Exercice 1

Les kératoses actiniques sont de petites lésions cutanées qui servent de précurseurs au cancer de la peau. Selon certaines théories, les adultes résidant dans des villes américaines proches de l'équateur sont plus susceptibles de développer des kératoses actiniques, et donc d'avoir un risque plus élevé de cancer de la peau, que les adultes résidant dans des villes américaines proches de l'équateur. risque de cancer de la peau, que les adultes résidant dans des villes américaines éloignées de l'équateur. Pour tester cette théorie, supposons une collecte de données comme suit

- les dossiers dermatologiques d'un échantillon aléatoire de n_1 résidents adultes d'une ville américaine particulière (disons, Ville 1) proche de l'équateur sont examinés pour déterminer le nombre de kératoses actiniques que chacun de ces n_1 adultes a développé.
- Les dossiers de dermatologie d'un échantillon aléatoire de n_0 adultes résidant dans une ville ville américaine particulière (disons, Ville 0) éloignée de l'équateur sont examinés pour déterminer le nombre de kératoses actiniques que chacun de ces adultes a développé

Comme modèle statistique pour évaluer cette théorie, pour un résident adulte j ($j = 1, \ldots, n_i$) dans la ville i ($i = 0, 1$), supposons que la variable aléatoire $Y_{ij} \sim \mathcal{P}\big(L_{ij}\lambda_i\big)$, où L_{ij} est la durée (en années) pendant laquelle l'adulte j a résidé dans la Ville i et où λ_i est le taux de développement des kératoses actiniques par an (c'est-à-dire le nombre attendu de kératoses actiniques qui se développent par an) pour un adulte résidant dans la ville i. Ainsi, la paire $\big(L_{ij}, y_{ij}\big)$ constitue l'information observée pour un résident adulte j dans la ville i.

1. Développer un intervalle de confiance à $100(1-\alpha)\%$ basé sur la normalité asymptotique de l'estimateur par maximum de vraisemblance pour le logarithme du rapport de taux $\ln \psi = \ln \frac{\lambda_1}{\lambda_0}$.
2. Si $n_1 = 30$ et $n_0 = 30$ et $\sum_{j=1}^{n_1} y_{1j} = 40$, $\sum_{j=1}^{n_1} L_{1j} = 350$, $\sum_{j=1}^{n_0} y_{0j} = 35$, et $\sum_{j=1}^{n_0} L_{0j} = 400$, calculer un IC à 95% pour le rapport de taux ψ. Commenter !

Exercice 2

Des chercheurs mènent une étude sur n nourrissons afin de déterminer si les nourrissons placés dans des crèches sont plus susceptibles d'être en surpoids que ceux qui sont gardés à la maison. Les nourrissons sont déclarés comme étant "en surpoids" s'ils se situent dans le 85e percentile ou plus sur la courbe de croissance de l'indice de masse corporelle (IMC) ajustée à l'âge et au sexe, établie par les Centers for Disease Control and Prevention (CDC).

$Y_i = 1$ ($i = 1, \ldots, n$) si le ième enfant est en surpoids et $Y_i = 0$ s'il ne l'est pas. Il est supposé que Y_i suit une loi de Bernoulli. Pour étudier l'association entre le mode de garde du nourrisson et la probabilité qu'il soit en surpoids, une modélisation par régression logistique est proposé :

$$\pi_i = \pi(x_i) = \mathbb{P}\left(Y_i = 1 | x_i\right) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}, \quad i = 1, \ldots, n,$$

où $x_i = 1$ si le ième nourrisson est gardé à la crèche et $x_i = 0$ si le ième nourrisson est gardé à la maison, et α et β sont des paramètres inconnus à estimer.

Supposons que les données collectés sont comme suit $(y_1, x_1 = 0), (y_2, x_2 = 0), \ldots, (y_{n_0}, x_{n_0} = 0), (y_{n_0+1}, x_{n_0+1} = 1), (y_{n_0+2}, x_{n_0+2} = 1), \ldots, (y_n, x_n = 1)$

1. Montrer que les estimateurs par maximum de vraisemblance des paramètres α et β sont données par

$$\widehat{\alpha} = \ln\left(\frac{p_0}{1-p_0}\right) \quad \text{et} \quad \widehat{\beta} = \ln\left[\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right],$$

où $p_0 = n_0^{-1} \sum_{i=1}^{n_0} y_i$ et $p_1 = n_1^{-1} \sum_{i=n_0+1}^{n} y_i$.

