Analyse de données : Rapport d'activité

Clara Borie-Bioulès
21511823
Analyse de données
Parcours Intermédiaires
Master Géographie, Aménagement, Environnement et Développement
GéoSuds

Choix du parcours et plan du dossier.....1

Séance 4 - Chapitre 3 : Les distributions statistiques..... 1

Questions de cours..... 1

Mise en œuvre avec Python..... 2

Séance 5 - Chapitre 4 : Les statistiques inférentielles..... 8

Questions de cours..... 8

Mise en œuvre avec Python.....11

1. Théorie de l'échantillonnage..... 11

2. Théorie de l'estimation.....12

3. Théorie de la décision..... 14

2.3 Bonus.....17

Séance 6 - Chapitre 5 : La statistique d'ordre des variables.....18

qualitatives..... 18

Questions de cours..... 18

Mise en œuvre avec Python..... 19

Bonus.....23

Séance 7 - Chapitre 6 : Régression et corrélation statistique de deux variables..... 25

Questions de cours..... 25

Mise en œuvre avec Python..... 27

Bonus.....31

Séance 8 - Chapitre 7 : Étude de deux variables qualitatives..... 35

Questions de cours..... 35

Mise en œuvre avec Python..... 36

Bonus.....38

Choix du parcours et plan du dossier

Pour ce dossier, j'ai choisi de réaliser le parcours intermédiaire, à la suite de nombreux cours de statistiques univariées et bivariées en licence de science politique. Il s'agit cependant de ma première utilisation de Python, ayant utilisé SPSS Statistics et R Studio par le passé. Il s'agit ainsi d'une découverte de ce langage informatique. Utilisant un MacBook, l'exécution du code Python s'est faite sur Visual Studio Code.

Ce dossier se compose de tous les exercices des séances 4, 5, 6, 7 et 8. Chaque section contient les questions de cours suivi par la mise en œuvre avec Python, suivi d'un court paragraphe explicatif des résultats obtenus. Les problèmes d'utilisation ou de mise en œuvre sont expliqués au début de la mise en œuvre. Toutes les consignes sont rédigées en italiques, et la rédaction des questions de cours et des exercices en texte droit. Tous les fichiers de code Python ayant permis de répondre aux questions sont disponibles sur mon Github, classés par sessions.

Séance 4 - Chapitre 3 : Les distributions statistiques.

Questions de cours

1. *Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues?*

Afin de choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues, il est important de prendre en compte la nature des données à étudier, ainsi que la nature de l'étude dans laquelle une distribution statistique est utilisée. Les distributions statistiques avec des variables discrètes sont à favoriser pour des données dénombrables, indiquant une valeur entière. Celles-ci permettent de calculer des fréquences, ou la répétition de certaines valeurs entières et quantités dénombrables. Les distributions statistiques avec des variables continues sont, à l'inverse, à favoriser pour des données d'intervalles, soit décimales, s'inscrivant sur une échelle pouvant être infinie, comme la température d'une zone donnée. Les distributions continues sont plus communément utilisées pour étudier des valeurs données dans un intervalle, pouvant être infini.

2. *Expliquez selon vous quelles sont les lois les plus utilisées en géographie?*

Au vu des critères de chaque distribution statistique et des exigences de la discipline de la géographie, les lois les plus utilisées en géographie sont :

- La loi normale : cette distribution statistique, au-delà du fait qu'elle soit l'une des plus utilisées, utilise directement une variable X prenant en compte des facteurs extérieurs et indépendants qui interviennent sur la variable donnée. En considérant le théorème

central limite, soit la somme de ces nombreux petits effets extérieurs, la loi normale permet une meilleure représentation des diverses facteurs intervenant dans un phénomène physique, spatial ou social.

- La loi Poisson : cette distribution permet de visualiser la fréquence d'éléments rares dans une succession d'épreuves très nombreuses, comme le nombre d'accidents d'avion par an, ou le nombre de séismes par décennie et par continent. En géographie, cette loi permet ainsi de visualiser la fréquence de ces éléments "anormaux" dans l'espace et le temps, très pratique en cas d'étude d'un phénomène particulier.

- La loi uniforme continue, celle-ci montrant la répartition d'une variable, permettant de simuler des échantillonnages sur celle-ci. En géographie, cette distribution statistique permet de simuler des positions ou des modèles.

- La loi log-normale, qui s'adapte à une variable X affectée par des processus multiplicatifs, créant un effet de croissance multiplicative. En géographie, cette distribution statistique permet de visualiser des distributions, comme celle du revenu, dans un espace donné.

- La loi Zipf-Mandelbrot, qui permet de visualiser la fréquence d'une variable inversement proportionnelle à son rang, permettant d'étudier le rapport rang-taille du nombre d'habitants dans une ville. Cette loi permet donc d'établir des systèmes hiérarchiques urbains.

