**Démarche statistique** **(à rendre au plus tard le dimanche 28/01/2024)**

# Devoir

*Enseignant:* Mohammed Sedki *pageweb:* masedki.github.io

**Instructions :**

Le devoir est à rendre sous format d'un seul fichier pdf (scans de vos notes) nommé `prenom_nom.pdf` ou `prenom1_nom1_prenom2_nom2.pdf` pour un binôme.

## Problème : Test du rapport des vraisemblances (10 points)

Supposons que $X$ et $Y$ soient des variables aléatoires continues représentant les temps de survie (en années) des patients suivant deux types différents d'interventions chirurgicales pour le traitement d'un cancer du côlon avancé. Les fonctions de densité associées aux variables aléatoires $X$ et $Y$ sont respectivement données par

$$f_X(x,\alpha) = \alpha e^{-\alpha x}, x > 0, \quad \alpha > 0 \quad \text{et} \quad f_Y(y,\beta) = \beta e^{-\beta y}, y > 0, \quad \beta > 0.$$

Soit $y_1, y_2, \ldots, y_n$ un échantillon de réalisations de même loi que $X$ et $y_1, y_2, \ldots, y_n$ un second échantillon de même loi que $Y$. On note $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ et $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$.

L'objectif de ce problème et de mettre en place un mécanisme statistique pour tester les hypothèses $H_0 : \alpha = \beta$ versus $H_1 : \alpha \neq \beta$. Pour cela, nous allons faire appel au test dit *de rapport de vraisemblances*. Les ingrédients indispensables à la réalisation de ce test sont

— La vraisemblance du jeu de données complet $x_1, \ldots, x_n, y_1, \ldots, y_n$ notée $\mathcal{L}(\gamma)$ sous l'hypothèse $H_0 : \alpha = \beta = \gamma$ (sous $H_0$, on considère les deux jeux de données de même loi de paramètre $\gamma$).

— La vraisemblance du jeu de données complet $x_1, \ldots, x_n, y_1, \ldots, y_n$ notée $\mathcal{L}(\alpha, \beta)$ sous l'hypothèse $H_1 : \alpha \neq \beta$ (sous $H_1$, les deux jeux de données proviennent de lois distinctes de paramètres $\alpha$ et $\beta$ respectivement).

— Les estimateurs par maximum de vraisemblance $\widehat{\alpha}, \widehat{\beta}$ et $\widehat{\gamma}$ des paramètres $\alpha, \beta$ et $\gamma$ respectivement.

— La statistique du rapport des vraisemblances donnée par [1]

$$\widehat{\lambda} = \frac{\mathcal{L}(\widehat{\gamma})}{\mathcal{L}(\widehat{\alpha}, \widehat{\beta})}.$$

— Un théorème de statistique mathématique : lorsque $n$ est grand, sous l'hypothèse $H_0$, la variable aléatoire $-2\ln\widehat{\lambda}$ suit la loi $\chi^2_1$ du khi-deux à 1 degré de liberté.

**Q.1** Vérifier que

$$\mathcal{L}(\gamma) = \gamma^{2n} e^{-n\gamma(\bar{x}_n + \bar{y}_n)} \quad \text{et} \quad \mathcal{L}(\alpha,\beta) = \alpha^n e^{-n\alpha\bar{x}_n} \beta^n e^{-n\beta\bar{y}_n}.$$

**Q.2** Calculer l'estimateur par maximum de vraisemblance $\widehat{\gamma}$ en maximisant $\mathcal{L}(\gamma)$.

**Q.3** Calculer les estimateurs par maximum de vraisemblance $\widehat{\alpha}$ et $\widehat{\beta}$ en maximisant $\mathcal{L}(\alpha, \beta)$.

**Q.4** Vérifier que

$$\mathcal{L}(\widehat{\gamma}) = e^{-2n}\left[\frac{2}{(\bar{x}_n + \bar{y}_n)}\right]^{2n} \quad \text{et} \quad \mathcal{L}(\widehat{\alpha}, \widehat{\beta}) = \left(\bar{x}_n \bar{y}_n\right)^{-n} e^{-2n}.$$

**Q.5** Déduire que

$$\widehat{\lambda} = \left[4u(1-u)\right]^n, \quad \text{où} \quad u = \frac{\bar{x}_n}{\bar{x}_n + \bar{y}_n}.$$

1. Dans cette expression, on remplace dans chaque expression de vraisemblance, le paramètre par son estimateur par maximum de vraisemblance.

