Exercices de statistique avec le logiciel R (S5)

Exercice 9. Intervalles de confiance
a) Générer $n = 10$ i.i.d d'une v.a. $\mathcal{N}(0,1)$. Donner un intervalle de confiance bilatéral de niveau 99% sur la moyenne, la variance étant supposée inconnue elle aussi.
b) Estimer le niveau de confiance par le niveau empirique. Le niveau empirique correspond à la fréquence de présence du vrai centre (ici 0) dans l'intervalle de confiance. Conclusion ?
c) Procéder de même avec la variance supposée connue.

Exercice 10. Tests d'adéquation
a) Réecrire le test d'adéquation du χ^2 à une loi normale en considérant les deux paramètres moyenne et variance comme inconnus.
b) On considère un échantillon $x_1, x_2, \ldots, x_n$ de $\mathcal{N}(0,1)$.
· Tester l'hypothèse d'adéquation à une loi normale. On utilisera trois tests différents : χ^2, Kolomogorov-Smirnov et Shapiro-Wilk. On prendra n+100 et un risque $\alpha = 1\%$. On pourra prendre le découpage en classes suivant pour le test du χ^2 : $(-\infty, -1)$,$[-1, 0)$,$[0, 1)$ et $[1, \infty)$.
· Quel est le risque empirique associé à chaque test ?
c) On considère maintenant un échantillon $x_1, \ldots, x_n$ de $\mathcal{U}_{[-2,2]}$. Dans le cadre de l'hypothèse d'adéquation à une loi normale, estimer la puissance (pour n=100) associée à chaque test quand $\alpha = 5\%$. Classer alors les trois tests par ordre de préférence.

Exercice 11. Tests d'adéquation à une loi à queue lourde
a) Créer trois fonctions en R qui calculent:
(1) $F(x) = \begin{cases} 0, & x < 1 \\ 1 - \frac{1}{x^p} & x \geq 1 \end{cases}$ (fonction **pareto(x, p)**)
(2) $F^{-1}(x) = (1-x)^{-1/p}, 0 < x < 1$ (fonction **qpareto(x, p)**)
(3) $R(n) = F^{-1}(Unif(n))$, (fonction **rpareto(n, p)**)
b) Réecrire le test d'adéquation du χ^2 à une loi de Paréto avec $p = 3$ en considérant p connu.
c) On considère un échantillon $x_1, x_2, \ldots, x_n$ de loi de Paréto avec $p = 3$.
· Tester l'hypothèse d'adéquation à la loi de Paréto avec $p = 3$. On utilisera deux tests différents : χ^2 et Kolomogorov-Smirnov. On prendra $n = 100$ et un risque $\alpha = 1\%$. On pourra prendre le découpage en classes suivant pour le test du χ^2 : $(-\infty, 1.5)$,$[1.5, 2)$ et $[2, \infty)$.

· Quel est le risque empirique associé à chaque test ?
d) On considère maintenant un échantillon $x_1, \ldots, x_n$ de $\mathcal{U}_{[1,3]}$. Dans le cadre de l'hypothèse d'adéquation à la loi de Paréto avec $p = 3$, estimer la puissance (pour n=100) associée à chaque test quand $\alpha = 5\%$. Quel test préférez vous ?

Exercice 12. Régression

On cosidère le modèle de régression $Y = 10 + x + 5\sin(x) + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, 4)$.
a) Pour des valeurs $x_i = i$, $(i = 1, \ldots, 30)$, générer un échantillon $y_1, \ldots, y_{30}$ i.i.d. à partir du modèle précédent. Tracer les couples de points (x_i, y_i) ainsi que la vraie courbe.
b) On suppose maintenant ignorer le modèle qui a généré les données.
· Estimer par moindres carrés les paramètres du modèle $Y = a + bx + \epsilon$. Tracer la courbe estimée sur le graphique précédent. Vérifier graphiquement la normalité des résidus. Que peut-on supposer ?
· Estimer par moindres carrés les paramètres du modèle $Y = a + bx + c\sin(x) + dx^2 + \epsilon$. Ajouter la courbe estimée au graphique précédent. Faire le test de nulité des coefficients. Que laisse supposer ce test ?
· Estimer par moindres carrés les paramètres du modèle $Y = a + bx + c\sin(x) + \epsilon$. Ajouter la courbe estimée au graphique précédent.
· Sélectionner le meilleur modèle au sens du critère AIC. Conclusion ? Aurait-on pu choisir le modèle avec le critère du R^2 ? Du $\bar{R}^2$? Essayer.
· Vérifier graphiquement la normalité des résidus de ce modèle, les résidus studentisés, la distance de Cook.