Mise en œuvre avec Python

1. *Utiliser des méthodes scipy.stats ou écrire une fonction (informatique) qui vous permettra de visualiser :*

— les distributions statistiques de variables discrètes suivantes :

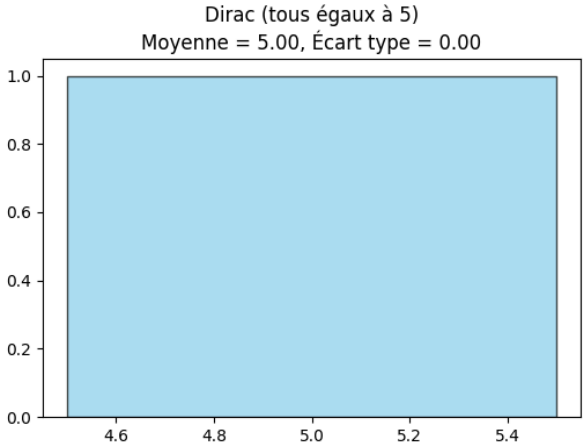
- *la loi de Dirac;*
- *la loi uniforme discrète;*
- *la loi binomiale ;*
- *la loi de Poisson ;*
- *la loi de Zipf-Mandelbrot.*

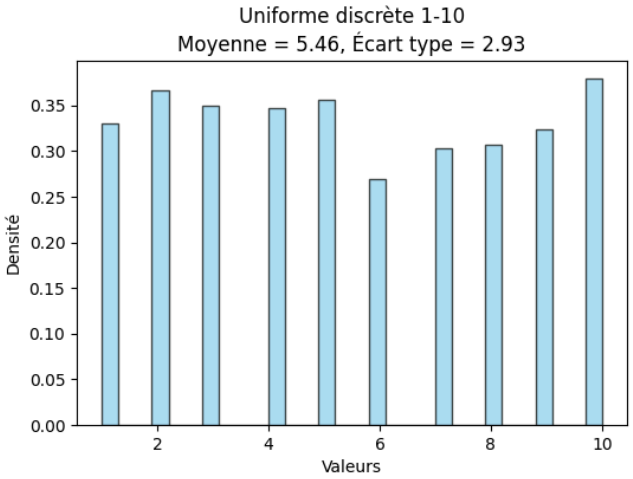
— les distributions statistiques de variables continues suivantes :

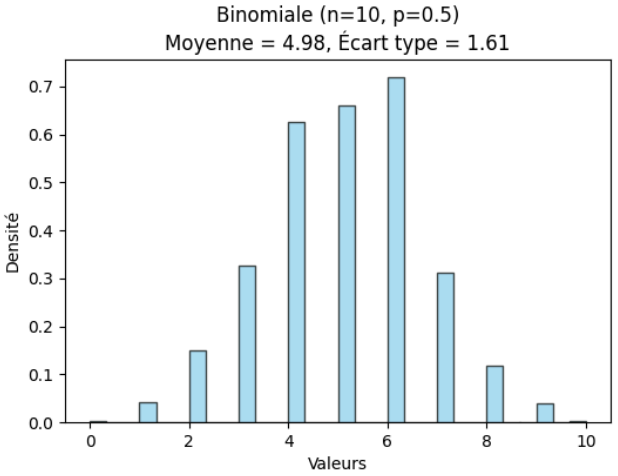
- *la loi de Poisson ;*
- *la loi normale ;*
- *la loi log-normale;*
- *la loi uniforme;*
- *la loi du $\chi 2$;*
- *la loi de Pareto.*

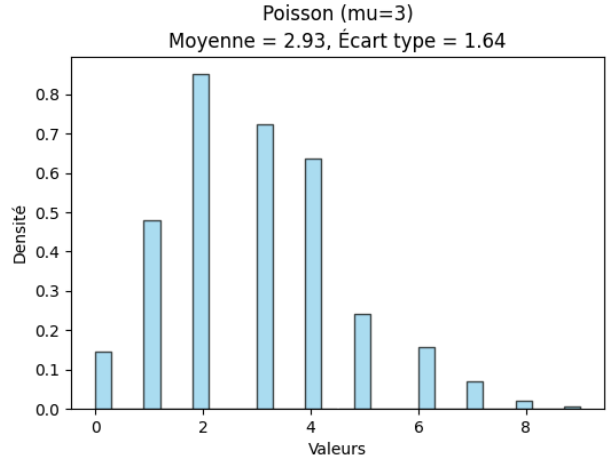
Le code utilisé pour visualiser les distributions statistiques de variables discrètes demandées est disponible sur le Github sous le nom main.py Séance 4.

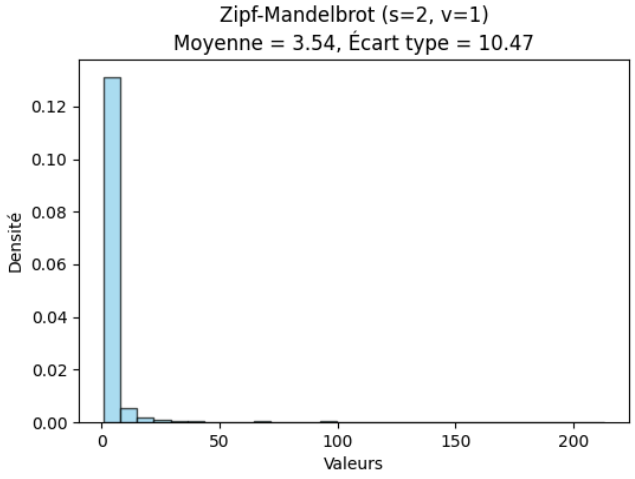
Les résultats obtenus après exécution du code se trouvent ci-dessous. En exécutant le code proposé, des graphiques représentant les différentes distributions discrètes demandées sont produits par python. Ceux-ci ont cependant été produits un à un, puisque la commande python rentrée dans Visual Studio Code ne demande pas explicitement de produire tous les graphiques en même temps :

