Semestre 2

Rapport SAE 2.04

Livrable n°3 : Statistiques

Stéphane Antoine & Jonas Brisson

1) Les données colleges.csv - Problématique

(a) Présentation des données

Le fichier colleges.csv regroupe plusieurs séries statistiques de collèges répertoriés dans notre base de données :

- La population est donc l'ensemble des collèges, qui sont représentés par leur nom et leur uai (leur matricule).
- La 1ère variable statistique est le nombre de candidats général
- La 2ème est le taux de réussite général
- La 3ème est le taux d'accès de la 6ème à la 3ème
- La 4ème est le nombre de mentions en général
- La 5ème est l'ips (indice de position sociale)
- La 6ème est l'écart-type de l'ips

(b) Problématique

En utilisant ces données, on va essayer de répondre à la problématique suivante :
Dans les différents collèges, est ce que le taux de réussite général peut être affecté par les autres données de notre fichier ?

2) Import des données, mise en forme

(a) Importer les données en Python

Notre vue est importée sous forme de DataFrame avec la commande python suivante :

```
#création du dataframe à partir du fichier CSV
collegesDF = pd.read_csv("/home/etuinfo/jbrisson/Documents/Sae/S2.04/partie3/colleges.csv")
```

(b) Mise en forme

Nous avons transformé notre DataFrame en array liste afin de pouvoir mieux le manipuler :

```
#création d'un np.array à partir du dataframe
collegesAR0 = collegesDF.to_numpy()
```

Nous avons supprimé les deux premières colonnes (uai et nom de l'établissement) dont nous n'avions pas besoin à part pour une meilleure lisibilité au début de l'exercice grâce aux commandes suivantes :

```
#on enlève l'uai et le nom de l'établissement de notre array
collegesAR = collegesAR0[:,[2,3,4,5,6,7]]
```

(c) Centrer-réduire

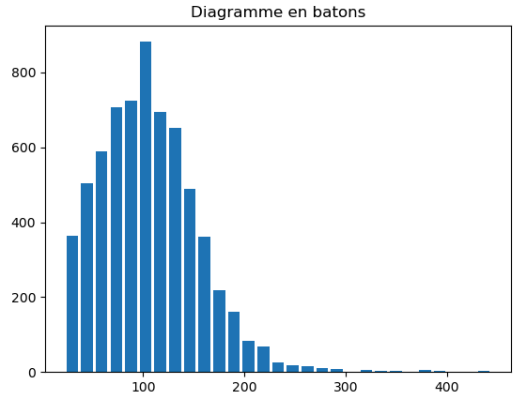
Nous avons utilisé une fonction que nous avions créée pour centrer-réduire nos données :

```
#fonction pour centrer réduire collegesAR
def centreduit(t):
    t = np.array(t, dtype=np.float64)
    lig, col = t.shape
    temp = np.zeros((lig, col))
    for i in range(0, lig):
        for j in range(0, col):
            temp[i][j] = (t[i][j] - np.mean(t[:,j], axis = 0)) / np.std(t[:,j], axis = 0)
    return temp
```

3) Exploration des données :

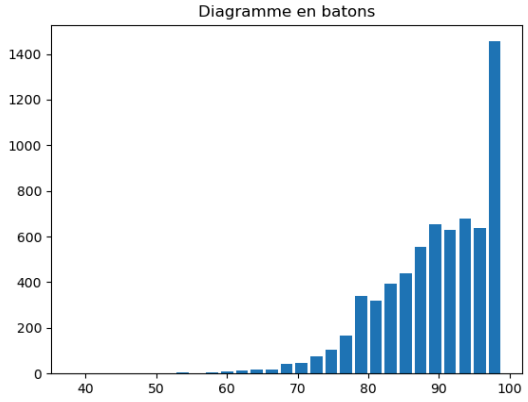
a. Représentations graphiques

On choisit d'étudier les diagrammes en bâtons des nos variables statistiques :



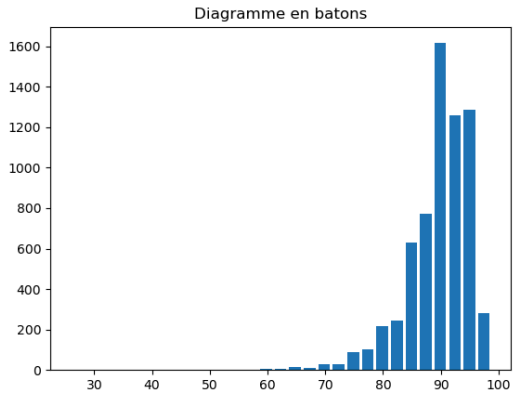
Le nombre de candidats en général

On remarque qu'il y a en moyenne 100 élèves par collège et qu'il y a beaucoup de collèges qui ont entre 30 et 180 élèves.



