MATH0487-2 - Éléments de statistiques

Projet

Généralités

Le projet porte sur l'étude statistique des relations entre certaines caractéristiques physiques et sociales de patients et les frais hospitaliers qu'ils doivent débourser. Les étudiants doivent d'une part utiliser l'analyse descriptive pour décrire les données et étudier des échantillons i.i.d. tirés à partir des données et d'autre part utiliser les statistiques inférentielles pour étudier différents estimateurs et réaliser des tests d'hypothèses.

Ce travail est à réaliser par groupe de deux. Chaque groupe doit rendre une archive nommée `projet_stats.zip` contenant, sans sous-dossier, un rapport au format `PDF`, ainsi que tous les codes sources MATLAB produits. Chaque résultat du rapport doit être appuyé par un code. De plus, pour chaque occurrence du symbole 💻, le code d'une fonction est à compléter. Les rapports inutilement longs sont à proscrire. La longueur conseillée du rapport est de 15 faces, hors annexes et page de garde. Il n'est ni nécessaire d'écrire une introduction, ni de faire des rappels des questions posées, ni de prévoir une table des matières. Toute sous-question posée dans l'énoncé devra comporter un élément de réponse dans le rapport, *en justifiant votre raisonnement.* **Toute forme de plagiat sera sanctionnée.**

Le projet est à rendre pour le lundi 09/12/2019 23h59 via la plateforme

`https://submit.montefiore.ulg.ac.be/`

Notez que tous les membres du groupe doivent s'inscrire sur la plateforme et rejoindre le groupe.

Au-delà de la deadline il ne sera plus possible de soumettre les projets.

Présentation du problème

Vous disposez d'un jeu de données `data.csv` [1] et d'une fonction `population.m` (à laisser intacte) vous permettant de générer *votre* population personnalisée. Pour ce faire, appelez la fonction avec en arguments le jeu de données et votre numéro de groupe [2]. Vous obtiendrez en sortie une `table` reprenant pour chaque patient : un nom factice (`LastName`), son age (`Age`), sexe (`Sex` avec 0 pour homme), BMI (`BMI` en $\mathrm{kg\,m^{-2}}$), nombre d'enfants (`Children`), s'il fume ou pas (`Smoker`) et ses frais hospitaliers (`Charges` en \$).

Les objectifs de ce projet sont les suivants : extraire différentes statistiques descriptives, apprendre à extraire un sous-ensemble aléatoire d'observations de manière répétitive et à comparer les statistiques à celles obtenues sur les données complètes, tirer plusieurs échantillons i.i.d. pour estimer différents paramètres et réaliser des tests d'hypothèses.

1. Source : `https://github.com/stedy/Machine-Learning-with-R-datasets`.
2. Votre numéro de groupe correspond à l'identifiant que vous recevrez lorsque vous créerez le groupe sur la plateforme de soumission.

Questions

1. **Analyse descriptive**

   Dans cette partie, vous vous intéresserez aux frais hospitaliers déboursés par les patients de *votre* population.

   (a) Générez l'histogramme des frais hospitaliers. Interprétez.

   (b) Calculez la moyenne, la médiane et l'écart-type de ces frais. Interprétez. Comparez avec les résultats obtenus pour Ms. Smith. [`sum_up.m`]

   (c) Définissez les caractéristiques des frais hospitaliers *normaux* (au sens de la loi normale) et calculez la proportion de patients ayant des frais *normaux* (au sens de la loi normale). Ms. Smith a-t-elle des frais *normaux* (au sens de la loi normale) ? [`normal_interval.m`]

   (d) Réalisez la boîte à moustaches relative aux frais hospitaliers. Y a-t-il des données aberrantes ? Que valent les quartiles ? [`quartiles.m`]

   (e) Réalisez le polygone des fréquences cumulées des frais hospitaliers et estimez la proportion de patients ayant des frais inférieurs ou égaux à 25 000 $ et supérieurs à ceux de Ms. Smith. [`proportion.m`]

   (f) Réalisez deux scatterplots comparant l'âge des patients, respectivement masculins et féminins, et leurs frais hospitaliers. Calculez les coefficients de corrélation. Interprétez ces résultats. [`correlation.m`]

2. **Génération d'échantillons i.i.d.**

   Dans cette partie du travail, vous devrez tirer un ou plusieurs échantillons i.i.d. de patients à partir de *votre* population et comparez différentes statistiques descriptives de ces échantillons avec celle-ci.