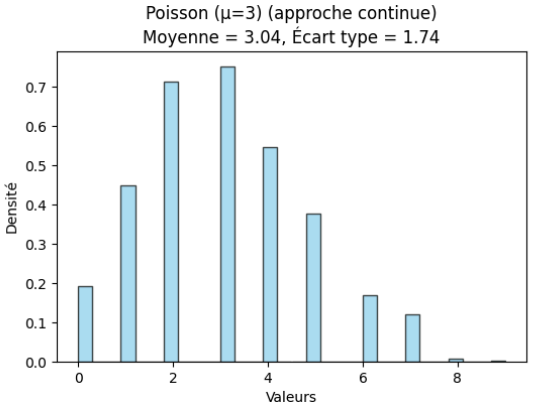


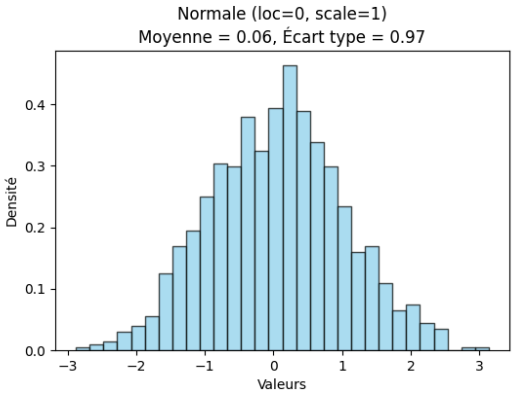


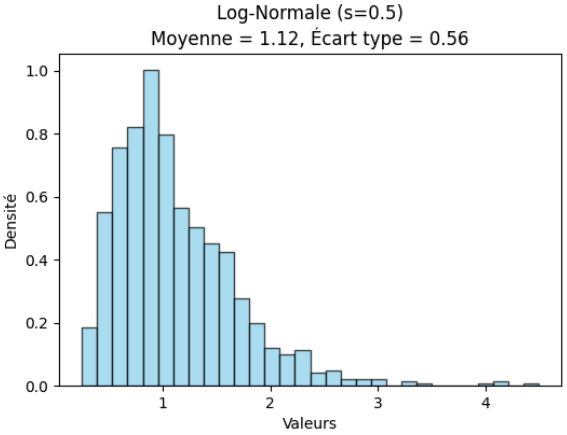

Binomiale (n=10, p=0.5)
Moyenne = 4.98, Écart type = 1.61
Densité
Valeurs
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10

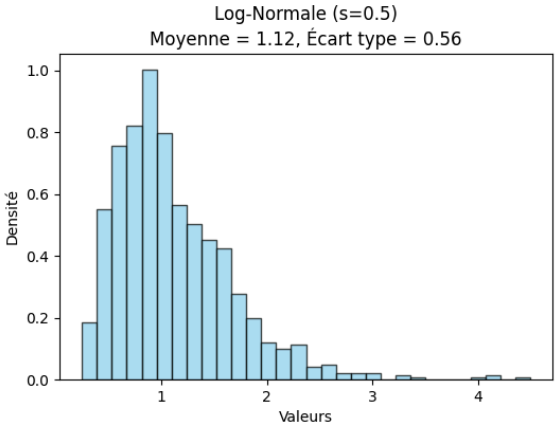

Poisson (mu=3)
Moyenne = 2.93, Écart type = 1.64
Densité
Valeurs
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
2
4
6
8

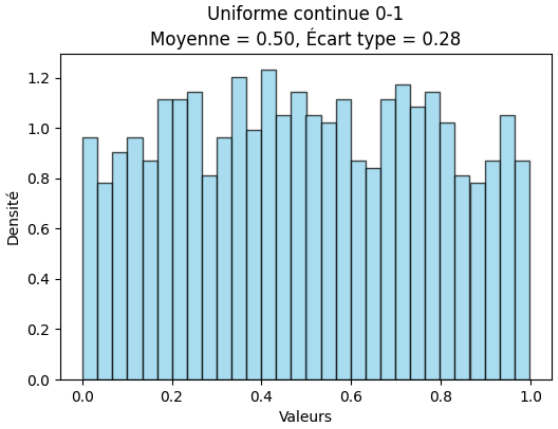

Zipf-Mandelbrot (s=2, v=1)
Moyenne = 3.54, Écart type = 10.47
Densité
Valeurs
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
50
100
150
200

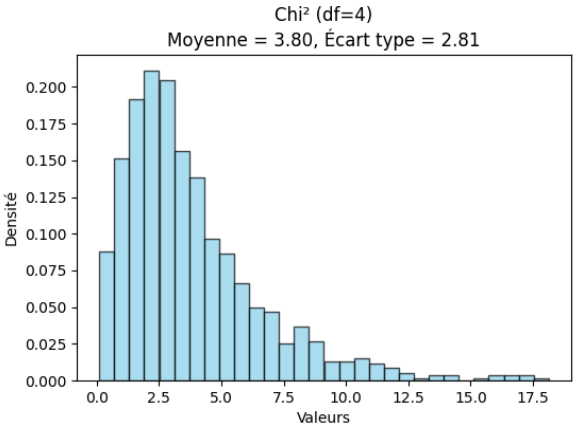

Poisson (μ=3) (approche continue)
Moyenne = 3.04, Écart type = 1.74
Densité
Valeurs

