TP1 : Statistiques descriptives univariées

Tanguy ROUDAUT — Tadios QUINIO FIPASE 24

30 Août 2022

1 Statistiques Descriptives univariées sur des données d'Iris

1.1 Analyse préalable

Comprendre les données (signification des individus et des variables)

Question 2 : Quel est le nombre d'individus statistiques ?

Le nombre d'individus statistique est de 150, il peut être obtenu grâce au code *Python* suivant :

```
print("Nombre d'individus statistiques: ", len(df.values))
```

Listing 1 – Code Python pour obtenir le nombre d'individus statistiques

```
Nombre d'individus statistiques:  150
```

Listing 2 – Résultat du code

Question 3 : Trouver les variables qualitatives et leurs modalités associées. Sont-elles nominales ou ordinales ?

La variable qualitative est la *class*. Il y a en tout trois modalités, qui sont *Iris-setosa, Iris-versicolor* et *Iris-virginica*. Les modalités sont nominales, une espèce n'est pas plus importante qu'une autre, il n'y a donc pas d'ordre.

```
print(speciesname)
```

Listing 3 – Code Python pour obtenir les modalités

```
['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']
```

Listing 4 – Résultat du code

Question 4 : Trouver les variables quantitatives. Sont-elles continues ou discrètes ?

Les variables qualitatives sont discrètes, elles correspondent aux différentes mesures de l'iris : *sepallength, sepalwidth, petallength* et *petalwidth*.

```
print(variablename)
```

Listing 5 – Code Python pour obtenir les variables qualitatives

```
['sepallength' 'sepalwidth' 'petallength' 'petalwidth']
```

Listing 6 – Résultat du code

1.2 Étude de la variable species

Question 5 : Quels sont les effectifs de chaque modalité ?

Le code *Python* suivant permet de calculer l'effectif de chaque espèce, soit les modalités de la variable qualitative *class*. On trouve qu'au final il y a 50 Iris de chaque espèce.

```
for spe in species:
    if spe == 'Iris-setosa':
        iris_setosa += 1
    elif spe == 'Iris-versicolor':
        iris_versicolor += 1
    elif spe == 'Iris-virginica':
        iris_virginica += 1

print("effectif de la modalité iris_setosa", iris_setosa)
print("effectif de la modalité iris_versicolor", iris_versicolor)
print("effectif de la modalité iris_virginica", iris_virginica)
```

Listing 7 – Code Pyton pour déterminer les effectifs de chaque modalités

```
effectif de la modalité iris_setosa 50
effectif de la modalité iris_versicolor 50
effectif de la modalité iris_virginica 50
```

Listing 8 – Résultat du code

Question 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)

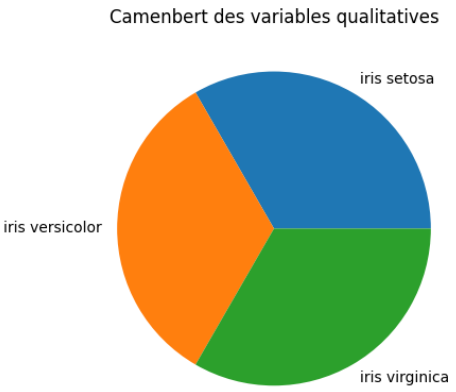

FIGURE 1 – Diagramme camembert des variables qualitatives

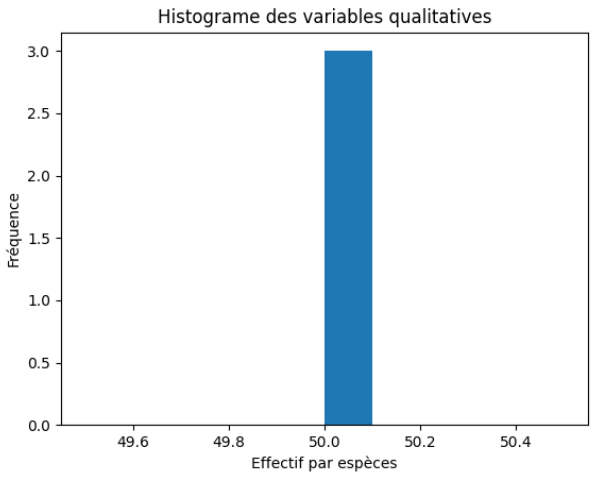

FIGURE 2 – Histogramme des variables qualitatives

```
nb_species = np.array([iris_setosa, iris_versicolor, iris_virginica])
plt.pie(nb_species, labels=["iris setosa", "iris versicolor", "iris virginica"])
plt.title("Camenbert des variables qualitatives")
plt.show()

plt.hist(nb_species)
plt.title("Histograme des variables qualitatives")
plt.show()
```