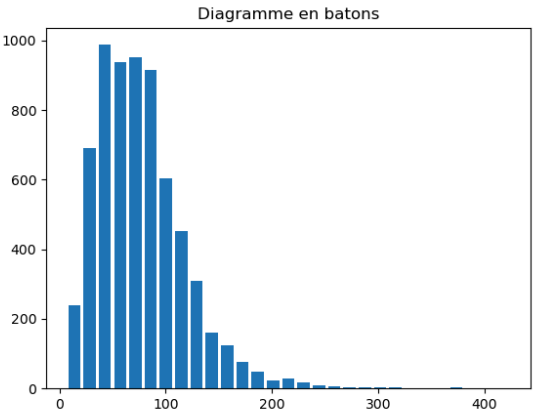
Le taux de réussite

On remarque qu'il y a un tiers des collèges qui ont entre 90 et 100% de taux de réussite, qu'un tiers est entre 80 et 90% de taux de réussite et qu'un tiers est entre 55 et 80% de taux de réussite.



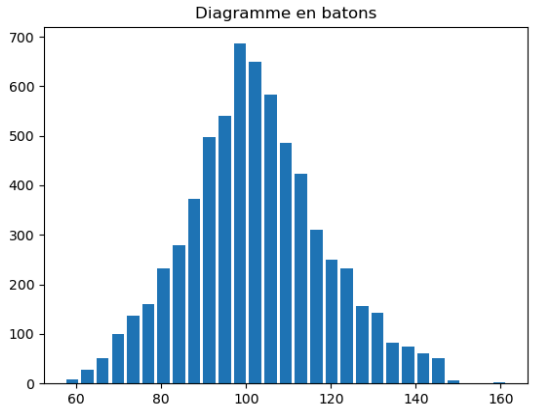
Taux d'accès des collèges

On remarque qu'il y a une majorité de collèges entre 90 et 95 % de taux d'accessibilité. Le reste des collèges est entre 60 et 97%.



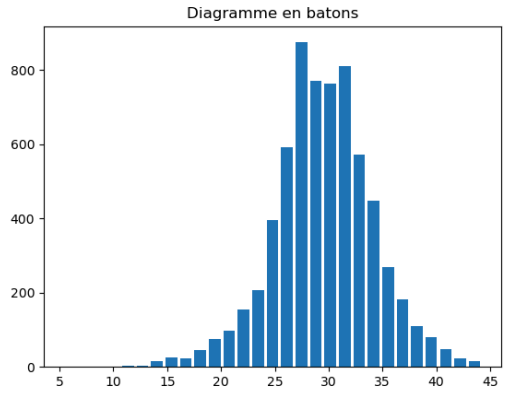
Nombre de mentions totales

On remarque que la moitié des collèges ont entre 20 et 100 mentions en tout. Il n'y a pas de collèges avec aucune mention en tout et il y en a environ 200 entre 150 et 200 mentions. Il y a moins de 10 collèges avec 370 mentions en tout.



Ips des collèges

On peut observer que les IPS des collèges se trouvent en moyenne aux alentours de 100, c'est à dire que la moyenne des élèves se trouve dans une situation sociale d'apprentissage correcte, cependant on voit qu'il y a quelques exceptions, avec des élèves qui sont dans de meilleures conditions et d'autres dans de pires conditions.



Écart-type de l'ips des collèges

On observe que l'écart-type de l'IPS se trouvent en moyenne autour de 30 ce qui veut dire qu'il s'écarte en moyenne de 30 autour de la moyenne des IPS des collèges.

b. Matrice de covariance

(a) Démarche

Dans cette partie, on calcule la matrice de covariance afin d'obtenir pour chaque variable, ses corrélations avec les autres variables et donc de repérer avec quelles variables une variable est corrélée.

(b) Matrice de covariance

Nous avons obtenu la matrice de covariance suivante :

	0	1	2	3	4	5
0	1.00015	-0.0152837	0.190011	0.902055	0.236893	0.162757
1	-0.0152837	1.00015	0.0528728	0.322318	0.614642	0.0316234
2	0.190011	0.0528728	1.00015	0.190141	0.0777967	0.102793
3	0.902055	0.322318	0.190141	1.00015	0.504664	0.155969
4	0.236893	0.614642	0.0777967	0.504664	1.00015	0.263031
5	0.162757	0.0316234	0.102793	0.155969	0.263031	1.00015

La variable 0 correspond au nombre de candidats général, la variable 1 au taux de réussite, la 2 au taux d'accès, la 3 au nombre de mentions obtenues, la 4 à l'ips et la 5 à l'écart-type de l'ips.
Cette matrice nous montre par exemple que le nombre de candidats est très corrélé avec le nombre de mentions obtenues ou encore, que le nombre de mentions obtenues est moyennement corrélé avec l'ips et pour finir, que l'écart-type de l'ips n'est pas du tout corrélé avec le taux de réussite.