   (a) Tirez un échantillon i.i.d. de 50 patients. [`iid_sample.m`]

      i. Calculez la moyenne, la médiane et l'écart-type des frais hospitaliers. Comparez aux résultats de la population. [`sample_sum_up.m`]

      ii. Réalisez la boîte à moustaches relative aux frais hospitaliers. Comparez à la population.

      iii. Réalisez le polygone des fréquences cumulées des frais hospitaliers. Comparez à la population. Calculez la distance de Kolmogorov Smirnov (c.-à-d. la distance maximale entre le polygone des fréquences cumulées) entre l'échantillon et la population. [`ks_distance.m`]

   (b) Tirez 500 échantillons i.i.d. de 50 patients.

      i. Calculez pour chaque échantillon les frais hospitaliers moyens et sauvegardez les 500 moyennes dans une nouvelle variable. Générez l'histogramme de cette nouvelle variable. L'allure de l'histogramme vous fait-elle penser à une loi théorique connue ? Que vaut la moyenne de cette nouvelle variable ? Cette moyenne est-elle proche de celle obtenue pour la population ?

      ii. Calculez pour chaque échantillon la médiane des frais hospitaliers et sauvegardez les 500 médianes dans une nouvelle variable. Générez l'histogramme de cette nouvelle variable. L'allure de l'histogramme vous fait-elle penser à une loi théorique connue ? Que vaut la moyenne de cette nouvelle variable ? Est-elle plus proche de celle de la population que la valeur calculée à la fin du point précédent ?

iii. Calculez pour chaque échantillon l'écart-type des frais hospitaliers et sauvegardez les 500 écart-types dans une nouvelle variable. Générez l'histogramme de cette nouvelle variable. L'allure de l'histogramme vous fait-elle penser à une loi théorique connue ? Que vaut la moyenne de cette nouvelle variable ? Cette moyenne est-elle proche de l'écart-type des frais hospitaliers de la population ? Interprétez.

iv. Concernant les frais hospitaliers, calculez pour chaque échantillon la distance de Kolmogorov-Smirnov entre les polygones des fréquences cumulées de la population et de l'échantillon considéré [3]. Sauvegardez les 500 distances obtenues dans une nouvelle variable. Réalisez l'histogramme de cette variable. Interprétez.

3. **Estimation**

Dans cette partie, vous vous intéresserez plus particulièrement au BMI (*Body Mass Index* ou *Indice de Masse Corporelle* en français) des patients.

Tirez 100 échantillons i.i.d. de 50 patients.

(a) Calculez pour chaque échantillon la moyenne m_X et sauvegardez les 100 valeurs dans une nouvelle variable. Utilisez cette nouvelle variable pour estimer le biais et la variance de l'estimateur m_X du BMI moyen de la population.

(b) Calculez pour chaque échantillon la médiane $median_X$ et sauvegardez les 100 valeurs dans une nouvelle variable. Utilisez cette nouvelle variable pour estimer le biais et la variance de l'estimateur $median_X$ du BMI moyen de la population.

(c) Répétez les deux points précédents avec des échantillons i.i.d. de taille 100. Que constatez-vous ? Interprétez.

(d) Construisez, pour 100 échantillons i.i.d. de taille 20, un intervalle de confiance à 95% du BMI de la population à partir de m_X en faisant l'hypothèse que la variable parente est Gaussienne et

i. en utilisant la loi de Student [`student_interval.m`]

ii. en utilisant la loi de Gauss [`gauss_interval.m`]

pour construire l'intervalle. Vérifiez dans les deux cas quelle proportion des 100 intervalles de confiance contient la valeur de la population. Interprétez. Était-il raisonnable de supposer que la variable parente était Gaussienne ?

4. **Tests d'hypothèse**

Une compagnie d'assurance souhaite étudier l'influence de la cigarette sur les frais médicaux afin d'ajuster ses tarifs. La compagnie engage donc un institut de recherche afin de tester l'hypothèse $H_0 =$ « Les frais hospitaliers des fumeurs sont en moyenne supérieurs de x aux frais hospitaliers des non-fumeurs. » où x correspond à la différence des deux moyennes dans *votre* population. L'institut réalise un sondage parmi un échantillon i.i.d. de 50 patients et utilise un seuil de signification $\alpha = 5\%$.

Tirez 100 fois un échantillon i.i.d. de 50 patients.

(a) Effectuez le test d'hypothèse demandé. Dans combien de cas l'hypothèse est-elle rejetée ? Comparez cette valeur à α. Interprétez.

(b) La compagnie souhaite maintenant cibler une clientèle de plus de 50 ans. Reeffectuez le test sur ce sous-ensemble *votre* population (en gardant le même x). Dans combien de cas l'hypothèse est-elle rejetée ? Comparez avec la valeur précédente. Interprétez.

3. On ne demande pas de générer les polygones des fréquences cumulées explicitement.

Suggestions

Après avoir généré votre population, il est conseillé de l'enregistrer afin de ne pas devoir la recalculer à chaque exécution. Par exemple, pour le groupe 13 :

```
pop = population(readtable('data.csv'), 13);
writetable(pop, 'population.csv');
```

Les fonctions suivantes de MATLAB peuvent vous être utiles : `abs`, `boxplot`, `cell`, `cdfplot`, `corrcoef`, `cumsum`, `hist`, `hold`, `icdf`, `kstest2`, `max`, `mean`, `median`, `min`, `mode`, `quantile`, `readtable`, `randsample`, `scatter`, `std`, `subplot`, `table`, `writetable` et `help`.

Il n'est pas nécessaire de réaliser un script différent pour chaque sous-question.