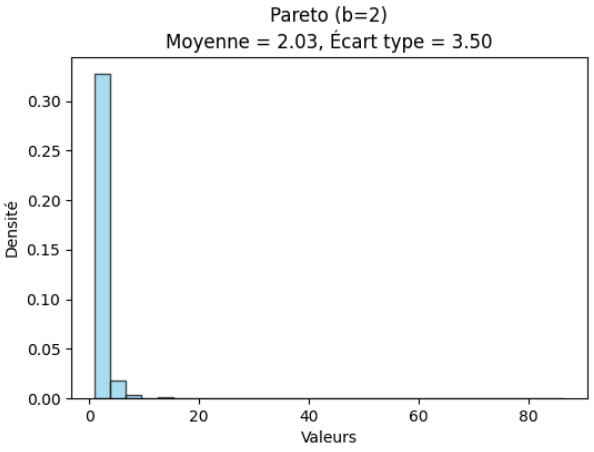

Normale (loc=0, scale=1)
Moyenne = 0.06, Écart type = 0.97
Densité
Valeurs


Log-Normale (s=0.5)
Moyenne = 1.12, Écart type = 0.56
Densité
Valeurs


Log-Normale (s=0.5)
Moyenne = 1.12, Écart type = 0.56
Densité
Valeurs


Uniforme continue 0-1
Moyenne = 0.50, Écart type = 0.28
Densité
Valeurs


Chi² (df=4)
Moyenne = 3.80, Écart type = 2.81
Densité
Valeurs


Pareto (b=2)
Moyenne = 2.03, Écart type = 3.50
Densité
Valeurs

2. Faire des fonctions (informatiques) pour calculer la moyenne et l'écart type des distributions précédentes.

```
def moyenne(data):
    return np.mean(data)

def ecart_type(data):
    return np.std(data)
```

En ajoutant cette commande dans mon code python, il est directement demandé à la machine de calculer les moyennes et écarts types de chaque distribution, affichant ainsi les moyennes et écarts types respectifs sur chaque graphique.

Séance 5 - Chapitre 4 : Les statistiques inférentielles.

Questions de cours

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier? Quelles sont les méthodes d'échantillonnage? Comment les choisir ?

L'échantillonnage est un sous-ensemble d'une population mère, faisant référence à l'entièreté de la population dont l'échantillon est extrait. Il est préférable d'utiliser des échantillons plutôt que la population entière, qui serait trop large pour être étudiée dans son ensemble. Devant être représentatif de la population mère, soit pouvant généraliser ses paramètres à l'entièreté de la population mère, l'échantillon doit être aléatoire et pertinent pour l'objet de recherche. Plusieurs méthodes d'échantillonnage existent :

- Les méthodes aléatoires, utilisant un tirage au sort depuis une base de sondage donnée. Les individus sélectionnés, classés de 1 à N, constituent l'échantillon. En fonction du type de tirage aléatoire utilisé, les individus peuvent être sélectionnés à nouveau (tirage avec remise) ou non (tirage sans remise). Avec la méthode aléatoire, chaque observation dans un échantillon à la même distribution de probabilité que la population mère. Ces méthodes, non biaisées, sont à choisir dans le cadre d'études où la répétition des tirages n'implique pas de conséquences importantes.

- Les méthodes non aléatoires, permettant de créer un échantillon dit représentatif, un "modèle réduit", de la population mère, en utilisant des procédés de tirage sélectifs. Ces méthodes non aléatoires peuvent faire appel à un échantillonnage systématique, soit la sélection d'individus dans une base de sondage en fonction d'une intervalle prédéfinie, ou encore à la méthode des quotas, qui vient à respecter les proportions de certaines caractéristiques distinctives (âge, genre, nationalité etc.) de la population mère. Les modèles aléatoires sont à choisir dans le cadre d'études visant à ce que l'échantillon représente les exactes mêmes éléments distinctifs que la population mère.

2. Comment définir un estimateur et une estimation?

L'estimation est le processus d'inférence statistique permettant de définir des estimations fiables quant aux caractéristiques de la population mère à partir de l'échantillon. Un estimateur est défini par rapport à une variable aléatoire, dont la valeur est proche de la vraie valeur du paramètre étudié. Il s'agit d'une fonction des observables permettant d'estimer les caractéristiques de la population mère, en fonction des données obtenues à travers l'échantillon.

3. Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?

Un intervalle de fluctuation suppose la connaissance de la vraie proportion théorique de la population. Il établit un intervalle au seuil de 95% de probabilité, dans laquelle la valeur théorique p est connue ou attendue. À l'inverse, un intervalle de confiance est la marge d'erreur d'un échantillon, dans le cas où la valeur de p n'est pas connue. Celui-ci permet ainsi de construire une marge permettant à l'estimation de l'échantillon d'être fausse, c'est-à-dire l'estimation de caractéristiques de la population mère en fonction de l'échantillon.

4. Qu'est-ce qu'un biais dans la théorie de l'estimation?

Dans la théorie de l'estimation, un biais fait référence à la différence entre l'espérance de l'estimateur d'un paramètre, et la valeur à estimer de ce même paramètre dans la population. Si un estimateur est biaisé, celui-ci crée alors une "erreur systématique", où la variance de l'estimateur dépend de son espérance mathématique plutôt qu'autour de la valeur du paramètre à étudier.

5. Comment appelle-t-on une statistique travaillant sur la population totale? Faites le lien avec la notion de données massives ?

Une statistique travaillant sur la population totale s'appelle une enquête exhaustive. Elle est le contraire d'un sondage,qui se base sur un échantillon représentatif de la population mère. De façon similaire, les données massives, soit des données particulières par leur large volume et variété, nécessitent des méthodes d'analyse et des technologies particulières afin de pouvoir être traitées (cf. *data centers*). Ainsi, le traitement d'une enquête exhaustive est similaire au traitement des données massives, puisque celle-ci impliquerait un très large volume et une très forte variété de données à analyser.

