Principes et Méthodes Statistiques
TP 2011

Un compte-rendu est à rendre pour le 25 novembre 2011 (date impérative). Il 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. Ce compte-rendu, sous forme d'un unique fichier pdf, sera déposé sur le site TEIDE. Tout retard sera pénalisé. Le travail sera conduit par groupes de 3 personnes, ces groupes étant tirés au hasard. Une soutenance aura lieu, selon les modalités décrites sur le Kiosk.

---

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 la commande `read.table("cuves.csv",sep=";",header=T)`, 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. Mettre en œuvre toutes les méthodes statistiques que vous jugerez appropriées pour d'une part valider la pertinence de la loi $\mathcal{P}a(a, b)$ pour ces données, et d'autre part estimer le mieux possible le paramètre a.

5. 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 ?

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. Donner l'expression d'un intervalle de confiance de seuil α pour a pour un échantillon de taille n de la loi $\mathcal{P}a(a, 2)$. 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. Proposer plusieurs méthodes pour estimer le paramètre a. 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. En déduire quel est le meilleur des estimateurs.

4. Soit une suite $\{X_n\}_{n\geq 1}$ de variables aléatoires réelles indépendantes et de même loi, d'espérance $E(X)$ finie.

   La loi faible des grands nombres dit que la suite $\{\bar{X}_n\}_{n\geq 1}$ converge en probabilité vers $E(X)$, ce qui s'écrit $\bar{X}_n \xrightarrow{P} E(X)$ et qui signifie que

   $$\forall\, \varepsilon > 0, \lim_{n\to+\infty} P\left(|\bar{X}_n - E(X)| > \varepsilon\right) = 0.$$

   Concrètement, cela signifie que quand on fait un très grand nombre d'expériences identiques et indépendantes, la moyenne des réalisations de la variable aléatoire à laquelle on s'intéresse est une bonne approximation de l'espérance de sa loi. La procédure suivante a pour but de vérifier ce fait expérimentalement.

   Simuler m échantillons de taille n de la loi $\mathcal{P}a(a, 2)$. Calculer le nombre de fois où l'écart en valeur absolue entre la moyenne empirique et l'espérance de la loi simulée est supérieure à un ϵ à choisir. Faire varier n en partant de $n = 5$ et conclure.

5. Soit une suite $\{X_n\}_{n\geq 1}$ de variables aléatoires réelles indépendantes et de même loi, d'espérance $E(X)$ et d'écart-type $\sigma(X) = \sqrt{Var(X)}$ finis.

   Pour tout $n \geq 1$, on pose :

   $$Z_n = \frac{\sum_{i=1}^{n} X_i - nE(X)}{\sqrt{nVar(X)}} = \sqrt{n}\,\frac{\bar{X}_n - E(X)}{\sigma(X)}$$

   Le théorème central-limite prouve que la suite $\{Z_n\}_{n\geq 1}$ converge en loi vers la loi normale centrée-réduite, ce qui signifie que la fonction de répartition de Z_n tend vers la fonction de répartition ϕ de la loi normale centrée-réduite quand n tend vers l'infini. Ce résultat s'écrit :

   $$\sqrt{n}\,\frac{\bar{X}_n - E(X)}{\sigma(X)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

   Concrètement, cela signifie que, pour n grand, $\bar{X}_n$ se comporte approximativement comme une variable aléatoire de loi normale $\mathcal{N}\left(E(X), \frac{Var(X)}{n}\right)$. La procédure suivante a pour but de vérifier ce fait expérimentalement.

   Simuler m échantillons de taille n de la loi $\mathcal{P}a(a, 2)$. Sur l'échantillon des m moyennes empiriques, tracer un histogramme et un graphe de probabilités pour la loi normale. Faire varier n en partant de $n = 5$ et conclure.