MATH0487-2 - Éléments de statistiques

Projet

Généralités

Le projet porte sur l'étude statistique du taux de natalité et mortalité dans les différents 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 11/12/2018 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 le nombre de naissance (par 1000 habitants) et le nombre de décès (par 1000 habitants) en 2013 [1]. 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 : World Health Organisation, Global Health Observatory data repository, 2013.

Questions

1. **Analyse descriptive**
   (a) Générez les histogrammes du taux de natalité et de mortalité dans le monde. Interprétez et comparez.
   (b) Calculez la moyenne, la médiane, le mode et l'écart-type du taux de natalité et de mortalité dans le monde. Interprétez. Comparez avec les résultats obtenu pour la Belgique.
   (c) Définissez les caractéristiques d'un taux "normal" (au sens de la loi normale) du nombre de naissance et de décès et calculez la proportion de pays ayant un taux de natalité "normal" d'une part et de mortalité "normal" d'autre part (au sens de la loi normale). La Belgique a-t-elle un taux de natalité et de mortalité normal (au sens de la loi normale) ?
   (d) Réalisez les boîtes à moustaches relatives au taux de natalité et de mortalité. Y a-t-il des données aberrantes ? Que valent les quartiles ?
   (e) Réalisez le polygone des fréquences cumulées du taux de natalité et estimez la proportion de pays ayant un taux de natalité inférieur ou égal à 20 pour 1000 habitants et supérieur à celui de la Belgique.
   (f) Réalisez un scatterplot comparant le taux de natalité et le taux de mortalité. Calculez 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 du taux de natalité et de mortalité de l'échantillon. Comparez aux résultats de la population.
      ii. Réalisez les boîtes à moustaches relatives au taux de natalité et de mortalité. Comparez à la population.
      iii. Réalisez le polygone des fréquences cumulées du taux de natalité et de mortalité. 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 celui relatif à la population).
   (b) Tirez 500 échantillons i.i.d. de 20 pays.
      i. Calculez pour chaque échantillon le taux moyen de natalité et de mortalité et sauvegardez pour chaque statistique les 500 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 du taux moyen de natalité et de mortalité obtenues par la population ?

ii. Calculez pour chaque échantillon la médiane du taux de natalité et de mortalité et sauvegardez pour chaque statistique les 500 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 du taux moyen de natalité et de mortalité de la population que les valeurs calculées à la fin du point précédent ?

iii. Calculez pour chaque échantillon l'écart-type du taux de natalité et de mortalité et sauvegardez pour chaque boisson les 500 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 du taux de natalité et de mortalité de la population ? Interprétez.

iv. Concernant le taux de natalité et de mortalité, calculez pour chaque échantillon et pour chaque statistique la distance de Kolmogorov Smirnov entre les polygones des fréquences cumulées de la population et de l'échantillon considéré [2]. Sauvegardez dans les deux cas les 500 distances obtenues dans une nouvelle variable. Réalisez l'histogramme de ces deux variables.

3. **Estimation**

Tirez 100 échantillons i.i.d. de 20 pays et considérez ici uniquement leur taux de natalité.

(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 taux de natalité 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 taux de natalité moyen de la population.

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

(d) Construisez, pour chaque échantillon de taille 20, un intervalle de confiance à 95% du taux de natalité 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.

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

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 - proportion**
   Les Belges sont connus pour avoir un faible taux de natalité. Dans cette partie du travail, nous imaginons que l'OMS (Organisation Mondiale de la Santé) se penche sur le taux de natalité en Belgique afin de vérifier si ces allégations sont vraies. L'OMS décide donc de demander à quatre instituts de statistique indépendants ainsi qu'à l'Etat belge de tester l'hypothèse H_0 = "la proportion de pays ayant un taux de natalité plus faible que la Belgique est de $x\%$" versus l'hypothèse alternative H_1 = "la proportion de pays ayant un taux de natalité plus faible que la Belgique est supérieure à $x\%$" (x correspond à la vraie proportion de pays ayant un taux de natalité inférieur à celui de la Belgique [3]). Tous les instituts ainsi que l'Etat belge tirent un échantillon i.i.d. de 40 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 n'ont pas un faible taux de natalité.
   Tirez 100 fois 5 échantillons i.i.d. de 40 pays [4].

   (a) Effectuez dans chaque cas le test d'hypothèse demandé. 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 n'ont pas un faible taux de natalité ? 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 [5] ?

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.

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