Mini-Projet

Consignes

La date limite pour rendre le projet est le 21 Octobre à 23h59. Le language à utiliser est `R` et le rendu est attendu sous la forme d'un notebook réalisé avec `R-studio` ou Jupyter notebook (avec le noyau `R`).

Rappel

Pour la reproductibilité des questions numériques, il est conseillé de fixer la « graine » du générateur de nombres pseudo-aléatoires, en haut de votre script, en utilisant la fonction `set.seed` de R, par exemple :

```
set.seed(1,kind="Marsaglia-Multicarry")
```

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 `R`, 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) Représenter en abscisse les quantiles d'une $\mathcal{N}(0, 1)$ de niveaux $(\frac{1}{2n}, \frac{3}{2n}, \frac{5}{2n}, ...\frac{2n-1}{2n})$ et en ordonnées les valeurs de l'échantillon rangées par ordre croissant (n la taille de l'échantillon). Superposer la droite affine qui passe par les quantiles 0.25 et 0.75.
   (c) Répéter en utilisant les fonctions `R` `qqnorm` et `qqline`. Vous devez obtenir exactement le même résultat.
   (d) Discuter si le modèle des lois normales vous paraît adapté.
3. On considère maintenant le modèle des lois exponentielles
   (a) Montrer que pour un quantile d'une loi exponentielle de paramètre λ on a
   $$\forall\, p \in (0,1)\,, \quad F^{-1}(p; \lambda) = \frac{1}{\lambda} F^{-1}(p; 1)\,.$$
   (b) Répéter (2.b) pour la loi exponentielle $\mathcal{E}(1)$.
   (c) Représenter en abscisse les valeurs de l'échantillon rangées par ordre décroissant et en ordonnées 0 pour toutes observations.
   (d) Construire un histogramme (en densité) des données observées à l'aide de la fonction `R` `hist`. Superposer la courbe de la densité de la loi $\mathcal{E}\Big(\frac{n}{\sum_{i=1}^n x_i}\Big)$.
   (e) Discuter si une loi exponentielle semble être plus plausible qu'une loi normale.

Exercice 2 (Estimation ponctuelle des paramètres d'une loi exponentielle)**:**
Selon Wheatley, Sovacool, and Sornette (2017), on peut utliser le modèle des losi 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, $X = (X_1, \dots, X_n)$, 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. 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 $T_1(X) = \frac{1}{n}\sum_{i=1}^n X_i$. Monter que la statistique $T_1(X)$ est un estimateur UVMB (uniformément de variance minimale parmi les estimateurs sans biais) de g_1.
2. Soit $\alpha > 0$. On considère le nouvel estimateur

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

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

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

   Pourquoi ce résultat n'est-il pas en contradiction avec la question précédente ?
3. Pour le même modèle, on cherche à estimer la médiane $g_2(\lambda) = \frac{\log 2}{\lambda}$. En utilisant $\varphi(x) = x$ avec les notations du cours, construire un estimateur $T_2(X)$ de g_2 par la méthode des moments. Au vu de la taille de l'échantillon considéré, peut-on dire que le risque quadratique du T_2 comme un estimateur de g_1 est inférieur à celui du T_1 pour tout $\lambda > 0$?
4. Estimer g_1 en utilisant T_1 pour l'échantillon donné.
5. Estimer g_2 en utilisant T_2 pour l'échantillon donné. Comparer à la médiane empirique.
6. Tracer les risques quadratiques de T_1 et T_2 comme estimateurs de g_1, sur le même graphique, en fonction de N la taille de l'échantillon, avec N variant de 1 à $n = 55$. On prendra valeur de λ celle de l'estimateur obtenu à la question (4.). Expliquer le graphique observé.

Exercice 3 (Test sur le paramètre d'un loi)**:**
On considère le modèle $\{\mathrm{P}_\lambda, \lambda > 0\}$ des lois exponentielle. On souhaite affirmer avec un faible risque d'erreur que le coût moyen d'un accident est inférieur à un milliard de dollars. Ceci équivaut à rejeter H_0 du test suivant :

$$H_0 \,:\, \lambda = \lambda_0 \quad vs. \quad H_1 \,:\, \lambda > \lambda_0\,.$$

On utilise comme statistique de test :

$$T(X) = \sum_{i=1}^n X_i\,.$$

1. Construire formellement la règle de décision pour le test au niveau α. (Commencer par montrer que la loi de $T(X)$ est une loi Gamma).

2. En R, 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 ?" (c'est-à-dire, peut-on rejeter H_0 pour $\lambda_0 = 0.001$?). Donner la p-valeur associée à la réalisation de la statistique de test $t = \sum_{i=1}^{n} x_i$.
3. Pour la taille de l'échantillon $n = 55$ et $\lambda_0 = 0.001$, tracer la densité de la statistique de test sous l'hypothèse nulle et indiquer la région du rejet au niveau $\alpha = 0.05$.
4. En utilisant théorème centrale limite, donner une approximation de la loi de T et répéter les questions (2.) et (3.) en utilisant cette loi.
5. En utilisant la loi exacte de T, tracer la fonction puissance pour $n = 10, 50, 100, 500, 100\,000$ et $\lambda \in (0, 3\lambda_0)$. Expliquer vos résultats.
6. Montrer que dans le cadre considéré (c'est-à-dire en utilisant $T(X) = \sum_{i=1}^{n} X_i$ comme statistique de test avec $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{E}(\lambda)\,, i = 1, ..., n$), pour tout $\lambda' < \lambda_0$, le test $H_0' : \lambda = \lambda'$ *vs.* $H_1 : \lambda > \lambda_0$ est de niveau $\alpha' < \alpha$.

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.