Principes et Méthodes Statistiques
TP 2019

Le travail sera conduit par groupes de 2 ou 3 personnes, ces groupes étant constitués au hasard. 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 issu directement de la compilation du fichier Rmd. L'archive devra être déposée sur Teide avant le vendredi 15 mars 2019 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 Analyse des défauts de cuves

Dans un état idéal, des cuves ont une surface parfaitement lisse. En pratique, et après quelques années d'utilisation, elles présentent un certain nombre de défauts qui peuvent s'avérer dangereux pour leur utilisation. Un défaut est caractérisé par une fissure. La taille du défaut correspond à la profondeur de la fissure, en mm. A l'aide d'un appareil A, on détecte et on mesure les défauts de taille supérieure à 2 mm.

Le fichier `cuves.csv` contient les relevés de défauts de 3 cuves différentes, contrôlées après 5 années d'utilisation.

Vous pouvez soit créer les jeux de données manuellement, soit charger le tableau de données dans `R` en utilisant les commandes `read.table("cuves.csv",sep=";",header=T)` puis `attach(cuves)` et en enlevant les valeurs "NA" des vecteurs créés.

1. Effectuer une étude de statistique descriptive (histogrammes et indicateurs) pour les défauts des 3 cuves. Commenter les résultats.

   Une loi de probabilité classique pour modéliser les profondeurs de fissures est 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.

   On supposera ici que la profondeur des fissures est de loi $\mathcal{P}a(a, 2)$. Soit X une variable aléatoire de loi $\mathcal{P}a(a, 2)$.

2. 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 ?

3. Donner la loi de probabilité de $Y = \ln \frac{X}{2}$.

4. Donner l'expression d'un intervalle de confiance de seuil α pour a.

5. Mettre en œuvre toutes les méthodes statistiques que vous jugerez appropriées pour d'une part valider la pertinence de la loi $\mathcal{P}a(a, 2)$ pour ces données, et d'autre part estimer le mieux possible le paramètre a. Quand c'est possible, donner les propriétés théoriques des estimateurs de a proposés.

2 Vérifications expérimentales à base de simulations

1. Expliquer comment simuler un échantillon de taille n de la loi $\mathcal{P}a(a, b)$, pour b quelconque.

2. Vérifiez expérimentalement que, quand on simule un grand nombre m d'échantillons 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 m, n et a.

3. Comparer les différents estimateurs pour le paramètre a en utilisant la méthodologie suivante :
   — Simuler m échantillons de taille n de la loi $\mathcal{P}a(a, 2)$.
   — Pour chaque échantillon, calculer les valeurs de toutes les estimations de a proposées. On obtient ainsi un échantillon de m valeurs pour chaque estimateur.

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

— Conclure : quel estimateur pensez-vous être le meilleur ?

4. Vérification de la convergence faible d'un estimateur.
   Un estimateur A_n de a est dit faiblement convergent (ou convergent en probabilité) si et seulement si
   $$\forall \varepsilon > 0, \lim_{n \to +\infty} P(|A_n - a| > \varepsilon) = 0.$$
   Choisir l'un des estimateurs proposés (en précisant lequel) et une valeur de a. Vérifier la convergence faible de cet estimateur en choisissant plusieurs valeurs adaptées de ε et de n, puis simulant m échantillons de taille n de la loi $\mathcal{P}a(a, 2)$ et enfin, en calculant le nombre de fois où l'écart en valeur absolue entre l'estimation et le vrai paramètre est supérieure à ϵ (vous pouvez tracer des courbes). Conclure.

5. Vérification de la normalité asymptotique d'un estimateur.
   Choisir l'un des estimateurs proposés (en précisant lequel) et une valeur de a. Simuler m échantillons de taille n de la loi $\mathcal{P}a(a, 2)$. Sur l'échantillon des m estimations, tracer un histogramme et un graphe de probabilités pour la loi normale. Faire varier n en partant de $n = 5$ et conclure.

3 Comparaison de modèles et certification des cuves

1. Finalement, que proposez vous comme modèle approprié pour les tailles des défauts dans chacune des 3 cuves ?

2. Les défauts sont classés dangereux lorsque leur taille est supérieure à 5 mm. Le constructeur assure que ses cuves après 5 années d'utilisation ne présenteront pas une proportion de défauts dangereux supérieure à 5%.

   (a) Que pensez-vous de cette affirmation ?

   (b) Un autre appareil B, moins précis, permet uniquement de dire, à chaque défaut détecté, s'il est dangereux ou non. Si l'entreprise ne possédait que l'appareil de mesure B, qu'aurait-elle conclu sur l'affirmation du constructeur ?