Analyse de données - TP 2

ISEP – 6 Octobre 2020

Instructions : Préparez un rapport incluant le code source et vos résultats, et déposez-le sur Moodle. Pas plus de 2 personnes par groupe. N'oubliez pas de mettre les 2 noms sur le rapport, ou de faire 2 rendus.

Bibliothèques

Ce TP nécessite les bibliothèques suivantes : Numpy, Matplotlib, Seaborn, et Scipy :

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns; sns.set()
```

A Analyse mutlivariée : Iris de Fisher

Dans cet exercice, on va utiliser le jeu de données Iris.

1. Ouvrez le fichier "iris.csv" avec un éditeur de texte classique afin de voir à quoi ressemblent les données (combien de lignes, de colonnes, quel attributs, etc.). Utilisez ensuite la commande de la librairie pandas **read_csv**$(\cdots)$ avec les bons paramètres afin de charger les données.
2. Affichez les histogrammes des différents attributs. Vous pouvez pour cela utiliser la fonction **distplot** de la bibliothèque seaborn. Que pouvez-vous dire sur leurs distributions ?
3. Calculez les coefficients de corrélations entre les différents attributs à partir de la fonction **corr()** de pandas. Commentez.
4. Visualisez vos données et la matrice de corrélation en utilisant les fonctions **pairplot()** et **heatmap()** de Seaborn. Commentez.
5. En supposant que les attributs suivent une distribution normale, calculez les intervalles de confiance pour les différentes corrélations. Commentez vos résultats.

B Données multivariées : anthropométrie

Dans cet exercice, nous allons étudier le jeu de données "mansize" qui contient des données anthopométriques relevées dans une prestigieuse université de médecine sur une population d'étudiants de licence volontaires.

1. Ouvrez le fichier “mansize.csv" avec un éditeur de texte classique afin de voir à quoi ressemblent ces données (nombre de lignes, colonnes, type de séparateurs, etc). Ensuite, utilisez la commande pandas **read_csv**$(\cdots)$ avec les bons paramètres afin de charger ces données sous forme d'une matrice.
2. Utilisez la fonction **describe**$(\cdot)$ sur ces données. Que fait cette fonction ? Commentez les résultats sur vos données.
3. Affichez les histogrammes des différents attributs. Que pouvez-vous dire sur leur distribution ?
4. Utilisez les commandes **corr()** et **pairplot()** pour confirmer vos résultats, puis **heatmap()** pour visualiser les corrélations. Commentez. Que pouvez-vous dire sur l'utilisation en archéologie de la longueur du fémur pour prédire la taille d'un individu ?
5. Calculez les intervalles de confiance pour vos coefficients de corrélation (on fera l'hypothèse que les attributs suivent tous des distributions normales). Commentez vos résultats.
6. A partir des questions précédentes (et de vos analyses des coefficients de corrélation et de détermination), que pouvez-vous dire des liens entre les différentes variables anthropométriques de ces données ?

C Test d'indépendance et variables catégorielles

Dans cet exercice, nous voulons étudier la possible dépendance entre plusieurs variables météorologiques relevées dans différentes villes.

1. Ouvrez le jeu de données “weather.csv" et décrivez les différentes variables et leurs valeurs à partir d'histogrammes.
2. Créez le tableau de contingence avec la commande **crosstab()** de pandas pour avoir le tableau croisé entre la variable “outlook" et la variable “temperature". Affichez et commentez la répartition des données dans le tableau résultant et déduisez le nombre de degrés de liberté de ce problème.
3. Utilisez la commande **chi2_contingency**$(\cdot)$ de la bibliothèque scipy.stats sur votre tableau. A partir des résultats et éventuellement en calculant d'autres indices, que pouvez vous conclure sur la dépendance entre les 2 variables ?
4. En vous inspirant des questions précédentes, établissez s'il existe un lien de dépendance entre les autres variables présentes dans les données (outlook/humity,temperature/humidity).