Mini-Projet

Consignes

La date limite pour rendre le projet est le 30 Octobre 2020 à 23h59. Le language à utiliser est `Python` et le rendu est attendu sous la forme d'un `Python` notebook téléchargé sur plateforme `eCampus` : https://ecampus.paris-saclay.fr/enrol/index.php?id=33526.

Rappel

Pour la reproductibilité des questions numériques, il est conseillé de fixer la « graine » du générateur de nombres pseudo-aléatoires.

Rappels et compléments de cours :

— La p-valeur est la probabilité que, sous l'hypothèse nulle, la statistique de test prenne une valeur au moins aussi extrême que celle qui a été observée.

— La fonction puissance est la probabilité de rejeter sous l'hypothèse alternative H_1 : $h(\tilde{\lambda}) = \mathbb{P}[T \in \bar{A} \,|\lambda = \tilde{\lambda}\,]$ pour $\lambda \in \Lambda_{H_1}$ où T est la statistique de test, $\bar{A}$ et la région de rejet, λ est le paramètre à tester et Λ_{H_1} signifie l'ensemble des paramètres appartenant à la région de l'hypothèse alternative.

— Pour k entier positif, la fonction de répartition d'une loi Gamma ($X \sim Gamma(k, \theta)$) peut être formulée comme

$$F(x; k, \theta) = \mathbb{P}[X < x] = 1 - e^{-\frac{x}{\theta}} \sum_{i=0}^{k-1} \frac{1}{i!} \Big(\frac{x}{\theta}\Big)^i .$$

— Pour $X \sim \mathcal{E}(\lambda)$ la fonction caractéristique est $\phi_X(t) = \frac{1}{1-\frac{it}{\lambda}}$.
Pour $X \sim Gamma(k, \theta)$ la fonction caractéristique est $\phi_X(t) = \frac{1}{(1-it\theta)^k}$.

Exercice 1 (Exploration des données, recherche de leur loi)**:**
On s'intéresse aux coût d'accidents nucléaire avant l'accident de Three Mile Island qui s'est produit le 28 mars 1979. Cet exercice est consacré à l'exploration des données et la recherche d'un modèle statistique pertinent. On utilisera la base de données accessible à l'adresse :
https://xyotta.com/v1/index.php/Nuclear_events_database

1. Télécharger les données des accidents nucléaires en utilisant le lien suivant https://innovwiki.ethz.ch/v1/images/NuclearPowerAccidents2016.csv.
   D'une manière automatique et en utilisant `Python`, former un vecteur des coût des accidents (strictement) avant l'accident de Three Mile Island, en million dollars 2013 et supprimer toutes les observations (avec données) manquantes. Vous devez obtenir $n = 55$ observations $x_1, ..., x_n$.
2. Construction d'un QQ-plot normal. On pourra consulter la page https://fr.wikipedia.org/wiki/Diagramme_Quantile-Quantile pour une explication détaillée sur les QQ-plots.

(a) Montrer que la fonction quantile d'une loi normale $\mathcal{N}(\mu, \sigma^2)$, notée $F^{-1}(p; \mu, \sigma^2)$, vérifie

$$\forall\, p \in]0,1[\,, \quad F^{-1}(p; \mu, \sigma^2) = \mu + \sqrt{\sigma^2} F^{-1}(p; 0, 1)\,.$$

On a montré que les quantiles de la loi normale avec les paramètres arbitraires et ceux de la loi normale centrée réduite sont linéairement dépendants. Ainsi, pour toute loi normale de paramètres inconnus, il existe $(a, b) \in \mathbb{R} \times \mathbb{R}_+$ tels que $F^{-1}(p; \mu, \sigma) = a + bF^{-1}(p; 0, 1)$. Cela suggère la méthode diagnostique suivante : si les données proviennent d'une loi normale, et si on trace les quantiles empiriques des données et les quantiles correspondants de la loi normale centrée réduite, le graphique devrait ressembler à une droite. Souvent, pour éviter l'étape d'estimation des paramètres, on trace la droite passant par deux points du graphe, ceux correspondant aux quantiles empiriques 0.25 et 0.75 par exemple.

(b) Tracer le QQ-plot de données pour la loi normale (vous pouvez utiliser la fonction `probplot` de la librairie `SciPy`).

3. On considère maintenant le modèle des lois exponentielles

(a) Montrer que pour un quantile d'une loi exponentielle de paramètre λ on a

$$\forall\, p \in (0,1)\,, \quad F^{-1}(p; \lambda) = \frac{1}{\lambda} F^{-1}(p; 1)\,.$$

Si les observations suivent la loi exponentielle, le QQ-plot doit ressembler à une droite.