Listing 9 – Code Python pour tracer le diagramme camenbert et l'histogramme

1.3 Étude de la variable petalLength

Première approche : graphique

Question 7 : Tracer l'histogramme en fréquences et l'histogramme des fréquences cumulées. Faire varier le nombre d'intervalles de l'histogramme.

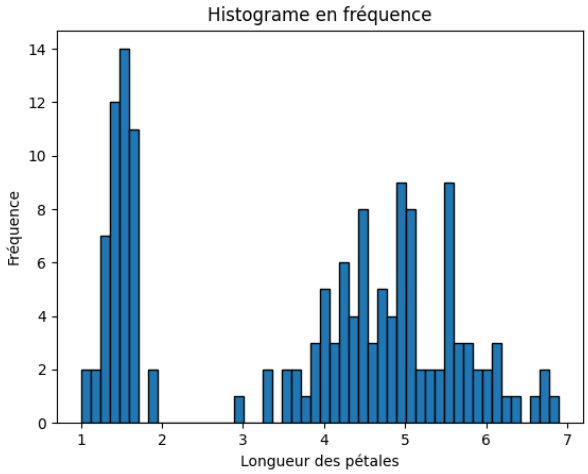

FIGURE 3 – histogramme en fréquences de la variable petalLength

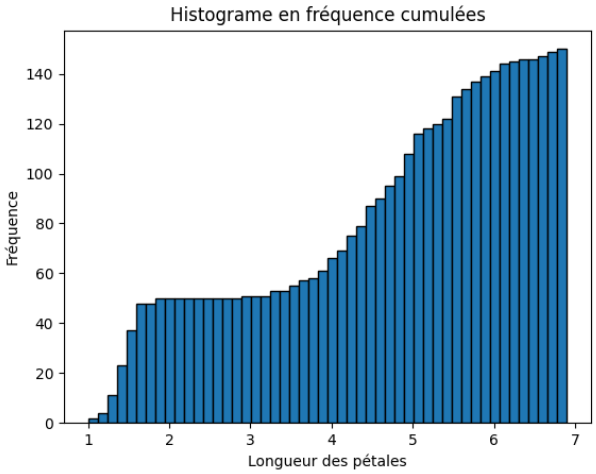

FIGURE 4 – histogramme en fréquences cumulées de la variable petalLength

```python
n, x, _ = plt.hist(petallength, 50, edgecolor='black')
plt.title("Histograme en fréquence")
plt.show()

n_cumul, _, _ = plt.hist(petallength, 50, cumulative=True, edgecolor='black')
plt.title("Histograme en fréquence cumulées")
plt.show()
```

Listing 10 – Code python pour tracer les histogrammes

Question 8 : Décrire les caractéristiques de l'histogramme et analyser ces caractéristiques en fonction du nombre de classes.

On remarque sur l'histogramme en fréquence qu'une majorité des longueurs de pétale est située entre 1 et 2 cm. Si l'on tient compte du nombre de classes, alors on peut penser à deux cas différents :

1. Le premier serait que deux des trois classes ont majoritairement une longueur de pétale qui varie entre 1 et 2 cm. Dans ce cas, la troisième espèce a une longueur de pétale qui varie entre 3 et 7 cm.
2. Le second cas serait qu'une des espèces a une longueur de pétale qui varie beaucoup moins que les deux autres. Par exemple l'espèce 1 varie entre 1 et 2 cm, tandis que l'espèce 2 et 3 varie entre 3 et 7 cm. Si l'on suit une logique de probabilité, il est donc évident que l'effectif des longueurs de pétale entre 1 et 2 cm soit plus important puisqu'il y a le même nombre d'effectifs dans chaque espèce et que l'intervalle est plus faible.

