

SAE 2.04 - RAPPORT

Tom Moriceau, Pierre Lechat

1) Nos données, notre problématique :

Notre fichier de vue `vue.csv` contient plusieurs séries statistiques sur les résultats scolaires de différentes régions pour l'année 2022-2023 :

- La **population** est l'ensemble des élèves inscrits dans les établissements scolaires des différentes régions pour l'année 2022-2023.
- La **première série** correspond au taux de réussite général (en pourcentage) dans chaque région.
- La **deuxième série** correspond au nombre d'élèves inscrits dans chaque région.
- La **troisième série** correspond au nombre de candidats au bac général dans chaque région.
- La **quatrième série** correspond à la note moyenne obtenue à l'écrit du bac général dans chaque région.
- La **cinquième série** correspond à la valeur ajoutée de la note pour chaque région.

1	libelle_region	taux_de_reussite	nombre_d_eleves	nb_candidats_g	note_a_l_ecrit_g	va_de_la_note_g
2	Martinique	100	285	41	10.3	-1.2
3	Martinique	100	132	32	8.3	-0.2
4	Bourgogne-Fran	100	301	67	11.2	-0.1
5	Hauts-de-France	100	400	68	10.9	0.3
6	La Réunion	100	740	111	9.2	1.1
7	Auvergne-Rhône	100	191	33	12	0.7
8	Bretagne	100	196	35	12.1	2.5
9	Bretagne	100	406	92	10.7	-0.8
10	Nouvelle-Aquitai	100	163	40	13.2	2.3

Les 10 premières lignes de notre fichier, avec le nom des colonnes

En utilisant ces données, nous allons essayer de répondre à la problématique suivante :

Le nombre d'élèves influe-t-il sur le taux de réussite sur l'année scolaire 2022-2023 ?

En choisissant la première série statistique (le taux de réussite) comme variable endogène et les autres séries comme variables explicatives, la régression linéaire multiple nous permettra d'obtenir une estimation du taux de réussite en fonction des autres séries statistiques.

Les paramètres de la régression linéaire multiple nous informeront des descripteurs qui influencent le plus le taux de réussite. Le coefficient de corrélation multiple nous permettra d'apporter une réponse à la problématique.

2) Import des données, mises en forme, centrage-réduction

Pour importer les données depuis le fichier `vue.csv`, on utilise la bibliothèque *pandas* de Python :

```
import pandas as pd
```

Et on lit les données grâce à la fonction *read_csv* de *pandas* :

```
vueDF = pd.read_csv("vue.csv")
```

On a décidé de supprimer les cases vides provoquant des valeurs manquantes pour éviter des éventuels problèmes de mise en forme :

```
vueDF = vueDF.dropna()
```

Pour centrer et réduire les données, on a créé une fonction *Centreduire* :

```
def Centreduire(T):  
    T = np.array(T, dtype = np.float64)  
    moyennes = np.mean(T, axis = 0)  
    ecarts_types = np.std(T, axis = 0)  
    Res = (T - moyennes) / ecarts_types  
  
    return Res
```

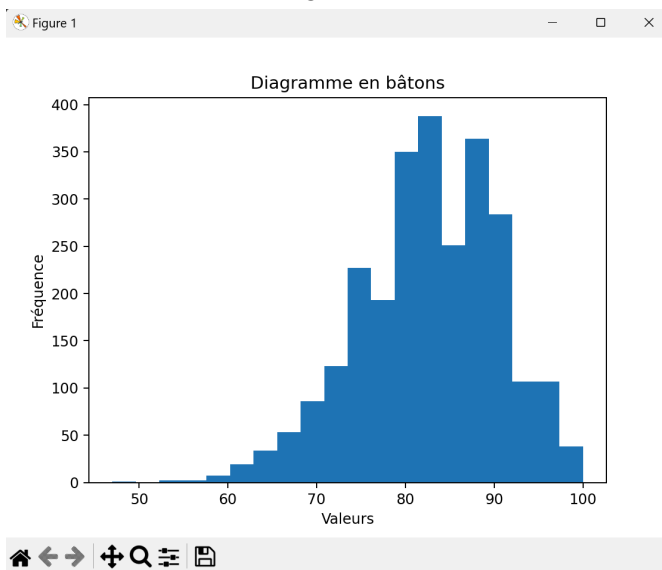
Dans cette fonction, on convertit le tableau *T* prit en entrée en un tableau numpy, on calcule la moyenne et l'écart-type de chaque colonne, puis on le soustrait par la moyenne et on le divise par l'écart-type de chaque colonne pour finalement renvoyer le tableau centré et réduit.

