Principes et Méthodes Statistiques
TP 2023

Le travail sera conduit par groupes de 3 personnes. Le délivrable de ce TP est une archive contenant deux fichiers. Le premier sera le compte-rendu du TP au format Rmd, qui comprendra le code R et vos réponses détaillées aux questions selon les règles présentées ci-dessous. Le second sera le fichier pdf ou html issu directement de la compilation (knit) du fichier Rmd. L'archive devra être déposée sur Teide avant le vendredi 21 avril 2023 à 22h. Tout retard sera pénalisé.

Le compte-rendu Rmd comprendra, suivant la nature des questions posées, des calculs mathématiques et/ou des sorties numériques et graphiques de R. Une grande importance sera accordée aux commentaires, visant à interpréter les résultats et mettre en valeur votre analyse du problème. Des conseils et des directives obligatoires pour la rédaction du compte-rendu sont disponibles sur Chamilo ; les enseignants pourront y faire référence dans leur correction.

---

1 Assurance maritime

Le fichier `sinistres.csv` contient les montants payés par une compagnie d'assurance pour des sinistres d'une flotte de navires sur une période de 4 ans. Ces montants, intialement mesurés en milliers d'euros, ont été affectés d'un coefficient multiplicateur pour anonymiser les données.

1. Charger le tableau de données dans `R` en utilisant la commande `x<-scan("sinistres.csv")`. Les montants des sinistres sont notés $x_1, \ldots, x_n$. On supposera que ce sont des réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi.

   A l'aide d'histogrammes et de graphes de probabilités, montrer que les lois normale et exponentielle ne sont pas des modèles plausibles pour modéliser les montants des sinistres.

2. Tracer la fonction de répartition empirique pour ces données et calculer les principaux indicateurs statistiques. En déduire les caractéristiques essentielles de ces données. Quelle est la principale explication du fait que les modèles usuels de loi normale et exponentielle ne sont pas adaptés ?

On propose de modéliser les montants des sinistres par la loi $\mathcal{P}a(a,b)$, dont la densité est :

$$f(x) = \frac{a\,b^a}{x^{1+a}}\,\mathbb{1}_{[b,+\infty[}(x)$$

où a et b sont deux paramètres strictement positifs. Soit X une variable aléatoire de loi $\mathcal{P}a(a,b)$.

3. Calculer la fonction de répartition, l'espérance et la variance de X. Quelle condition doit vérifier a pour que cette loi admette une espérance et une variance finies ? On admettra que cette condition est vérifiée.

Soit $X_1, \ldots, X_n$ un échantillon de la loi $\mathcal{P}a(a,b)$.

4. Déterminer l'expression d'un graphe de probabilités pour cette loi et en déduire des estimateurs graphiques a_g et b_g de a et b.
5. Déterminer les estimateurs de a et b par la méthode des moments, $\tilde{a}_n$ et $\tilde{b}_n$.
6. Déterminer les estimateurs de a et b par la méthode du maximum de vraisemblance, $\hat{a}_n$ et $\hat{b}_n$.
7. La loi $\mathcal{P}a(a,b)$ vous parait-elle un modèle plausible pour les montants de sinistres ? Si oui, calculer toutes les estimations de a et b pour ces données. Le calcul théorique du biais et de la variance des estimateurs étant trop complexe, on déterminera quels sont les meilleurs estimateurs dans la partie 2, à l'aide de simulations.

2 Vérifications expérimentales à base de simulations

1. Donner la loi de probabilité de $Y = \ln \frac{X}{b}$. En déduire comment simuler un échantillon de taille n de la loi $\mathcal{P}a(a,b)$.

2. On suppose uniquement dans cette question que b est connu.

   (a) Pour un échantillon $x_1, \ldots, x_n$ de la loi $\mathcal{P}a(a,b)$, donner l'expression d'un intervalle de confiance de seuil α pour a.

   (b) Vérifier expérimentalement que, quand on simule un grand nombre m d'échantil-lons de taille n de la loi $\mathcal{P}a(a,2)$, alors une proportion approximativement égale à $1-\alpha$ des intervalles de confiance de seuil α obtenus contient la vraie valeur du paramètre a. Prendre plusieurs valeurs de α, m, n et a et commenter les résultats.

3. Comparer les différents estimateurs obtenus pour les paramètres a et b dans les questions 4, 5 et 6 de la première partie, en utilisant la méthodologie suivante.

   — Choisir des valeurs de a et b. Simuler m échantillons de taille n de la loi $\mathcal{P}a(a,b)$.

— Pour chaque échantillon, calculer les valeurs de toutes les estimations de a et b proposées. On obtient ainsi un échantillon de m valeurs pour chaque estimateur.

— Estimer le biais et l'erreur quadratique moyenne de ces estimateurs.

— Conclure : quels sont les meilleurs estimateurs de a et de b ?

4. Vérification de la convergence en probabilité d'un estimateur.
   Un estimateur T_n d'un paramètre θ est dit convergent en probabilité si et seulement si

$$\forall \varepsilon > 0, \lim_{n \to +\infty} P(|T_n - \theta| > \varepsilon) = 0.$$

   Choisir des valeurs de a et b. Vérifier la convergence en probabilité de $\tilde{a}_n$ et $\hat{a}_n$ en choisissant plusieurs valeurs adaptées de ε et de n, puis en simulant m échantillons de taille n de la loi $\mathcal{P}a(a, b)$ et enfin, en calculant le nombre de fois où l'écart en valeur absolue entre l'estimation et le vrai paramètre est supérieur à ϵ (vous pouvez tracer des courbes). Quel est l'estimateur qui converge le plus vite ?

5. Vérification de la normalité asymptotique d'un estimateur.
   Choisir des valeurs de a et b. Simuler m échantillons de taille n de la loi $\mathcal{P}a(a, b)$. Sur les échantillons des m estimations $\tilde{a}_n$ et $\hat{a}_n$, tracer un histogramme et un graphe de probabilités pour la loi normale. Faire varier n en partant de $n = 5$ et conclure.