TP1: Statistiques descriptives univariées

Le Chenadec Gilles

August 30, 2022

1 Informations préliminaires (mais néanmoins importantes)

- Le TP est probablement noté. 1 ou 2 TPs de la séquence seront entièrement notés. La note finale de TP comprendra des points :
  - sur la réponse aux questions et l'argumentation des TPs notés;
  - sur la forme et la qualité de la rédaction des TPs notés;
  - sur le rendu en qualité et en ponctualité de tous les TPs.
- Il vous est demandé de commencer le compte-rendu de ce TP en séance et de le déposer dans Moodle, 1 semaine jour pour jour après la séance (si la séance a lieu, le lundi 7 de 8h10 à 12h15, la date limite est le lundi 14 à 12h15, un rendu après est à ranger dans la catégorie "risque").
- Vous devez déposer dans Moodle un **fichier PDF** (**Mettre dans le nom du fichier, les deux noms en cas de binome**) dans lequel vous reporterez tous vos résultats:
  - réponses textuelles aux questions,
  - si vous avez besoin de mettre des équations :
    - word est équipé d'un éditeur d'équations
    - utilisez l'appareil photo de votre smartphone si vous préférez le papier crayon
  - codes python,
  - figures avec indication des abscisses et ordonnées,
  - valeurs numériques des résultats des calculs, etc.
- **Il est très fortement conseillé :**
  - de **prendre le clavier** à tour de rôle pour progresser
  - de **justifier** la réponse aux questions posées
  - de **ne pas plagier** le code, les phrases ni du cours, ni des camarades (l'école est dotée du logiciel anti-plagiat compilatio qui compare les documents rendus par les étudiants et les documents disponibles sur internet).

2 Statistiques Descriptives univariées sur des données d'Iris

Les techniques de statistiques descriptives vont être illustrées via l'analyse des données célèbres, collectées par Edgar Anderson. Il s'agit des mesures en centimètres des variables suivantes : longueur du sépale (sepalLength), largeur du sépale (sepalWidth), longueur du pétale (petalLength) et largeur du pétale (petalWidth) pour trois espèces d'iris : setosa, versicolor et virginica.

Figure 1 – Iris Setosa, Versicolor, Virginica

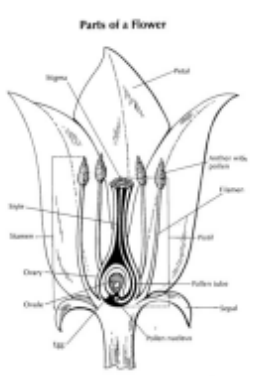

Figure 2 – Nomenclature des parties d'une fleur

2.1 Analyse préalable

Chargement des données

Question 2.1: Taper et comprendre les commandes Python suivantes:

```
import numpy as np
import pandas as pd
df = pd.read_csv('iris.csv', sep=',')
print(df.values)

sepallength = df["sepallength"].values
sepalwidth = df["sepalwidth"].values
petallength = df["petallength"].values
petalwidth = df["petalwidth"].values
species = df["class"].values

speciesname = np.unique(species)
variablename = df.keys().values[0:-1]
```

Comprendre les données (signification des individus et des variables)

Question 2.2: Quel est le nombre d'individus statistiques?

Question 2.3: Trouver les variables qualitatives et leurs modalités associées. Sont-elles nominales ou ordinales ?

Question 2.4: Trouver les variables quantitatives. Sont-elles continues ou discrètes?

2.2 Étude de la variable species

Question 2.5: Quels sont les effectifs de chaque modalité ?

Question 2.6: Les représentations graphiques classiques liées aux variables qualitatives sont la représentation en secteurs ou camembert (pie), la représentation en bâtons (hist). Représenter ces graphiques. (pour Python vous pouvez utiliser matplotlib.pyplot)

2.3 Étude de la variable petalLength

Deux approches permettent d'étudier cette variable ; il s'agit de l'approche graphique et celle créant des résumés numériques. Se rapporter aux cours pour les définitions des outils proposés.

Première approche : graphique

Question 2.7: Tracer l'histogramme en fréquences et l'histogramme des fréquences cumulées. Faire varier le nombre d'intervalles de l'histogramme.

Question 2.8: Décrire les caractéristiques de l'histogramme et analyser ces caractéristiques en fonction du nombre de classes.

Question 2.9: Tracer la boite à moustaches (*boxplot*) et rappeler les différents éléments la constituant.

Deuxième approche : résumés numériques

Question 2.10: Calculer les résumés numériques de localisation (moyenne et médiane) et ceux de dispersion : (écart-type, variance et quartiles). Retrouver en particulier, les valeurs des éléments de la boite à moustache.