6. Quels sont les enjeux autour du choix d'un estimateur ?

Le choix d'un estimateur se justifie par les facteurs du biais, de l'efficacité, ainsi que de la convergence de celui-ci. Le biais de l'estimateur indique la similitude entre l'estimation et le vrai paramètre à analyser : il est préférable que l'espérance de l'estimation soit égale à celle du paramètre analysé, rendant l'estimateur dit "sans biais". L'efficacité d'un estimateur, soit la valeur de sa variance, est également préférable lorsqu'elle est la plus petite possible, indiquant des variations moins importantes au sein de l'estimateur. Enfin, la convergence d'un estimateur, soit la convergence de la distribution autour de la valeur du paramètre étudié lorsque l'échantillon est agrandi vers l'infini, indique également un estimateur plus fiable et plus représentatif des propriétés du caractère étudié.

7. Quelles sont les méthodes d'estimation d'un paramètre? Comment en sélectionner une ?

Il existe plusieurs méthodes afin d'estimer un paramètre. La méthode la plus simple, celle des moindres carrés, minimise la somme des déviations de l'estimation d'un paramètre, notamment les valeurs ajustées et les résidus. Cependant, la méthode du maximum de vraisemblance (M.V.) est préférable puisque plus générale : la vraisemblance s'apparente au tri des différentes valeurs du paramètre selon leur probabilité, fondé sur des observations et le paramètre. La méthode du maximum de vraisemblance maximise la vraisemblance du paramètre étudié, afin d'inférer les paramètres de probabilité de l'échantillon donné.

8. Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?

Il existe plusieurs tests en statistique inférentielle, comme la théorie des tests, formulant des hypothèses sur les paramètres ou les lois intervenant dans les problèmes étudiés. Ainsi, la théorie de la décision se construit autour des différents tests statistiques suivant :

- Les tests paramétriques, soit des tests où la forme des distributions testées est supposément connue, et portant sur un paramètre précis. Lors des tests paramétriques, une hypothèse_0 est supposée sur la loi des données. Les tests paramétriques relatent à des paramètres comme la moyenne, l'écart type ou le type de distribution utilisé.

- Les tests non paramétriques, où la forme de la distribution n'est pas considérée, et s'appliquant à des variables pouvant être qualitatives et quantitatives. Ces tests non paramétriques relatent des paramètres comme l'effectif ou la médiane.

- Les tests robustes, composés des tests libres, valables qu'importe la loi de la variable étudiée. Ces tests libres sont utilisés pour mener des tests lorsque la loi de la variable aléatoire X n'est pas connue.

- Les tests de signification, où la valeur observée du score lors d'une expérience quantifie l'écart entre la distribution des observations et l'hypothèse nulle formulée avant le test.

- Les tests d'hypothèse, servant à vérifier si les données présentes dans l'échantillon sont compatibles avec l'hypothèse formée sur la population mère. Les tests d'hypothèse servent à rejeter ou non l'hypothèse étudiée, celle-ci étant étudiée comme valeur de référence lors du test.

- Les tests d'ajustement, permettant d'évaluer la cohérence entre une situation réelle donnée et un modèle théorique. Ces tests permettant d'évaluer si l'hypothèse d'une loi de probabilité est en adéquation avec la réalisation d'un échantillon d'une variable X.

- Les tests de comparaison, utilisés pour comparer des échantillons entre eux.

- Les tests d'indépendance, intervenant lorsque plusieurs variables aléatoires sont à prendre en question dans le même test.

9. Que pensez-vous des critiques de la statistique inférentielle ?

Les critiques de la statistique inférentielle estiment que certaines méthodes utilisées en statistique inférentielle auraient tendances à tirer des conclusions hâtives quant à l'adéquation entre l'échantillon testé et la population mère, ou le rejet de l'hypothèse 0. Bien que certaines de ces méthodes soient discutables, leur utilité reste significativement importante : la statistique inférentielle permet de visualiser un phénomène de grande échelle à petite échelle, tout en prenant les mesures nécessaires pour pallier la marge d'erreur de chaque test.

Mise en œuvre avec Python

1. Théorie de l'échantillonnage

***Problème à résoudre**. Vous réalisez une enquête d'opinion. On suppose que la population mère est composée de 2 185 individus. Si vous leur posez une question quelconque, nécessitant une opinion tranchante de type : « Pour », « Contre » ou « Sans opinion ». On suppose que, pour la population mère, il existe 852 personnes « Pour », 911 personnes « Contre » et 422 personnes « Sans opinion ».*

***Exercice simulant une situation réelle.** Vous n'avez pas accès à la totalité de l'information précédente. Ici, on parle de 2 185 individus, mais imaginez que votre population soit celle de la France entière ou du monde. Vous n'avez pas accès à ce que pense tout le monde. Il faut de fait l'échantillonner. Avec des outils de simulation, il est facile de générer 100 échantillons. aléatoires.*

— Dans le dossier src, introduire le dossier data le fichier Echantillonnage-100-Echantillons.csv disponible dans la Seance-05 du GitHub

— Ouvrir le fichier en utilisant la fonction locale ouvrirUnFichier(). Elle prend en paramètre une chaîne de caractères matérialisant l'adresse et le nom du fichier. Le fichier contient le résultat d'un tirage au hasard sans remise dans la population initiale de 100 échantillons. Sur chaque ligne, il est compté le nombre d'individus ayant telle ou telle opinion.

— Pour chaque colonne, calculer la moyenne obtenue. Le nombre de personnes devant être entier, il faut arrondir votre calcul avec aucune décimale en respectant la règle de l'arrondi avec la fonction native round().