**Q.6** Sachant que $n = 100$, $\bar{x}_n = 1.25$ années et $\bar{y}_n = 0.75$ années, calculer la valeur numérique de $\widehat{\lambda}$ et déduire la p-value donnée par

$$\mathbb{P}\big[\chi_1^2 > -2\ln\widehat{\lambda}\big],$$

pour conclure au rejet (ou pas) de $H_0$ en faveur de $H_1$.

**Q.7** Expliquer en quelques mots (3 lignes maximum) l'intuition à l'origine du test du rapport des vraisemblances.

**Problème : Modèle géométrique** (10 points)

Supposons que $n$ patients adultes de sexe masculin, sélectionnés au hasard et souffrant d'hypertension, se voient administrer un nouveau médicament d'hypotension au cours d'un essai clinique destiné à évaluer l'efficacité de ce nouveau médicament pour favoriser la rémission à long terme de l'hypertension artérielle. De plus, une fois que la pression artérielle de chaque patient est revenue à un niveau normal, chaque patient est examiné tous les mois pour voir si l'hypertension réapparaît. Pour le ième patient de l'étude, $x_i$ représente l'âge du patient au début de l'essai clinique et $Y_i$ est la variable aléatoire représentant le nombre de mois de suivi jusqu'à la première réapparition de l'hypertension. Il est raisonnable de supposer que $Y_i$ suit une loi géométrique.

$$p_{Y_i}(y_i;\theta_i) = (1-\theta_i)^{y_i-1}\theta_i, \quad y_i = 1, 2, \dots, \infty, \quad 0 < \theta_i < 1 \quad \text{et} \quad i = 1, 2, \dots, n.$$

Il est bien établi que l'âge est un facteur de risque d'hypertension. Pour tenir compte des différences d'âge des patients au début de l'essai, il est proposé d'exprimer $\theta_i$ comme la fonction suivante de l'âge :

$$\theta_i = \frac{\beta x_i}{1+\beta x_i}, \quad \beta > 0.$$

Étant donné les $n$ paires d'observations $(x_i, y_i), i = 1, 2, \dots, n$, l'objectif de l'analyse est d'utiliser l'estimateur par maximum de vraisemblance $\widehat{\beta}$ du paramètre $\beta$ pour répondre à des questions d'inférence statistique autour du paramètre $\beta$.

**Q.1** Vérifier que [2]

$$\ln\mathcal{L}(\beta) = \sum_{i=1}^{n}\Big[\ln(\beta x_i) - y_i\ln(1+\beta x_i)\Big].$$

**Q.2** Déduire que l'estimateur par maximum de vraisemblance $\widehat{\beta}$ vérifie l'équation

$$\widehat{\beta} = \frac{n}{\sum_{i=1}^{n} x_i y_i \Big(1+\widehat{\beta}x_i\Big)^{-1}}.$$

**Q.3** Sachant que $\mathbb{E}(Y_i) = \theta_i^{-1} = \frac{1+\beta x_i}{\beta x_i}$, vérifier que l'information de Fisher associée à l'échantillon entier est donnée par

$$\mathcal{I}_n(\beta) = \frac{1}{\beta^2}\sum_{i=1}^{n}\big(1+\beta x_i\big)^{-1}.$$

**Q.4** Déduire la variance asymptotique (lorsque $n$ est grand) de l'estimateur $\widehat{\beta}$.

**Q.5** Si l'essai clinique concerne 50 patients âgés de 30 ans et 50 patients âgés de 40 ans au début de l'essai, proposer un intervalle de confiance à 95% de confiance pour $\beta$ lorsque $\widehat{\beta} = 0.5$.

2. Calculer la vraisemblance $\mathcal{L}(\beta)$ en premier !