(b) Tracer le QQ-plot de données pour la loi exponentielle.

4. Discuter, en vu de QQ-plots obtenus, quelle loi semble être plus plausible pour les données.

Exercice 2 (Estimation ponctuelle des paramètres d'une loi exponentielle)**:**
Selon Wheatley, Sovacool, and Sornette (2017), on peut utiliser le modèle des lois exponentielles pour modéliser les coût des accidents avant l'accident de Three Mile Island.

Remarque : *Après l'accident les données suivent une loi de Pareto et ne sont pas le sujet du projet. Il est connu que amélioration de sécurité permet d'éviter les événement modérés, mais souvent au détriment des événement extrêmes occasionnels.*

Il est acceptable de supposer que les accident sont indépendants. Les observations à disposition sont les n coût des accidents, $X = (X_1, \ldots, X_n)$, où les X_i sont indépendants et identiquement distribués, de loi exponentielle $\mathcal{E}(\lambda)$ donnée par

$$\mathrm{P}_\lambda(]x, \infty[) = \mathbb{P}_\lambda(X_1 > x) = \begin{cases} e^{-\lambda x} & (x \geq 0) \\ 1 & (x < 0), \end{cases}$$

où $\lambda > 0$ est le paramètre (inconnu) du modèle.

1. Calculer l'estimateur du maximum de vraisemblance $\widehat{\lambda}_n$ pour le paramètre de la loi exponentielle λ.
2. Tracer sur le même graphique l'histogramme de données (sur l'échelle de densité) et la densité de probabilité de la loi exponentielle avec le paramètre $\widehat{\lambda}_n$.

3. On cherche à estimer la grandeur d'intérêt $g_1(\lambda) = \frac{1}{\lambda}$. On admet que le modèle $\{\mathrm{P}_\lambda, \lambda > 0\}$ est régulier, au sens des hypothèses du théorème de Cramér-Rao. On note $T_1(X) = \frac{1}{n}\sum_{i=1}^n X_i$. Montrer que la statistique $T_1(X)$ est un estimateur efficace pour $g_1(\lambda)$.
4. Calculer g_1 en utilisant T_1 pour l'échantillon donné.
5. Soit $\eta > 0$. On considère le nouvel estimateur

$$\widetilde{T}_{1,\eta}(X) = \eta T_1(X)\,.$$

Montrer que pour certaines valeurs de η (que vous préciserez), et pour le risque quadratique, on a

$$\forall\, \lambda > 0\,,\ R(\lambda, \widetilde{T}_{1,\eta}) < R(\lambda, T_1)\,.$$

Pourquoi ce résultat n'est-il pas en contradiction avec la question précédente ?

6. Pour quelle valeur de η l'estimateur $\widetilde{T}_{1,\eta}$ est l'estimateur sans biais de la médiane. Calculer cette estimateur pour l'échantillon donné. Comparer avec la médian empirique.
7. Pour la valeur de η de la question précédente, comparer les risques quadratiques des estimateurs T_1 et $\widetilde{T}_{1,\eta}$ comme estimateurs pour $g_1(\lambda)$, en fonction de la taille d'échantillon n.

Exercice 3 (Test sur le paramètre d'une loi)**:**
On considère le modèle $\{\mathrm{P}_\lambda, \lambda > 0\}$ des lois exponentielles. On souhaite affirmer avec un faible risque d'erreur que le coût moyen d'un accident est inférieur à un milliard de dollars.

1. Formuler l'hypothèse null et l'hypothèse alternative.
2. En utilisant le principe de Neyman-Pearson donner le test le plus puissant pour le niveau α.
3. Appliquer le test pour l'échantillon considéré au niveau $\alpha = 0.05$ et donner une réponse à la question "Peut-on affirmer que le coût moyen d'accident est inférieur à un milliard de dollars ?" Donner la p-valeur.
4. Pour la taille de l'échantillon $n = 55$ et la valeur de λ associée au coût moyen d'un accident égal à un milliard de dollars, tracer la densité de probabilité de T_1 et indiquer la région du rejet au niveau $\alpha = 0.05$.
5. Tracer la fonction puissance de test pour le niveau α en fonction de λ pour l'échantillon de la taille $n = 10, 50, 100, 500, 100\,000$. Expliquer vos résultats.
6. En utilisant théorème centrale limite, donner une approximation de la loi de T_1 et proposer un nouveau test. Donner le résultat de ce test.

Références

S. Wheatley, B. Sovacool, and D. Sornette. Of disasters and dragon kings : A statistical analysis of nuclear power incidents and accidents. *Risk Analysis*, 37(1), 99–115, 2017.