— Pour comparer l'échantillon avec sa population mère, il faut calculer les fréquences. Pour ce, calculer la somme des trois moyennes obtenues, puis diviser ce résultat à l'ensemble de vos moyennes. Calculer également suivant le même principe les fréquences de la population mère afin de comparer les deux résultats. Arrondir les fréquences obtenues à deux décimales.

2— Vous devez constater un écart entre les valeurs observées en moyennant les échantillons et les valeurs réelles de la population mère. Calculer l'intervalle de fluctuation de chacune des fréquences à un seuil de 95 %, soit zC = 1, 96.
— Expliquer le lien entre l'intervalle de fluctuation et les valeurs réelles de la population mère. Que pouvez-vous en conclure par rapport aux échantillons utilisés pour le calcul?

Lors de cet exercice, j'ai eu des difficultés à importer le dossier CSV sur Visual Studio Code. J'ai utilisé la commande curl et le "raw file" Github dans le Terminal :

curl -L -o data/Echantillonnage-100-Echantillons.csv "https://raw.githubusercontent.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/refs/heads/main/Seance-05/Exercice/src/data/Echantillonnage-100-Echantillons.csv"

Ayant eu également du mal à créer un container Docker pour cette étape, j'ai codé dans un environnement personnel créé avec Python : (.venv).

Le code utilisé dans le src afin de commander à python de réaliser tous les calculs demandés dans l'exercice est le suivant disponible sur le Github sous le nom main.py Séance 5. Les résultats obtenus après avoir exécuté la commande:
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 src/main.py

Moyennes par opinion : [**391**, **416**, **193**]
Fréquences de l'échantillon : [**0.39**, **0.42**, **0.19**]
Fréquences de la population mère : [**0.3**, **0.5**, **0.2**]
Intervalles de fluctuation à 95% : [(**0.27**, **0.33**), (**0.47**, **0.53**), (**0.18**, **0.22**)]

Conclusion : Les fréquences de l'échantillon doivent se situer dans l'intervalle de fluctuation des fréquences de la population mère. Des écarts sont possibles à cause de la variabilité des échantillons.

2. Théorie de l'estimation

Problème à résoudre. *Dans le cas d'espèce, et c'est un des cas les plus fréquents en sciences humaines et sociales, on ne possède qu'un échantillon de la population mère. Comment avoir confiance en lui? Ici, la méthode consiste à construire des intervalles de confiance.*
— Prendre le premier échantillon de la liste précédente en utilisant la méthode Pandas iloc(0), le paramètre 0 correspondant à votre première ligne de données. Il faudra utiliser des fonctions natives, donc convertissez l'objet Pandas en castant une list().
— Calculer la somme de la ligne, puis comme avec le cas précédent, les fréquences en utilisant l'effectif total de l'échantillon isolé.
— L'intervalle de confiance ne dépend que de la taille de l'échantillon, c'est-à-dire la somme précédente. Calculez cet intervalle pour chaque opinion.
— Dans votre rapport, vous interpréterez le résultat obtenu et vous le comparerez avec

le résultat précédent.

Pour cette étape de l'exercice, la théorie de l'estimation, j'ai ajouté les commandes demandées dans la consigne au code src de l'étape précédente. Ainsi, le code src ajouté au code précédent (afin de visualiser les différences plus facilement) et exécuté avec la commande python3 src/main.py pour cette étape est disponible sur le Github sous le nom main.py Séance 5.

Le résultat de l'exécution de ce code est la suivante :

(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 src/main.py
Moyennes par opinion : [**391**, **416**, **193**]
Fréquences de l'échantillon : [**0.39**, **0.42**, **0.19**]
Fréquences de la population mère : [**0.3**, **0.5**, **0.2**]
Intervalles de fluctuation à 95% : [(**0.27**, **0.33**), (**0.47**, **0.53**), (**0.18**, **0.22**)]

Conclusion : Les fréquences de l'échantillon doivent se situer dans l'intervalle de fluctuation des fréquences de la population mère. Des écarts sont possibles à cause de la variabilité des échantillons.

	Pour	**Contre**	**Sans opinion**
0	395	396	209
1	379	432	189
2	384	426	190
3	395	407	198
4	389	413	198

Premier échantillon : [**395**, **396**, **209**]
Fréquences du premier échantillon : [**0.395**, **0.396**, **0.209**]
Intervalle de confiance à 95% : [(**0.36**, **0.43**), (**0.37**, **0.43**), (**0.18**, **0.23**)]

Les résultats obtenus avec le premier échantillon en comparaison avec les premiers résultats permettent de déduire si le premier échantillon est représentatif de la population mère. Dans les premiers résultats produits, la moyenne de tous les échantillons, [**0.39**, **0.42**, **0.19**], apparaît comme légèrement différente de la population mère, [**0.3**, **0.5**, **0.2**]. En s'intéressant au premier échantillon seulement, les intervalles de confiance produits par python, [(**0.36**, **0.43**), (**0.37**, **0.43**), (**0.18**, **0.23**)] sont globalement proches des fréquences de la population mère. Ainsi, le premier échantillon permet de fournir une estimation fiable de la population mère.

