BUT1 – OUTILS NUMÉRIQUES POUR LES STATISTIQUES DESCRIPTIVES TRAVAUX DIRIGÉS

IUT DE NANTES – DÉPARTEMENT INFORMATIQUE – 2021/2022

Table des matières

Organisation du temps de travail

Première semaine :

- TD1 (1h20) : cours sur les statistiques descriptives – premiers indicateurs numériques
- TP1 (2h40) : visualisation de données

Deuxième semaine :

- TD2 (1h20) : cours sur les statistiques à deux variables
- TP2 (2h40) : statistiques à deux variables

Troisième semaine :

- TD3 (1h20) : séance de transition
- TP3 (2h40) : analyse fréquentielle de chiffrement

Quatrième semaine :

- TP4 (2h40) : visualisation de données géographiques (Sae)
- TD4 (1h20) : *évaluation terminale sur machine*

Au total : 4 créneaux de TD et 8 créneaux de TP par étudiant.

Les deux premiers TD comprendront une partie cours et une partie travail sur feuille. La troisième séance de TD sera une séance vous permettant notamment de poser des questions sur des aspects du cours des deux premières semaines que vous n'auriez pas compris, de reprendre des exercices ou alors d'avancer sur le TP de la semaine.

TD 1 – Indicateurs numériques

Cette première feuille de TD a pour but de vous familiariser avec les premiers indicateurs numériques statistiques. Une partie de ce TD se fera sur machine.

Exercice 1.1 – Rappels des permiers indicateurs statistiques –

1. *À faire sur feuille.* On considère la série statistique suivante

$$X \coloneqq \{17, 13, 9, 6, 13, -3, -2, 3, 18, 0, 17, -1, 16\} .$$

   a) Calculer la moyenne $\overline{X}$ de la série statistique X.

   b) Calculer la variance $\mathbb{V}(X)$ de la série statistique X ; en déduire son écart type σ_X.

   c) Calculer la médiane de la série statistique X.

   d) Calculer le premier et le troisième quartile de X.

2. *Passage à Python.* On considère L une liste de nombres.

   a) Écrire une fonction `mean(L)` qui renvoie la moyenne arithmétique (ou empirique) de la série statistique `L`.

   b) Écrire une fonction `variance(L)` qui renvoie la variance de la série statistique `L`.

   c) Écrire une fonction `ecart_type(L)` qui renvoie l'écart type de la série statistique `L`.

   d) Écrire une fonction `median(L)` qui renvoie la médiane de la série statistique `L`.

   e) Écrire une fonction `first_quartile(L)` qui renvoie le premier quartile de la série statistique `L`.

   f) Écrire une fonction `third_quartile(L)` qui renvoie le troisième quartile de la série statistique `L`.

   Tester ces différentes fonctions pour retrouver les résultats de la question 1.

3. *Illustration de la sensibilité de la moyenne aux valeurs extrêmes.* On construit ci-dessous deux listes de nombres `listX1` et `listX2` grâce à `numpy`. Utiliser les fonctions précédentes pour afficher les moyennes et médianes des deux listes de nombres. Que pouvez-vous dire ?

```python
import numpy as np
listX1 = list(np.random.standard_normal(50))
listX2 = listX1 + [18,19,20,30]
```

Exercice 1.2 – Le but de cet exercice est de classer un certain nombre d'élèves qui candidatent à une formation sélective (comme par exemple un BUT d'informatique). Dans la suite, on utilisera les fichiers de notes suivants

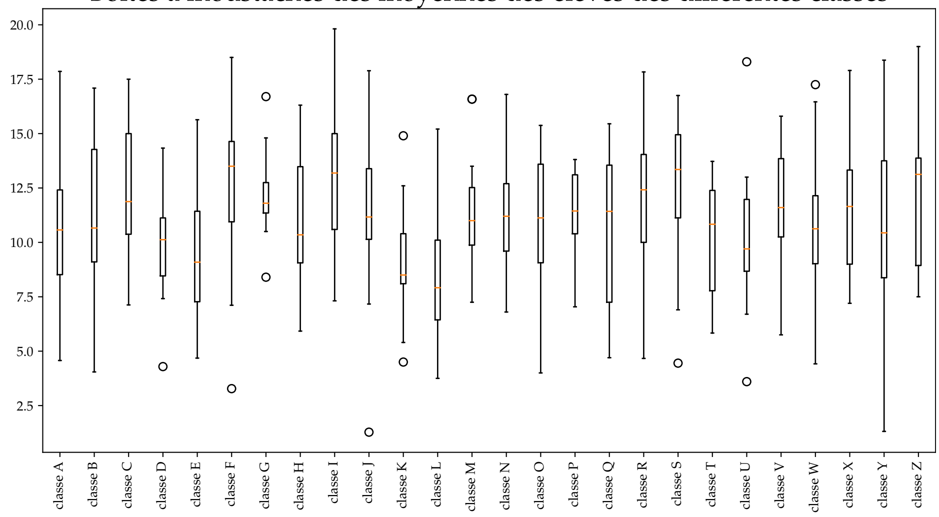
`classe_*.csv`

qui contiennent les notes de mathématiques d'élèves de différentes classes au cours de leur année de terminale. Dans la suite de l'exercice, nous utiliserons les fonctions présentes dans la bibliothèque `scipy` que l'on importera comme suit ; on utilisera encore la librairie `pandas`.

```python
import scipy as sp
import pandas as pd
```

1. Faire une fonction `visualisation()` qui affiche une boîte à moustache (boxplot) pour chaque classe en utilisant les moyennes des élèves de la classe. On donne une illustration du graphique attendu en Figure 1.