3) Exploration des données par représentations graphiques

Pour visualiser les données, on a décidé de créer un diagramme en bâtons, dont vous pouvez retrouver le code dans la fonction *DiagBatons* de notre fichier :

```
def DiagBatons(Colonne):  
    m = np.min(Colonne)  
    M = np.max(Colonne)  
    inter = np.linspace(m, M, 21)  
  
    plt.figure()  
    plt.hist(Colonne, bins=inter)  
    plt.title('Diagramme en bâtons')  
    plt.xlabel('Valeurs')  
    plt.ylabel('Fréquence')  
    plt.show()  
  
DiagBatons(vueAR[:,[0]])
```

Après exécution du programme, voici ce à quoi notre diagramme en bâtons ressemble :



Ici, on analyse la distribution du taux de réussite, puisqu'on a choisi la première colonne des valeurs. On pourrait aussi analyser d'autres distributions pour les comparer, comme la distribution du nombre d'élèves dans les établissements.

Pour le taux de réussite en tout cas, on peut voir une concentration des valeurs autour de 80-90, ce qui nous montre une tendance vers un taux de réussite élevé de manière générale, mais on remarque quand même une variation qu'il faudra analyser ensuite.

On a également choisi de calculer le Coefficient de Détermination R^2 et le Coefficient de corrélation multiple pour mieux comprendre la corrélation.

```
modele_regression = LinearRegression()
modele_regression.fit(X, y)
coefficient_determination = modele_regression.score(X, y)
coefficient_correlation_multiple = np.sqrt(coefficient_determination)

print("Coefficient de détermination R^2 : " + str(coefficient_determination))
print("Coefficient de corrélation multiple : " + str(coefficient_correlation_multiple))
```

L'affichage des résultats :

```
Coefficient de détermination R^2 : 0.1823097368558001
Coefficient de corrélation multiple : 0.4269774430292543
```

Le coefficient de détermination R^2 est de 0.182, donc seulement 18.2% de la variance du taux de réussite est expliquée par les variables indépendantes (donc : nombre d'élèves, nombre de candidats, note à l'écrit, valeur ajoutée de la note).

Le coefficient de corrélation multiple, lui, est de 0.427, ce qui montre une relation linéaire modérée entre les variables explicatives et le taux de réussite.

Donc les valeurs obtenues montrent que le modèle de régression linéaire multiple n'explique pas la variance du taux de réussite, on peut donc conclure sur une absence de corrélation significative.

Conclusion

Pour rappel, notre problématique était la suivante : “Le nombre d'élèves influe-t-il sur le taux de réussite sur l'année scolaire 2022-2023 ?”

Notre réponse :

Nous pensons que les résultats obtenus montrent qu'il n'y a pas de corrélation significative entre le nombre d'élèves et le taux de réussite pour l'année scolaire 2022-2023.

En effet, nous avons calculé deux paramètres clés à partir de notre modèle de régression linéaire multiple :

1) **Coefficient de détermination R^2 : 0.182**

Cela signifie que seulement 18.2% de la variance du taux de réussite est expliquée par les variables indépendantes qu'on a sélectionnées, et ça indique que notre modèle ne parvient pas à capturer la majeure partie des variations dans le taux de réussite.

2) **Coefficient de corrélation multiple : 0.427**

La valeur montre une relation linéaire modérée entre les variables explicatives et le taux de réussite, mais un coefficient de corrélation multiple de 0.427 n'est pas assez élevé pour indiquer une forte corrélation.

Nos interprétations personnelles :

Ce que nous n'avons pas pris en compte, c'est la complexité des facteurs éducatifs : on sait évidemment que le taux de réussite scolaire est influencé par beaucoup d'autres facteurs, comme la qualité de l'enseignement, le cadre familial, et bien d'autres. Se baser seulement sur le nombre d'élèves est peu trop précis.

Une interprétation absurde pourrait être que la météo pendant les périodes d'examen a un impact sur le taux de réussite. On pourrait se dire que les élèves réussissent mieux lorsqu'il fait beau, et que nos données n'ont pas pris en compte cette variable météorologique.

Pour conclure, les résultats obtenus par notre analyse statistique indiquent qu'il n'existe pas de corrélation significative entre le nombre d'élèves et le taux de réussite scolaire pour l'année 2022-2023. Les variables explicatives que nous avons choisies n'expliquent qu'une petite partie de la variance du taux de réussite, il faudrait donc prendre en compte d'autres facteurs qui jouent un rôle plus déterminant.