2. Proposer une approximation de la matrice de variance covariance asymptotique associée au couple d'estimateurs $\left(\widehat{\alpha}, \widehat{\beta}\right)$.
3. Le jeu de données est composé d'observations 100 nourrissons bénéficiant de garde à domicile, dont 18 sont en surpoids, et 100 autres nourrissons en crèche, dont 26 sont en surpoids. Calculer des IC à 95% pour α et β sur un large échantillon. Les données fournissent-elles une preuve statistique que les nourrissons gardés en crèche sont plus susceptibles d'être en surpoids que les nourrissons gardés à domicile ?

Exercice 3

Pour $\alpha, \beta > 0$, on pose, $\theta = (\alpha, \beta)$ et on définit la densité de probabilité f_θ par

$$f_\theta(x) = \alpha \beta^{-\alpha} x^{\alpha-1} \mathbb{1}_{[0,\beta]}(x).$$

On donne

$$\mathbb{E}\big[X\big] = \frac{\alpha\beta}{\alpha+1}, \quad \mathrm{Var}\big(X\big) = \frac{\alpha\beta^2}{(\alpha+1)^2(\alpha+2)}.$$

On pose $U = -\alpha \log\left(\frac{X}{\beta}\right)$. La variable aléatoire U suit une la loi $\Gamma(1,1)$ (i.e exponentielle de paramètre 1). On observe un échantillon i.i.d $X_1, X_2, \ldots, X_n$ de loi de densité f_θ.

1. Suggérer un estimateur de α. Indication : on posera $c = \dfrac{\Big(\mathbb{E}\big[X\big]\Big)^2}{\mathrm{Var}\big(X\big)}$.

2. **On suppose maintenant que β est connu**. Déterminer l'estimateur par maximum de vraisemblance $\widehat{\alpha}_n$ de α.
3. Montrer sans calcul que l'estimateur $\widehat{\alpha}_n$ est biaisé. Pour cela, on exprimera $\widehat{\alpha}_n$ en fonction de la moyenne empirique des variables $U_i = -\alpha \log\left(\frac{X_i}{\beta}\right)$ et on fera appel à l'inégalité de Jensen qu'on peut énoncer comme suit. Soit g une fonction convexe et X une variable aléatoire telle que $\mathbb{E}\big[g(X)\big]$ existe : alors, $g\Big(\mathbb{E}[X]\Big) \leq \mathbb{E}\big[g(X)\big]$. L'inégalité précédente est stricte lorsque la fonction g est non linéaire. Dans notre cas, il suffit de poser $g(x) = \frac{1}{x}$ et utiliser l'inégalité $\mathbb{E}\big[\frac{1}{X}\big] > \frac{1}{\mathbb{E}[X]}$.

4. Montrer que l'estimateur $\tilde{\alpha}_n = \frac{n-1}{n}\widehat{\alpha}_n$ est sans biais. Indication : on sait que $\sum_{i=1}^n U_i$ suit une loi gamma $\Gamma(n,1)$ [1].

5. **Supposons maintenant que β est inconnu**. Déterminer l'estimateur par maximum de vraisemblance $\widehat{\beta}_n$ de β. Indications : vérifier à l'aide d'un graphique que la vraisemblance est nulle si $\beta \in \big[0, \max_{1\leq i\leq n} X_i\big[$ et décroissante sur $\big[\max_{1\leq i\leq n} X_i, +\infty\big[$.
6. L'estimateur $\widehat{\beta}_n$ est-il biaisé ?. Indication : examiner la probabilité $\mathbb{P}\big(\widehat{\beta}_n < \beta\big)$.

1. Rappelons la définition de la fonction gamma d'Euleur $\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} \mathrm{d}x$, pour tout $a > 0$.

