f(0) = 0$ et $f'(0) = i\mathbb{E}(X) = 0$, on en déduit que $f(t) = -t^2/2$ sur U_0, ou encore $\phi(t) = e^{-t^2/2}$ sur U_0. Posons maintenant $t_+ = \sup\{t \mid \forall |u| < t \quad \phi(u) = e^{-u^2/2}\}$, et supposons $t_+ < +\infty$. Par continuité on a $\phi(t_+) = e^{-t_+^2/2}$, et donc que $\phi(t) \neq 0$ sur $]t_+ - \delta, t_+ + \delta[$. En utilisant le même raisonnement qu'auparavant on a que $\phi(t) = e^{-t^2/2}$ sur $]t_+ - \delta, t_+ + \delta[$, pour un $\delta > 0$. Symétriquement on a un autre $\delta > 0$ tel que $\phi(t) = e^{-t^2/2}$ sur $]-t_+ - \delta, -t_+ + \delta[$, ce qui est contradictoire avec la définition de t_+.

On en déduit alors que $\forall t \in \mathbb{R} \quad \phi(t) = e^{-t^2/2}$, et donc que $X_1 \sim \mathcal{N}(0, 1)$.