MATH0487-2 - Éléments de statistique
Partie 1 du projet personnel

Généralités

Le projet portera sur l'étude des résultats de première session du cours d'*éléments de probabilité* lors de l'année académique 2015-2016 [1] [2]. Ce projet se déroule en deux phases. Lors de la première partie, dont fait l'objet cet énoncé, les étudiants devront d'une part utiliser l'analyse descriptive pour décrire les données et d'autre part étudier des échantillons i.i.d. tirés à partir des données. Lors de la deuxième partie de ce projet, pour laquelle un nouvel énoncé sera généré et présenté en novembre, les étudiants devront utiliser les statistiques inférentielles.

Ce travail devra être réalisé individuellement. Chaque étudiant devra rendre une archive `.zip` contenant d'une part un rapport au format pdf et d'autre part ses codes sources MATLAB. Les rapports inutilement longs sont a proscrire. 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.

Cette partie du projet doit être rendue pour le jeudi 03/11/2016 23 :59 via la plateforme

`submit.run.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 Excel reprenant les résultats anonymisés des étudiants ayant participé à l'examen de première session du cours d'éléments de probabilité du professeur Wehenkel lors de l'année académique 2015-2016. L'objectif premier de ce projet est d'extraire différentes statistiques descriptives. Le second objectif est d'apprendre à extraire un sous-ensemble aléatoire d'observations de manière répétitive et à comparer les statistiques à celles obtenues sur les données complètes.

1. on ne considère que les étudiants ayant participé à toutes les parties de l'examen écrit
2. les points ont été arrondis de manière à n'avoir que des nombres entiers

Questions

1. **Analyse descriptive**
   (a) Générez les trois histogrammes des résultats des questions de théorie. Comparez et interprétez.
   (b) Calculez les trois moyennes, les trois médianes, les trois modes et les trois écart types des résultats des exercices. Comparez et interprétez. Définir pour chaque exercice quels résultats sont "normaux" et la proportion d'étudiants ayant réalisé un résultat "normal" (au sens de la loi normale).
   (c) Réalisez les trois boîtes à moustaches relatives aux résultats des projets. Y a-t-il des données aberrantes ? Que valent les quartiles ?
   (d) Générez deux nouvelles variables : la moyenne de chaque étudiant pour la théorie d'une part et pour les exercices d'autre part. Réalisez les polygones des fréquences cumulées de ces deux nouvelles variables et estimez dans les deux cas la proportion d'étudiants ayant une cote comprise dans l'intervalle $[10, 14]$.
   (e) Réalisez un scatterplot comparant les résultats obtenus au rapport du projet 2 et les résultats obtenus lors de la question sur le projet 2. Calculez le coefficient de corrélation. Interprétez ce résulat.
2. **Génération d'échantillons i.i.d.**
   Dans cette partie du travail, nous considérons que les résultats des données complètes représentent la population. Nous tirons un ou plusieurs échantillons i.i.d. d'étudiants à partir de cette population et comparons différentes statistiques descriptives de ce(s) échantillon(s) avec la population.
   (a) Tirer un échantillon i.i.d. de 20 étudiants.
      i. Calculez les trois moyennes, les trois médianes et les trois écart types des résultats des exercices. Comparez aux résultats de la population.
      ii. Réalisez les trois boîtes à moustaches relatives aux résultats des projets. Comparez à la population.
      iii. Générez une nouvelle variable : la moyenne de chaque étudiant pour la théorie. Réalisez le polygone des fréquences cumulées. Comparez à la population. Calculez la distance de Kolmogorov Smirnov entre ces deux courbes (la distance maximale entre les deux courbes).
   (b) Tirez 100 échantillons i.i.d. de 20 étudiants.
      i. Calculez pour chaque échantillon la moyenne obtenue à l'exercice 1 et sauvegardez les 100 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 la nouvelle variable ? Est-elle proche de la moyenne obtenue par la population à l'exercice 1 ?

ii. Calculez pour chaque échantillon la médiane obtenue à l'exercice 1 et sauvegardez les 100 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 la nouvelle variable ? Est-elle plus proche de la moyenne obtenue par la population à l'exercice 1 que la valeur calculée à la fin du point précédent ?

iii. Calculez pour chaque échantillon l'écart-type obtenu à l'exercice 1 et sauvegardez les 100 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 la nouvelle variable ? Est-elle proche de l'écart-type obtenu par la population à l'exercice 1 ? Interprétez.

iv. Concernant l'exercice 1, 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 100 distances obtenues dans une nouvelle variable. Réalisez l'histogramme de cette variable.

v. Répétez la procédure décrite au point iv. pour les exercices 2 et 3. Comparez l'allure des trois histogrammes obtenus. Interprétez votre comparaison sur base des résultats théoriques présentés en cours.

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.

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