4) Régression linéaire multiple

(a) Utilisation de la Régression linéaire multiple : comment ?

En choisissant la 2eme variable statistique comme variable endogène et certaines des autres variables comme variables explicatives, la régression linéaire multiple nous permettrait d'obtenir une estimation du taux de réussite en fonction d'autres informations sur ces collèges.

(b) Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible le taux de réussite des collèges, qui se trouve dans la colonne 1 de collegesAR.
La colonne 1 de la matrice de covariance donne les coefficients de corrélation du taux de réussite avec chacune des autres variables de collegesAR.
Nous allons choisir comme variables explicatives celles qui ont le coefficient de corrélation le plus grand (en valeur absolue) avec le taux de réussite.
Les coefficients de corrélation les plus grands en valeur absolue dans la colonne 1 de matr-cov sont : 0.615, 0.322 (arrondis au millième). Ils correspondent aux variables numéro 3 et 4.

Les colonnes 3 et 4 de collegesAR correspondent aux :

- nombre de mentions obtenues,
- l'ips.

On choisit donc ces 2 variables comme variables explicatives.

(c) Lien avec la problématique

Les paramètres de la régression linéaire multiple nous informeront des variables explicatives qui influencent le plus le taux de réussite. En calculant le coefficient de corrélation multiple, on saura de plus si cette influence permet de prédire la réalité, on saura ainsi ce qui influence réellement le taux de réussite.

(d) Régression Linéaire Multiple en Python

On fait maintenant la régression linéaire multiple avec Python :

```
#variable endogène de la régression linéaire
Y = np.array(collegesAR_CR[:,1])

print(Y)

#variables explicatives de la régression linéaire
X = np.array(collegesAR_CR[:, [3,4]])

print(X)

#calcul du coefficient de corrélation multiple
linear_regression = LinearRegression()
linear_regression.fit(X, Y)
coeff = linear_regression.coef_
```

(e) Paramètres, interprétation

On obtient les paramètres a0 = 0.016, a1 = 0.61
Le signe du paramètre a0 est faible, ce qui nous permet de voir que le nombre de mentions obtenues par collège influe peu sur le taux de réussite. On en déduit donc que seuls les élèves qui ont des mentions dans le collège réussissent.

Celui du paramètre a1 est moyen, ce qui nous permet de voir que l'ips des collèges influe moyennement sur le taux de réussite de leurs élèves. On en déduit donc qu'un élève qui va dans un collège avec un indice de position sociale élevée à plus de chance de réussir dans la vie qu'un élève qui va dans un collège avec un ips faible.

Comme les variables endogène et explicatives sont centrées-réduites, on peut de plus voir que le coefficient de corrélation multiple sera faible.

(f) Coefficient de corrélation multiple, interprétation

Notre Coefficient de corrélation multiple est d'environ 0.61. Les mentions obtenues par collège et l'indice de position sociale influencent donc faiblement le taux de réussite lorsqu'elles sont réunies.

Nous pouvons donc en déduire qu'un élève dans un collège avec un fort indice de position sociale où il y a beaucoup de mentions en tout n'a pas beaucoup plus de chance de réussir qu'un élève qui va dans un autre collège.

5) Conclusion

(a) Réponse à la problématique

La problématique que nous avions posée est : Dans les différents collèges, est ce que le taux de réussite général peut être affecté par les autres données de notre fichier ?

Nous avons vu que le nombre de candidats général, le taux d'accès et l'écart-type de l'ips avaient très peu d'influence sur le taux de réussite au collège avec la matrice de covariance.

Nous avons donc effectué une régression linéaire multiple avec comme variables explicatives le nombre de mentions totales par collège et l'indice de position sociale des élèves.

Grâce à la régression linéaire multiple nous avons constaté que le taux de réussite général était peu affecté par les autres données de notre fichier.

(b) Argumentation à partir des résultats de la régression linéaire

En effet, les résultats de la régression linéaire nous ont montré que le nombre total de mention par collège influençait peu le taux de réussite par collège et nous avons vu que l'indice de position sociale d'un collège influençait moyennement le taux de réussite de ses élèves.

(c) Interprétations personnelles

Nous en avons donc déduit que les élèves qui allaient dans des collèges avec des ips élevés et qui ont un grand nombre de mentions en tout ont des chances moyennes de réussir par rapport aux élèves qui vont dans d'autres collèges.