Fiche TP 02 :
Analyse de données

Master 1 informatique
2017-2018

Exercice 1 : Description de données

Cet exercice a pour but d'introduire aux outils de statistiques descriptives d'une seule variable (moyenne, quartiles, histogramme, etc.). Les étapes proposées dans cet exercice sont nécessaires pour commencer à décrire un ensemble de données impliquant une seule variable.

1 fichiers est disponible : `frame_features_qap.dat`. Chaque fichier contient les mesures de caractéristiques . Les observations peuvent donc être considérées comme indépendantes.

Questions :

a - Pour chaque série de mesures, pour certaines instances de problèmes, décrire qualitativement la distribution des mesures à l'aide de la fonction `hist`.

b - Pour chaque série de mesures, pour certaines instances de problèmes, calculer le nombre de données, la moyenne, la variance, l'écart-type et les quartiles.

c - Comparer les distributions de mesures à l'aide de boîte à moustache pour différentes instances de problème. Commenter.

Exercice 2 : Influence de la taille d'un échantillon

Cet exercice a pour but de montrer comment l'estimation de la moyenne dépend de la taille de l'échantillon (nombre d'observations) sur lequel repose l'estimation.

Dans le fichier `temp_1.csv`, nous disposons de 1000 mesures indépendantes de température. Nous allons extraire des sous-échantillons (aléatoires) de différentes tailles et montrer que la dispersion des moyennes observées dépend de cette taille. Cet exercice est une illustration de la loi faible des grands nombres.

Questions :

a - Lire les données du fichier `temp_1.csv` et les enregistrer dans le vecteur `temperatures`.

b - Ecrire une fonction qui calcule la moyenne d'un sous-échantillon aléatoire de taille n d'un vecteur donné.

c - Ecrire une fonction qui calcule un vecteur de taille *size* dont les éléments sont des estimations de la moyenne basées sur un échantillon de taille n.

d - Comparer les distributions des moyennes estimées pour des échantillons de taille 10 et de taille 100

e - Calculer un vecteur s dont les éléments sont les écart-types des distributions des moyennes estimées pour des échantillons de taille $\{5, 10, 15, \ldots, 500\}$. Représenter graphiquement ce vecteur : `plot`(x, y) permet de dessiner un vecteur de points où x et y sont des vecteurs contenant respectivement les abscisses et les ordonnées des points.

f - Calculer l'intervalle de confiance à 95% des moyennes estimées pour des échantillons de taille 10, 30 et 1000.

Exercice 3 : Comparaison de moyenne

Cet exercice a pour but d'illustrer les outils de comparaison de moyennes à l'aide de tests statistiques.

Les données utilisées sont celles des mesures de performances contenues dans les fichiers `perf_ea_qap.dat`, `perf_ils_qap.dat`, `perf_ts_qap.dat` et `perf_sa_qap.dat`.

Questions :

a - Les mesures de performance suivent-elles une distribution normale ? Quel(s) test(s) peut-on utiliser pour comparer les moyennes de ces distributions ?

b - Comparer les moyennes estimées sur les 100 observations des 4 séries de données à l'aide d'un test statistique adéquat.

c - Comparer de nouveau, à l'aide d'un test statistique, les moyennes des 4 séries de données à partir de sous-échantillons aléatoires de taille 30.

Exercice 4 : Corrélation

A l'aide de recherche sur le web, répondez aux questions suivantes :

a - Quelles sont les mesures possibles de corrélation ?

b - Existe-t-il une différence entre corrélation et causalité ?

Exercice 5 : Analyse

Le but de cet exercice est d'introduire aux outils de statistique descriptive à deux variables en analysant statistiquement les résultats de simulations d'une procession.

a - Analyser les corrélations entre les différentes caractéristiques d'instances de problèmes.

b - Existe-t-il une différence entre corrélation et causalité ?