Le cas numéro 2 semble le plus évident. Si l'on regarde l'histogramme en fréquence cumulé, on constate que le nombre d'effectifs augmente de 2/3 quand le pétale mesure entre 4 et 7 cm.
Grâce à l'histogramme en fréquences et l'histogramme en fréquences cumulées on peut conclure qu'une des espèces à une longueur de pétale plus petite que les deux autres, mais qui varie également beaucoup moins

Question 9 : Tracer la boite à moustaches (boxplot) et rappeler les différents éléments la constituant.

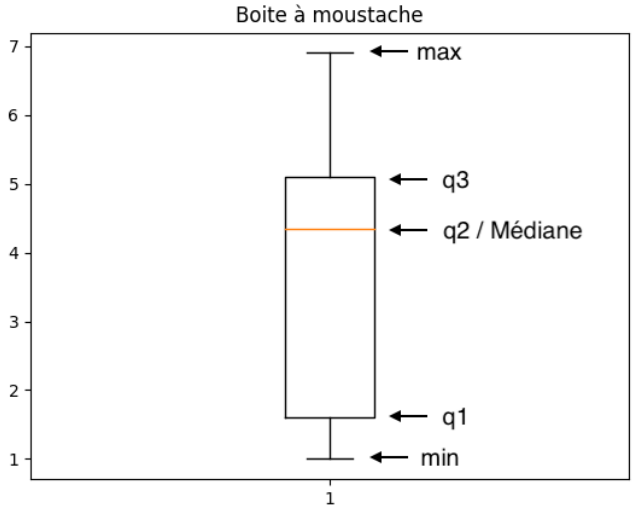

FIGURE 5 – Boite à moustache de la variable petalLength

```python
plt.boxplot(petallength)
plt.title("Boite à moustache")
plt.show()
```

Listing 11 – Code Python pour tracer la boite à moustache

Deuxième approche : résumés numériques

Question 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.

1. **Formules utilisées :**

– Moyenne :

$$xn = \frac{1}{n}\sum_{i=0} ni.xi \tag{1}$$

– Variance :

$$s^2_{n-1} = \frac{1}{n-1}\sum_{i=0} ni(xi - xn)^2 \tag{2}$$

– Médian :

$$m = \text{valeur de } x \text{ à la position } \frac{n_{cumul}}{2} \tag{3}$$

– Écart-type :

$$s_{n-1} = \sqrt{s^2_{n-1}} \tag{4}$$

– Quartiles :

$$q1 = \text{valeur de } x \text{ à la position } \frac{n_{cumul}}{4} \tag{5}$$

$$q2 = m \tag{6}$$

$$q3 = \text{valeur de } x \text{ à la position } \frac{3 * n_{cumul}}{4} \tag{7}$$

2. **Valeurs obtenues :**

Résumés numériques de localisation	moyenne	3.701
	médian	4.186
Résumés numériques de dispersion	écart-type	1.756
	variance	3.084
	q1	1.472
	q2	4.186
	q3	4.894

