MATH0487-2 - Éléments de statistiques

Projet

Généralités

Le projet porte sur l'étude statistique de la consommation d'alcool [1] par habitant sur une année de chaque pays du monde. Les étudiants devront 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 devra être réalisé par groupe de deux. Chaque groupe devra rendre une archive `.zip` contenant un rapport au format pdf, le fichier Excel *resultats.xlsx* résumant les résultats numériques obtenus ainsi que ses codes sources MATLAB. Les rapports inutilement longs sont à proscrire. La longueur conseillée du rapport est de 15 pages, 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*. Vous devez rendre un code source MATLAB pour toutes les sous-questions et le mettre en annexe du rapport. **Toute forme de plagiat sera sanctionnée.**

Le projet est à rendre pour le lundi 04/12/2017 23 :59 via la plateforme

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

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

Présentation du problème

Vous disposez d'un fichier .csv reprenant la consommation de bière (en nombre de canettes), de vin (en nombre de verres), d'alcool fort (en nombre de shots) et la consommation totale d'alcool pur (en litres) par habitant de 100 pays en 2010 [2]. 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. L'abus d'alcool est dangereux pour la santé, consommez avec modération.
2. Source : World Health Organisation, Global Information System on Alcohol and Health (GISAH), 2010

Questions

1. **Analyse descriptive**
   (a) Générez les histogrammes de la consommation de bière et d'alcool fort dans le monde. Interprétez et comparez.
   (b) Calculez la moyenne, la médiane, le mode et l'écart-type de la consommation de bière et d'alcool fort dans le monde. Interprétez. Comparez avec la consommation belge.
   (c) Définissez les caractéristiques d'une consommation "normale" (au sens de la loi normale) pour la bière et l'alcool fort et calculez la proportion de pays ayant une consommation de bière "normale" d'une part et d'alcool fort "normale" d'autre part (au sens de la loi normale). La Belgiqe a-t-elle une consommation de bière et d'alcool fort normale (au sens de la loi normale) ?
   (d) Réalisez les boîtes à moustaches relatives à la consommation de bière et d'alcool fort. Y a-t-il des données aberrantes ? Que valent les quartiles ?
   (e) Réalisez le polygone des fréquences cumulées de la consommation de bière et estimez la proportion de pays ayant une consommation à la fois supérieure à 200 canettes et strictement inférieure à celle de la Belgique.
   (f) Réalisez trois scatterplots comparant à chaque fois la consommation d'alcool pur à la consommation de bière, de vin et d'alcool fort. Calculez dans chaque cas le coefficient de corrélation. Interprétez ces résultats.
2. **Génération d'échantillons i.i.d.**
   Dans cette partie du travail, nous considérons que la base de données reçue représente la population. Nous tirons un ou plusieurs échantillons i.i.d. de pays à partir de cette population et comparons différentes statistiques descriptives de ces échantillons avec la population.
   (a) Tirez un échantillon i.i.d. de 20 pays.
      i. Calculez la moyenne, la médiane et l'écart-type de la consommation de bière et d'alcool fort de l'échantillon. Comparez aux résultats de la population.
      ii. Réalisez les boîtes à moustaches relatives à la consommation de bière et d'alcool fort. Comparez à la population.
      iii. Réalisez le polygone des fréquences cumulées de la consommation de bière et d'alcool fort. Comparez à la population. Calculez la distance de Kolmogorov Smirnov dans les deux cas (i.e. la distance maximale entre le polygone des fréquences cumulées relatif à l'échantillon et à celui relatif à la population).
   (b) Tirez 100 échantillons i.i.d. de 20 pays.
      i. Calculez pour chaque échantillon la consommation moyenne de bière et d'alcool fort et sauvegardez pour chaque boisson les 100 moyennes dans une nouvelle variable. Générez les deux histogrammes de ces nouvelles variables. L'allure des histogrammes vous fait-elle penser à une loi théorique connue ?

