TD : Fouille de données

1 Exercice 1 : Étude de l'efficacité d'herbicides

Nous souhaitons étudier l'efficacité de trois herbicides sur trois plantes : blé, chiendent et liseron. Pour cela, des cultures de ces plantes ont été mises en présence de l'un des trois herbicides, ou d'aucun d'entre eux. Le nombre de plants vivants dans la culture a été compté avant l'expérience, et 10 jours après. Chaque combinaison plante - herbicide a fait l'objet de 20 expérimentations, plus un témoin sans herbicide (soit 240 expérimentations en tout).

Le tableau de données est disponible dans le fichier `herbicide.csv` [1].

Question 1.1

Charger le fichier `herbicide.csv` et afficher les données.

Nous allons utiliser le package `pandas` (*Python Data Analysis Library*)[2] afin de lire les données. Pandas est une bibliothèque permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

Commencer par importer pandas :

```
1 import pandas as pd
```

Lire la documentation sur la fonction `read_csv()` de pandas afin d'importer `herbicide.csv` dans un `DataFrame`. Les `DataFrame` sont une des structures de données utilisées par pandas pour stocker des données à analyser.

Question 1.2

Calculer la moyenne du nombre de plants initial, sa variance, son écart type, ainsi que les valeurs minimum et maximum

Importer le package numpy étudié en cours :

```
1 import numpy as np
```

Utiliser les fonctions `np.mean()`, `np.std()`, `np.min()` et `np.max()` afin de répondre à la question.

Question 1.3

Calculer le pourcentage de plantes ayant survécu pour chaque expérimentation, et l'ajouter dans une nouvelle colonne

Afin d'ajouter une colonne à un `DataFrame` pandas, il suffit de la définir comme suit :

```
1 data["nouvelle colonne"] = ... # definition de la colonne
```

[1] http://germain-forestier.info/dataset/herbicide.csv
[2] https://pandas.pydata.org/

Question 1.4

Combien d'expérimentations ont donné lieu à moins de 5% de plants survivants ?

Vous pouvez spécifier un critère de sélection dans un `DataFrame` pandas comme suit :

```
data[data["colonne"] = 0.5]
```

Question 1.5

Le témoin correspond à l'absence d'herbicide. Extraire les lignes du tableau qui correspondent au témoin et les mettre dans une nouvelle variable que l'on appellera "temoin".

Afin de faire une sélection sur une chaîne de caractères dans un `DataFrame`, utiliser la fonction `str.contains()`

```
data2 = data[data["colonne"].str.contains("filtre")]
```

Question 1.6

Calculer la moyenne et l'écart type du pourcentage de plants ayant survécus sur le témoin.

Question 1.7

De la même manière, calculer la moyenne et l'écart type du pourcentage de plants ayant survécus pour chacun des trois herbicides. Quel herbicide vous semble le plus efficace globalement ?

Question 1.8

Quelle plante est celle qui a le mieux résisté aux herbicides, dans l'ensemble ?

Question 1.9

Tracer un histogramme représentant le taux de plantes survivantes.

Question 1.10

Tracer un camembert représentant la proportion des différentes plantes dans l'étude.

Utiliser la fonction `pie()` de matplotlib et `value_counts()` de pandas.

Question 1.11

Tracer un graphique représentant le taux de survivants en fonction de l'espèce de plante. Dans l'ensemble, quelle espèce résiste le mieux aux trois herbicides ?

Commencer par importer le package `matplotlib` :

```
import matplotlib.pyplot as plt
```

Lire la documentation et utiliser la fonction `boxplot()` afin de produire un diagramme en boîte.

Question 1.12

Tracer un graphique représentant le taux de survivants en fonction de l'herbicide. Dans l'ensemble, quel herbicide semble le plus efficace ? Utiliser encore une fois la fonction `boxplot()`.

Question 1.13

Tracer un diagramme de dispersion des taux de survivants en fonction de l'herbicide puis en fonction des types de plantes.

Installer `seaborn`[3] qui est un package Python pour effectuer de la visualisation de données (`pip3 install seaborn`), puis importer le package :

```
1 import seaborn as sns
```

Lire la documentation et utiliser la fonction `stripplot()` de `seaborn`.

Question 1.14

Tracer un graphique représentant le taux de survivants en fonction de l'herbicide et de l'espèce de plante. Commenter l'efficacité de chaque herbicide sur chaque type de plante.

Commencer par construire un tableau 2D contenant les taux moyens de survie pour chaque plante et chaque herbicide (3×3). Utiliser pour cela la fonction `pivot_table()` de pandas. Afficher ensuite une heatmap de ce tableau 2D à l'aide de la méthode `seaborn` : `heatmap()`.

Question 1.15

En vous aidant du graphique précédent, répondre aux questions suivantes :

1. Quel herbicide est le plus approprié pour appliquer sur une route, où l'on souhaite qu'aucune plante ne pousse ?
2. Quel herbicide est le plus approprié pour appliquer sur un champ de blé, où l'on souhaite que le blé pousse, mais pas le chiendent ni le liseron ?
3. Quelle expérience peut-on envisager pour améliorer l'efficacité du traitement de ce champ de blé ?

2 Exercice 2: Clairance de Cockcroft

La formule de Cockcroft permet de calculer la clairance rénale à partir d'un dosage de la créatininémie (dans le sang). Chez la femme, la formule est la suivante :

$$clairance = \frac{(140 - age) \times poids \times 1.04}{creatininemie}$$

Nous souhaitons vérifier la validité de cette formule pour des patientes âgées. Pour cela, la clairance rénale a été mesurée dans les urines (de 24h) sur un échantillon de 80 patientes âgées, et nous allons comparer cette mesure à l'estimation fournie par la formule de Cockroft.

[3] `https://seaborn.pydata.org/`

1. Charger le fichier `cockroft.csv`[4].
2. Représenter graphiquement l'âge des patientes, et leur poids.
3. Calculer la clairance rénale pour chaque patient, et la mettre dans une nouvelle colonne du tableau de donnée, que l'on appellera "clairance cockroft".
4. Représenter graphiquement la clairance rénale mesurée chez les patientes en fonction de la clairance calculée.
5. Ajouter un titre au graphique précédent, et modifier les axes pour qu'ils indiquent "Clairance calculée par la formule de Cockroft" et "Clairance mesurée sur les urines de 24h".

[4] `http://germain-forestier.info/dataset/cockroft.csv`