FIGURE 1
Boîtes à moustaches des moyennes des élèves des différentes classes



   Peut-on utiliser ces moyennes pour faire un classement des élèves ?

2. Faire une fonction `make_Resume()` qui renvoie un fichier `Resume.csv` qui contient trois colonnes :
   — une première contenant la lettre référençant la classe ;
   — une seconde contenant la moyenne de la classe que vous pourrez calculer en utilisant la fonction `sp.mean` ;
   — la dernière contenant l'écart type des moyennes des élèves de la classe, que vous pourrez calculer en utilisant la fonction `sp.std` .

3. *À faire sur feuille :* considérons une série statistique X, c'est-à-dire un ensemble de n nombres $x_1, x_2, \ldots, x_n$, de moyenne $\mu := \overline{X}$ et d'écart type $\sigma := \sigma_X$.
   a) Calculer la moyenne et l'écart type de la série statistique
   $$X' := \{x_1 - \mu, x_2 - \mu, \ldots, x_n - \mu\} \ .$$
   b) Faites de même pour la série statistique
   $$X'' := \left\{\frac{x_1 - \mu}{\sigma}, \frac{x_2 - \mu}{\sigma}, \ldots, \frac{x_n - \mu}{\sigma}\right\} \ .$$
   Que constatez-vous ?

4. En vous servant de la question précédente, faire une fonction `classement()` qui renvoie un fichier `.csv` avec trois colonnes contenant la classe de l'élève, son nom et sa moyenne renormalisée ; et qui affiche les 20 premiers étudiants sélectionnés avec ce classement. Pour cela, on

pourra utiliser la méthode `.sort_values` applicables sur les variables de type `dataframes` de la librairie `pandas`.

TD 2 – Série statistique à deux variables

Le but des exercices suivants est de vous familiariser avec la régression linéaire.

Exercice 2.1 – On considère la série statistique à deux variables suivante :

X	1.01	2.03	3.1	4.04	5.02
Y	3.3	5.8	9.4	11.7	14.6

1. Déterminer la droite de régression linéaire pour cette série statistique.
2. Calculer le coefficient de corrélation linéaire.

Exercice 2.2 – Voici la distribution des tailles et poids d'un groupe d'enfants de 6 ans.

Taille X (cm)	121	123	108	118	111	109	114	103	110	115
Poids Y (Kg)	25	22	19	24	19	18	20	15	20	21

1. Est-il judicieux d'effectuer une régression linéaire de Y en X ?
2. Effectuer cette régression et prédire le poids d'un enfant de 6 ans mesurant 120 cm.

Exercice 2.3 – Douze personnes sont inscrites à une formation. Au début de cette formation, ces stagiaires subissent une épreuve A notée sur 20. À la fin de la formation, elles subissent une épreuve B de niveau identique et également notée sur 20. Les résultats sont donnés dans le tableau suivant :

Épreuve A	3	4	6	7	9	10	9	11	12	13	15	4
Épreuve B	8	9	10	13	15	14	13	16	13	19	6	19

1. Représenter le nuage de points. Déterminer la droite de régression et calculer le coefficient de corrélation. Commenter.
2. Deux stagiaires semblent se distinguer des autres. Les supprimer et déterminer la droite de régression sur les dix points restants. Calculer le coefficient de corrélation. Commenter.

Exercice 2.4 – Le tableau suivant donne la distance d'arrêt d'une voiture sur revêtement moyen :

Vitesse X ($\mathrm{Km} \cdot \mathrm{h}^{-1}$)	20	30	50	70	90	110	130
Distance Y (m)	13	28	42	66	96	129	167

1. Représenter le nuage de points associés. Calculer le coefficient de corrélation.
2. Substituer à Y sa racine carré $Z = \sqrt{Y}$ et représenter le nouveau nuage de points. Analyser le résultat.

3. Calculer les coefficients de la droite de régression linéaire pour les variables X et Z puis tracer là. Calculer le coefficient de corrélation linéaire.
4. Estimer la distance nécessaire à l'arrêt d'une voiture à 80 $\mathrm{Km} \cdot \mathrm{h}^{-1}$ puis 140 $\mathrm{Km} \cdot \mathrm{h}^{-1}$.

EXERCICE 2.5 – Un hypermarché dispose de 20 caisses. On s'intéresse au temps moyen d'attente en fonction du nombre de caisses ouvertes. Le tableau ci-dessous donne X le nombre de caisses ouvertes et Y le temps moyen d'attente correspondant :

Nb de caisses ouvertes : X	3	4	5	6	8	10	12
Temps moyen d'attente (en min) Y	16	12	9.6	7.9	6	4.7	4

1. Calculer le coefficient de corrélation linéaire entre les deux variables.
2. Faut-il effectuer un ajustement par une droite ?
3. On effectue le changement de variable suivant : $X_1 = \ln(X)$ et $Y_1 = \ln(Y)$; Calculer les valeurs prises par X_1 et Y_1, puis calculer le coefficient de corrélation de (X_1, Y_1). Commenter.
4. Déterminer la droite de régression de Y_1 par rapport à X_1 et en déduire une relation ajustant Y par rapport à X du type $y = \alpha x^\beta$.

Email address: `johan.leray@univ-nantes.fr`

DÉPARTEMENT D'INFORMATIQUE – IUT DE NANTES