MST 2

Projet Logiciel R

$\sim$ o $\sim$ o $\sim$

— *Le projet donnera lieu à la rédaction d'un court rapport présentant les principaux résultats obtenus à l'aide du logiciel R ainsi que les commentaires essentiels pour l'interprétation des résultats.*

— *Il est impératif d'insérer dans votre rapport les résultats produits par le logiciel ainsi que les commandes ou programmes écrits en R.*

— *Utiliser la police* `courrier` *pour présenter les sorties du logiciel R (elle permet une meilleure mise en forme des résultats.)*

$\sim$ o $\sim$ o $\sim$

On s'intéresse à la population des 496 communes de moins de 5000 habitants du Bas-Rhin. Le fichier `com67.txt` fournit le nombre d'habitants X de chacune de ces communes en 2014.

Questions

1. Importer le jeu de données dans le logiciel R puis fournir les indicateurs statistiques suivants pour la variable `pop` (nombre d'habitants) : min, max, moyenne, médiane, variance et écart type.

On suppose dans la suite que les paramètres de la population sont inconnus et que l'on souhaite estimer ceux-ci à partir d'un échantillon.

2. Extraire aléatoirement et sans remise $n = 30$ communes parmi les $N = 496$ puis construire l'échantillon des nombres d'habitants des 30 communes obtenues.
3. À partir de l'échantillon construit à la question précédente, calculer l'intervalle de confiance au seuil de 95% pour μ, le nombre moyen d'habitants d'une commune du Bas-Rhin de moins de 5000 habitants. On précisera la formule générale utilisée ainsi que la valeur de chaque terme intervenant dans cette formule.
4. L'intervalle de confiance obtenu contient-il la vraie valeur du nombre moyen d'habitants d'une commune ? Était-ce attendu ? En répétant la construction d'un intervalle de confiance un grand nombre de fois, à quel résultat doit-on s'attendre ?
5. En utilisant le langage de commande de R, écrire un programme permettant de calculer, pour un nombre K d'échantillons aléatoires de taille n, la proportion d'intervalles de confiance obtenus encadrant bien la vraie valeur de μ.

   Présenter très précisément le programme en R ainsi que les résultats numériques obtenus pour différentes valeurs de K et de n. Pour chacune des trois valeurs de taille d'échantillon $n = 10, 30$ et 200, fournir l'histogramme des K moyennes d'échantillons obtenues (avec par exemple $K = 100000$).

6. Faire de même (programme R + résultats numériques) pour l'intervalle de confiance de la variance σ^2 du nombre d'habitants d'une commune.

On s'intéresse maintenant à la forme de la distribution de la variable X.

7. En vous aidant d'une représentation graphique appropriée, commenter l'allure de la distribution de la variable `pop`.
8. Peut-on admettre l'hypothèse selon laquelle le nombre d'habitants d'une commune X est distribué selon une loi de probabilité log-normale ?

   – On mettra en œuvre le test d'hypothèse d'adéquation de Kolmogorov-Smirnov et on interprétera précisément les résultats obtenus.

   – Sur le graphique de la question 7., superposer une loi log-normale appropriée.