Exercice 4

On suppose que la durée T (en mois) de la rémission des patients atteints de leucémie qui ont suivi un certain type de traitement par chimiothérapie suite une loi exponentielle

$$f_T(t;\theta) = \theta e^{-\theta t}, \quad t > 0, \theta > 0.$$

Supposons que le suivi d'un échantillon aléatoire de n patients atteints de leucémie ayant suivi ce traitement de chimiothérapie conduise aux n temps de rémission observés $t_1, t_2, \ldots, t_n$.

1. Calculer l'expression explicite de la variance (lorsque n est grand) de l'estimateur par maximum de vraisemblance $\widehat{\theta}_n$ de θ.
2. Un biostatisticien chargé d'analyser ces données se rend compte qu'il n'est pas possible de connaître avec certitude le nombre exact de mois pendant lesquels chaque patient est en rémission après avoir terminé le traitement de chimiothérapie. Ainsi, ce biostatisticien suggère la procédure alternative suivante pour estimer θ : Après un certain temps spécifié (période en mois) de longueur t^* (une constante positive connue) après la fin du traitement de chimiothérapie, soit $Y_i = 1$ si le ième patient est toujours en rémission après t^* mois et soit $Y_i = 0$ sinon. Nous avons

   $$\mathbb{P}(Y_i = 1) = \mathbb{P}(T_i > t^*), i = 1, 2, \ldots, n.$$

   Calculer un second estimateur $\widehat{\theta}_n^*$ par maximum de vraisemblance pour θ basé sur l'observation de l'échantillon $Y_1, Y_2, \ldots, Y_n$.
3. En supposant que $t^* \geq \mathbb{E}(T)$, lequel des deux estimateurs $\widehat{\theta}_n$ et $\widehat{\theta}_n^*$ a la plus petite variance, et pourquoi cela devrait-il être le résultat attendu ? Y a-t-il des circonstances où l'estimateur avec la plus grande variance pourrait être préféré ?

Exercice 5

Un économiste postule que la distribution des revenus (en milliers de dollars) dans une certaine grande ville américaine peut être modélisée par la loi de Pareto de fonction de densité

$$f_Y(y;\gamma,\theta) = \theta\gamma^\theta y^{-(\theta+1)}, \quad 0 < \gamma < y < +\infty \quad \text{et} \quad 2 < \theta < +\infty,$$

où γ et θ sont des paramètres inconnus. Supposons que $Y_1, Y_2, \ldots, Y_n$ est un échantillon suivant la loi de Pareto.

1. Si $n = 50$, $\bar{y}_n = n^{-1}\sum_{i=1}^n y_i = 30$, et $s^2 = (n-1)^{-1}\sum_{i=1}^n \left(y_i - \bar{y}_n\right)^2 = 10$, calculer la valeur numérique exacte des estimateurs par la méthode des moments $\widehat{\gamma}_{mm}$ et $\widehat{\theta}_{mm}$ de γ et θ.
2. Un statisticien suggère que la statistique d'ordre, le plus petit $Y_{(1)} = \min\{Y_1, Y_2, \ldots, Y_n\}$ est également un estimateur possible pour γ. $Y_{(1)}$ est-il un estimateur qyu converge en probabilité [2] vers γ ?
3. Supposons maintenant que $\theta = 3$ et γ est le seul paramètre inconnu. Il est souhaitable d'utiliser la variable aléatoire$Y_{(1)}$ pour calculer un intervalle de confiance unilatéral supérieur exact pour γ. En particulier, calculer l'expression explicite de la variable aléatoire $U = cY_{(1)}, 0 < c < 1$, telle que $\mathbb{P}(\gamma < U) = (1-\alpha), 0 < \alpha < 0.1$. Si $n = 5$, $\alpha = 0.1$ et la valeur observée de $Y_{(1)}$ est $y_{(1)} = 20$, utiliser ces données pour calculer un intervalle de confiance unilatéral supérieur pour le paramètre inconnu γ.

2. Il suffit de montrer que l'estimateur asymptotiquement sans biais et sa variance tend vers 0 quand n tend vers l'infini.