TP 10 : Intervalles de confiance, normalité asymptotique, tests

Ce TP 10 continue de faire travailler sur la construction d'intervalles de confiance, la normalité asymptotique, et les tests classiques avec R. Veillez à bien justifier vos calculs et à commenter vos résultats.

Exercice 1. Taux d'incidence dans deux sous-populations.

On considère deux populations A et B disjointes, dont tous les individus sont testés pour détecter la présence d'un virus. La sous-population A est composée de 730 individus, dont 12 sont positifs au test. La sous-population B est composée de 1256 individus, dont 3 sont positifs au test. On souhaite répondre à la question suivante : la différence du taux d'incidence entre ces deux sous-populations est-elle significative ?

1. Modéliser le problème : on introduira deux paramètres p_A et p_B. Quel est le test que l'on souhaite faire ? (préciser l'hypothèse nulle et l'hypothèse alternative).

2. Pour $(X_i)_{1\leq i\leq n}$ des variables i.i.d. de Bernouilli de paramètres p, on note $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Quelle est la limite en loi de

$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \quad ?$$

(On justifiera proprement). En déduire, pour tout $\alpha \in [0,1]$, un intervalle de confiance de niveau $1-\alpha$ pour le paramètre p. Cet intervalle est-il exact/par excès, asymptotique/non-asymptotique ?

3. Calculer et afficher les intervalles de confiance de niveau 95% pour les paramètres p_A et p_B. Répondre à la question posée initialement.

Exercice 2. Maximum de vraisemblance et intervalle de confiance

On reprend les powers laws vues dans le TP 6: une variable X à valeurs dans $[1,+\infty[$ suit une loi $PL(a)$ si elle a pour densité $f_a(x) := (a-1)x^{-a}\mathbf{1}_{x\in[1,+\infty[}$. On rappelle que pour $a>1$ et $X_1,\dots,X_n$ un échantillon i.i.d. de loi $PL(a)$, l'estimateur du maximum de vraisemblance de a était donné par :

$$\hat{a} := 1 + \left(\frac{1}{n}\sum_{i=1}^n \log X_i\right)^{-1}.$$

On admet de plus que, lorsque n est grand, on a la convergence p.s. $\hat{a}\to a$, ainsi que la convergence en loi suivante :

$$\sqrt{n}\frac{\hat{a}-a}{a-1} \to \mathcal{N}(0,1).$$

1. Pourquoi peut-on aussi affirmer que

$$\sqrt{n}\frac{\hat{a}-a}{\hat{a}-1} \to \mathcal{N}(0,1) \quad ?$$

En déduire, pour tout $\alpha\in[0,1]$, un intervalle de confiance de niveau $1-\alpha$ pour le paramètre p. Cet intervalle est-il exact/par excès, asymptotique/non-asymptotique ?

2. On rappelle que si U est uniforme sur $[0,1]$ alors $U^{-1/(a-1)} \sim PL(a)$. Simuler un échantillon de taille n de loi $PL(a)$ avec $n = 1000, a = 2.2$, et calculer et afficher un intervalle de confiance à 99% pour le paramètre a.

3. On reprend le même jeu de données moby_dick.csv, que l'on ouvrira de la façon suivante :

```
data = read.csv("moby_dick.csv")
X=data$occurrences
```

Donner un intervalle de confiance à 95% pour le paramètre a de la loi $PL(a)$ qui fitte le mieux les données.

Exercice 3. Le Titanic : à vous de jouer

Le jeu de données que l'on peut importer de la façon suivante :

```
titanic <- read.csv("https://github.com/pmagwene/Bio723/raw/master/datasets/titanic.csv")
```

donne le profil des passagers du Titanic (âge, classe, ville d'embarquement, sexe, numéro de chambre, etc.) ainsi que leur survie ou non. On souhaite répondre aux questions suivantes :

- L'âge des passagers a-t-il joué sur leurs chances de survies ? Si oui, dans quel sens ?
- La classe dans laquelle se situaient les passagers a-t-elle joué sur leurs chances de survies ? Si oui, dans quel sens ?

A vous de jouer pour répondre à ces deux questions.