3. Théorie de la décision

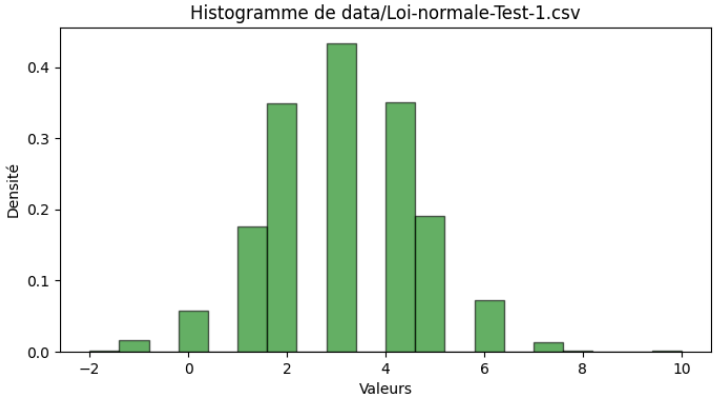
Problème à résoudre. *Intervalles de fluctuation ou intervalles de confiance peuvent être insuffisants pour valider un résultat. En fonction de la nature des variables (quantitatives ou qualitatives), les statisticiens ont inventé et inventent encore de nombreux tests statistiques. Le test de Shapiro-Wilks permet de tester si une distribution suit la loi normale.*
— Vous disposez de deux fichiers Loi-normale-Test-1.csv et Loi-normale-Test-2.csv. Il s'agit d'une série de nombres aléatoires traduisant une distribution statistique. En utilisant la méthode de scipy.stats shapiro(). Laquelle est une distribution normale ?
— Vous expliquerez dans votre rapport pourquoi.
N.B. Il vous faudra essentiellement connaître les types de tests en fonction des situations.

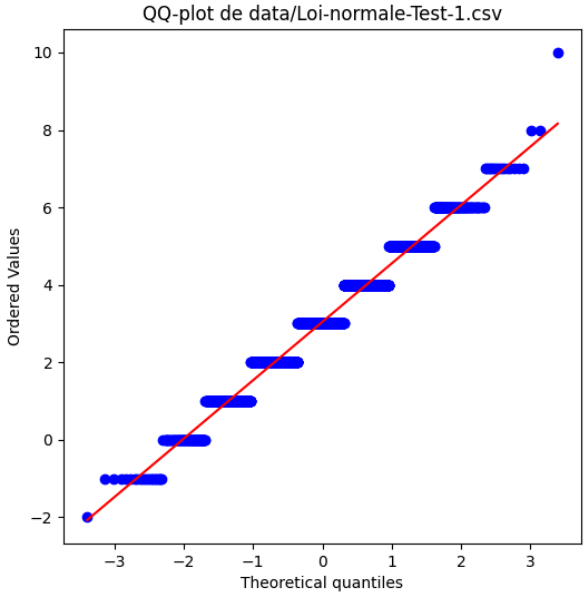
Pour cette dernière étape, j'ai fait face au même problème que lors de la première dans l'importation des données dans Visual Studio Code. Ainsi, j'ai fait appel à la même commande curl et le "raw file" de Github dans le Terminal :

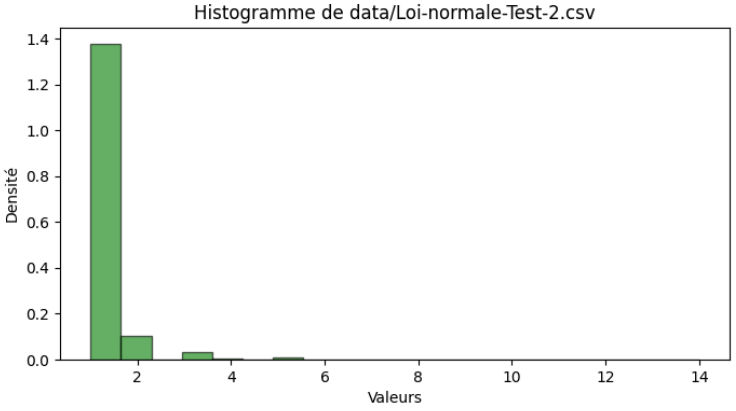
```
claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % source .venv/bin/ac
tivate
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % curl -L -o
data/Loi-normale-Test-1.csv
"https://raw.githubusercontent.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/refs/h
eads/main/Seance-05/Exercice/src/data/Loi-normale-Test-1.csv"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  4029  100  4029    0     0  11569      0 --:--:-- --:--:-- --:--:-- 11644

(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % curl -L -o
data/Loi-normale-Test-2.csv
"https://raw.githubusercontent.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/refs/h
eads/main/Seance-05/Exercice/src/data/Loi-normale-Test-2.csv"

  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  4009  100  4009    0     0  19625      0 --:--:-- --:--:-- --:--:-- 19556
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % ls -la data

total 24
drwxr-xr-x   5 claraboriebioules  staff   160 Dec  4 15:14 .
drwxr-xr-x@ 11 claraboriebioules  staff   352 Dec  4 12:37 ..
-rw-r--r--   1 claraboriebioules  staff  1231 Dec  4 12:46
Echantillonnage-100-Echantillons.csv
-rw-r--r--   1 claraboriebioules  staff  4029 Dec  4 15:13 Loi-normale-Test-1.csv
-rw-r--r--   1 claraboriebioules  staff  4009 Dec  4 15:14 Loi-normale-Test-2.csv
```

