

Statistiques

Introduction à Python & Statistiques descriptives

Exercice 1 : Pièces défectueuses parmi 40

Les données se trouvent dans le fichier PiecesDef.txt.

```
import numpy as np
import matplotlib.pyplot as plt

Ech = np.loadtxt('PiecesDef.txt')
```

Il contient les valeurs prises par une variable qui associe à un lot de 40 pièces le nombre de pièces défectueuses. 100 lots de 40 pièces ont été analysés.

1. Etudier cette variable en donnant : (Cf. Chapitre 2 Statistiques descriptives d'une variable aléatoire discrète)
 - les valeurs prises par la variable étudiée
 - les effectifs absolus associés
 - les effectifs relatifs
 - les effectifs relatifs cumulés
2. Tracer la distribution de la variable étudiée.
3. Quel type de loi cela inspire-t-il pour ajuster la variable (ou distribution observée) ?
Estimer alors les effectifs théoriques d'avoir 0, 1, 2, 3, 4 etc... pièces défectueuses parmi 40.
4. Tracer sur une même figure la distribution théorique et la distribution observée

Exercice 2 : Loi normale

1. Créer un vecteur de 100 nombres aléatoires issus d'une loi normale centrée réduite (`np.random.randn` ou `np.random.normal`)

1.1. Calculez la moyenne, l'écart-type et la variance de la variable. Pour chacune de ces grandeurs vous utiliserez les fonctions de NumPy ; vous écrirez également vos propres lignes de code PYTHON qui utiliseront simplement des opérations de somme, division et multiplication.

Remarque : essayer d'écrire sans faire de boucle

1.2. Tracer la distribution de la variable étudiée

1.3. Donner la médiane et les quartiles et tracer la boîte à moustaches

2. Loi normale de moyenne 14 et d'écart-type 3

2.1. Tracer la densité de probabilité d'une variable aléatoire qui suit une loi normale de moyenne 14 et d'écart-type 3 et sa fonction de répartition.

2.2. Générer un vecteur de 100 nombres aléatoires issus d'une loi normale de moyenne 14 et d'écart-type 3. Tracer son histogramme et sa fonction de répartition.

2.3. Quelle est la probabilité que cette variable soit comprise entre 11 et 17 ?

Exercice 3 : Jeu de données <http://www.statsci.org/data/oz/ms212.html>

Les données se trouvent dans le fichier Exercices1.xlsx.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# On charge les données
X = pd.read_excel("Exercices1.xlsx", sheet_name=0, header=0, index_col=0)
# Pour le TP1 on travaillera sur les Array
XArray=pd.DataFrame.to_numpy(X)
```

<u>Variables</u>	<u>Description</u>
Height	Height (cm)
Weight	Weight (kg)
Age	Age (years)
Gender	Sex (1 = male, 2 = female)
Smokes	Regular smoker? (1 = yes, 2 = no)
Alcohol	Regular drinker? (1 = yes, 2 = no)
Exercise	Frequency of exercise (1 = high, 2 = moderate, 3 = low)
Ran	Whether the student ran or sat between the first and second pulse measurements (1 = ran, 2 = sat)
Pulse1	First pulse measurement (rate per minute)
Pulse2	Second pulse measurement (rate per minute)
Year	Year of class (93 - 98)

1. Etudier les variables séparément: Height, Weight, Age, Pulse1 and Pulse2 (attention certaines variables ont des données manquantes). Supprimer les individus pour lesquels il manque des valeurs (is.nan ou ~is.nan)
2. Représenter les individus en fonction des variables Height et Weight (plt.scatter)
3. Donner le coefficient de corrélation entre ces 2 variables
4. Représenter maintenant les individus en fonction de Pulse1 et Pulse2. Donner une couleur ou un symbole différent aux participants en fonction de la variable Ran
5. Si vous avez répondu à toutes les questions vous pouvez reprendre les points de cet exercice 3 et refaire en travaillant directement sur les données au format dataframe.