Que vaut la moyenne de chaque nouvelle variable ? Ces moyennes sont-elles proches respectivement de la consommation moyenne de bière et d'alcool fort obtenues par la population ?

ii. Calculez pour chaque échantillon la médiane de la consommation de bière et d'alcool fort et sauvegardez pour chaque boisson les 100 médianes dans une nouvelle variable. Générez les deux histogrammes de ces nouvelles variables. L'allure des histogrammes vous fait-elle penser à une loi théorique connue ? Que vaut la moyenne de chaque nouvelle variable ? Sont-elles plus proches respectivement de la consommation moyenne de bière et d'alcool fort de la population que les valeurs calculées à la fin du point précédent ?

iii. Calculez pour chaque échantillon l'écart-type de la consommation de bière et d'alcool fort et sauvegardez pour chaque boisson les 100 écart-types dans une nouvelle variable. Générez les deux histogrammes de ces nouvelles variables. L'allure des histogrammes vous fait-elle penser à une loi théorique connue ? Que vaut la moyenne de chaque nouvelle variable ? Ces moyennes sont-elles proches de l'écart-type de la consommation de bière et d'alcool fort de la population ? Interprétez.

iv. Concernant la consommation de bière et d'alcool fort, calculez pour chaque échantillon et pour chaque boisson la distance de Kolmogorov Smirnov entre les polygones des fréquences cumulées de la population et de l'échantillon considéré [3]. Sauvegardez dans les deux cas les 100 distances obtenues dans une nouvelle variable. Réalisez l'histogramme de ces deux variables.

v. Répétez la procédure décrite au point iv. pour la consommation de vin et d'alcool pur. Comparez l'allure des quatre histogrammes obtenus. Interprétez votre comparaison sur base des résultats théoriques présentés en cours.

3. **Estimation**

Tirez 100 échantillons i.i.d. de 20 pays et considérez ici uniquement leur consommation de vin.

(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 de la consommation moyenne de vin 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 $median_X$ de la consommation moyenne de vin de la population.

(c) Répétez les deux points précédents avec des échantillons i.i.d. de taille 50. Que constatez-vous ? Interprétez.

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

(d) Construisez pour chaque échantillon de taille 20 un intervalle de confiance à 95% de la consommation de vin 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 pour construire l'intervalle.

ii. en utilisant la loi de Gauss pour construire l'intervalle.

Vérifiez dans les deux cas combien des 100 intervalles de confiance contiennent la valeur de la population. Interprétez. Etait-il raisonable de supposer que la variable parente était Gaussienne ?

4. **Tests d'hypothèse**

Les Belges sont fiers de leurs nombreuses bières. Dans cette partie du travail, nous imaginons que l'OMS (Organisation Mondiale de la Santé) se penche sur la consommation de bière par belge afin de vérifier si ces derniers font effectivement partie des plus gros buveurs de bière de la planète. L'OMS décide donc de demander à cinq instituts de statistique indépendants ainsi qu'à l'Etat belge de tester l'hypothèse H_0 = "la proportion de pays consommant plus de bière que la Belgique est de x%" versus l'hypothèse alternative H_1 = "la proportion de pays consommant plus de bière que la Belgique est supérieure à x%" (x correspond à la vraie proportion de pays ayant une consommation de bière supérieure à celle de la Belgique [4]). Tous les instituts ainsi que l'Etat belge tirent un échantillon i.i.d. de 50 pays et utilisent le même seuil de signification $\alpha = 5\%$. Si au moins un des instituts rejette l'hypothèse H_0, il sera alors considéré par l'OMS que les Belges ne font pas partie des plus gros consommateurs de bière du monde.

Tirez 100 fois 6 échantillons i.i.d. de 50 pays [5].

(a) Effectuez dans chaque cas le test d'hypothèse demandé [6]. Dans combien de cas l'Etat belge a-t-il rejeté l'hypothèse ? Comparez cette valeur à α.

(b) Dans combien de cas l'OMS a-t-elle considéré que les Belges ne font pas partie des plus gros consommateurs de bière du monde ? Comparez cette valeur à celle de la question précédente. Interprétez.

(c) Quelle(s) méthode(s) aurait-on pu utiliser pour éviter que les instituts de statistique indépendants soient avantagés par rapport à l'Etat belge [7] ?

Suggestions

Les fonctions suivantes de Matlab peuvent vous être utiles : abs, boxplot, cdfplot, corrcoef, cumsum, findobj, get, help, hist, hold, interpl, kstest2, max, mean, median, min, mode, quantile, randsample, scatter, std, subplot, tableread, table2array.

4. Vous obtenez la valeur de x en considérant, comme dans le reste du projet, que votre base de données correspond à la population.
5. Le premier des six correspondant à chaque fois à l'Etat belge.
6. Vous pouvez utiliser l'approximation vue au cours théorique.
7. En ce sens qu'à eux cinq ils avaient plus de chance de tomber sur un échantillon rejetant H_0