3. **Valeurs reportées sur les graphiques :**

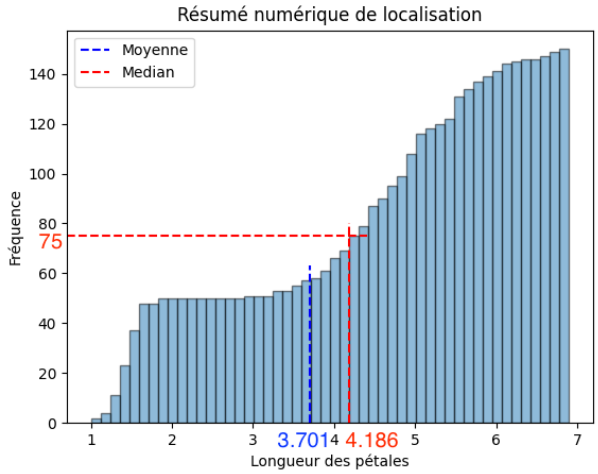

FIGURE 6 – histogramme en fréquences de la variable petalLength

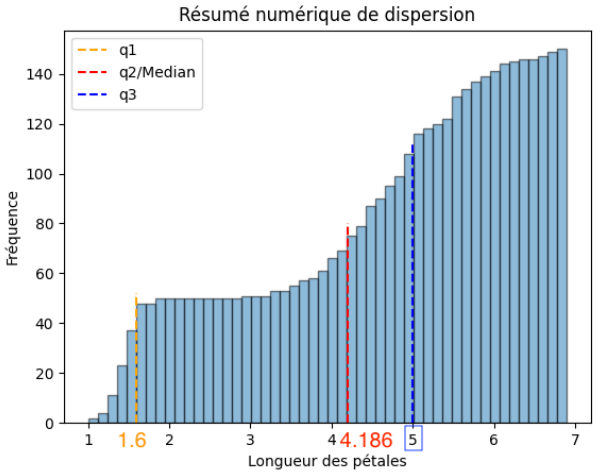

FIGURE 7 – histogramme en fréquences cumulées de la variable petalLength

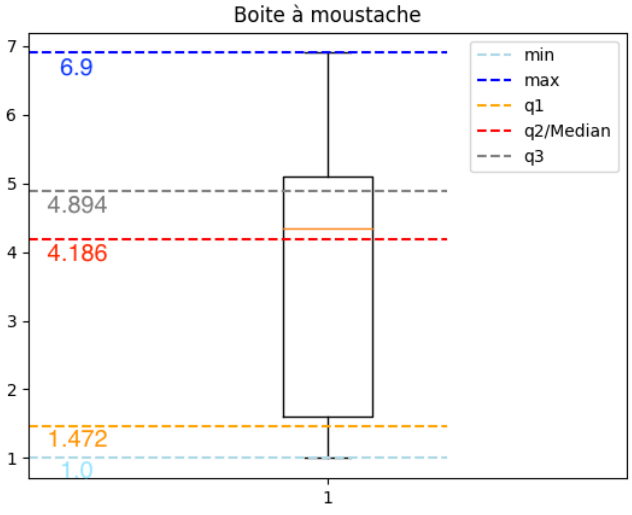

FIGURE 8 – Boite à moustache de la variable petalLength

4. **Code python :**

```python
def mean(n, x):
    n_max = len(df.values)
    res = 0
    for i in range(len(n)):
        res += (n[i] * x[i])
    ret = (1 / n_max) * res

    return round(ret, 3)

def median(n, x):
    n_median_index = np.where(n == (len(df.values) / 2))

    return float(x[n_median_index])

def variance(n, x, mean):
    n_max = len(df.values)
    res = 0
    for i in range(len(n)):
        res += n[i] * (x[i] - mean) ** 2
    ret = (1 / (n_max - 1)) * res

    return round(ret, 3)

def ecart_type(var):
    return round(np.sqrt(var), 3)

def quartiles(n, x):
    res = x[np.where(n < len(df.values) / 4)]
    q1 = res[-1]

    res = x[np.where(n < (3 * len(df.values)) / 4)]
    q3 = res[-1]

    q2 = median(n_cumul, x)

    return q1, q2, q3

print("RESUME NUMERIQUE DE LOCALISATION :")
print("Moyenne de petallenght:", mean(n, x))
print("Median de petallenght:", median(n_cumul, x), end="\n\n")

q1, q2, q3 = quartiles(n_cumul, x)
print("RESUME NUMERIQUE DE DISPERSION :")
print("Ecart-Type:", ecart_type(variance(n, x, mean(n, x))))
print("Variance:", variance(n, x, mean(n, x)))
print("Quartiles:", '\tq1 =', q1, '\tq2 =', q2, '\tq3 =', q3)
```

Listing 12 – Code python pour le résumé numérique

```
RESUME NUMERIQUE DE LOCALISATION :
Moyenne de petallenght: 3.701
Median de petallenght: 4.186

RESUME NUMERIQUE DE DISPERSION :
Ecart-Type: 1.756
Variance: 3.084
Quartiles:  q1 = 1.472  q2 = 4.186  q3 = 4.894
```

Listing 13 – Résultat numérique du code python