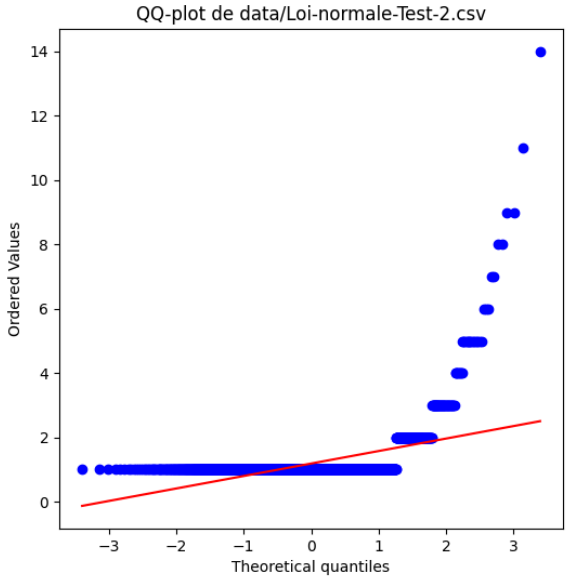
Une fois les données importées, le code python dans le src pour l'étape 3 a été ajouté à la suite du code des deux autres étapes, non répété dans le rapport puisque présenté ci-dessus, mais répété dans VSC. Étant donné le large nombre de données présentes dans les dossiers csv des test 1 et 2, j'ai exécuté une commande demandant à python de produire des histogrammes ainsi que des qq-plots pour les deux tests, avant de réaliser le test de Shapiro-Wilk. Ainsi, le code de l'étape 3 est disponible sur le Github sous le nom de main.py Séance 5.

Le résultat de l'exécution du code dans VSC est le suivant :









Test de Shapiro-Wilk pour Loi-normale-Test-1.csv:
Statistique = **0.9639**, p-value = **0.0000**
→ Loi-normale-Test-1.csv ne suit pas une loi normale.

Test de Shapiro-Wilk pour Loi-normale-Test-2.csv:
Statistique = **0.2609**, p-value = **0.0000**
→ Loi-normale-Test-2.csv ne suit pas une loi normale.

Les résultats de l'exécution du code python de l'étape 3 démontre qu'aucune des deux distributions des test1 et test2 n'est normale. En effet, comme inscrit dans le code proposé, pour que l'un des deux tests puisse suivre une loi normale, il faudrait que la p-value soit supérieure ou égale à 0.05 (choix du seuil de signification $\alpha = 0{,}05$, test au seuil de 5 %). Or, la p-value des deux tests étant de 0.0000, on rejette les hypothèses nulles, selon lesquelles "les données proviennent d'une loi normale". Une p-value si faible indique la très faible propriété d'obtenir un tel échantillon si la loi était distribuée normalement. On conclut ainsi que les test1 et test2 ne suivent pas une loi normale.

Cependant, la série du test1 semble suivre une distribution normale sur l'histogramme et le qq-plot. La forme de cloche de l'histogramme porte à confusion, donnant une impression visuelle de normalité, comme le qq-plot qui semble relativement aligné. Cependant, ce sont les petites déviations qui impactent la distribution, ce que le test de Shapiro-Wilk détecte.

2.3 Bonus

Lors du test de la décision, l'une des lois n'est pas normale. En vous aidant de la séance précédente, quelle est sa distribution?

Les résultats obtenus à la question précédente ne me permettent pas de répondre à cette question bonus, étant donné que les deux lois ont été prouvées comme non normales par python. Il s'agit certainement d'une erreur de ma part dans la rédaction de mon code ou la transmission des données sur Visual Studio Code, mais après plusieurs tentatives, je n'arrive pas à obtenir d'autres résultats.

Séance 6 - Chapitre 5 : La statistique d'ordre des variables qualitatives.

Questions de cours

1. *Qu'est-ce qu'une statistique ordinale? À quel autre statistique catégorielle s'oppose-t-elle ? Quel type de variables utilise-t-elle ? En quoi cela peut matérialiser une hiérarchie spatiale ?*

Une statistique ordinale est un type de statistique analysant la descente, stagnation, ou montée d'une entité donnée dans un classement, utilisant des variables ordinales, organisées de préférences dans un ordre croissant (aussi appelé "ordre naturel"). Ce type de statistique peut permettre de matérialiser une hiérarchie spatiale puisqu'elle permet de mesurer ou visualiser la fréquence ou l'évolution d'une entité dans un espace donné. Les statistiques ordinales s'opposent aux statistiques nominales, où les catégories concernées ne sont pas organisées dans un ordre particulier.

2. *Quel ordre est à privilégier dans les classifications ?*

L'ordre étant à privilégier dans les classifications est un ordre croissant, aussi appelé "ordre naturel", puisque celui-ci permet d'identifier les valeurs aberrantes dans une série d'observations donnée, et permet également de visualiser la loi de la plus grande valeur.

3. *Quelle est la différence entre une corrélation des rangs et une concordance de classements ?*

Bien que la corrélation de rangs et la concordance de *p* classement cherchent toutes les deux à identifier si les classements étudiés sont identiques, elles étudient des facteurs différents. La corrélation des rangs fait référence à la mesure de la similitude de deux classements ordinaux, et peut également servir à mesurer signification statistique entre les deux, en évaluant si les rangs sont liés. A l'inverse, une concordance de *p* classement exprime la généralisation à plus de deux classements, ainsi que la cohérence des mêmes individus entre eux.

4. *Quelle est la différence entre les tests de Spearman et de Kendall ?*

Le test de Spearman est utilisé pour étudier la corrélation des rangs, et permet de calculer si les classements étudiés sont identiques, inverses ou indépendants. Établissant deux hypothèses (le coefficient de corrélation n'est significativement pas différent de zéro, ou le coefficient de corrélation est significativement différent de zéro), le test de Spearman peut être utilisé comme outil géographique dans la classification des villes par rapport à leur population, ou de comparer plusieurs variables en lien avec un objet d'étude définit, évaluant la dépendance du classement par rapport à d'autres.

À l'inverse, le test de Kendall étudie deux classements correspondant à deux séries de valeurs distinctes, en créant des couples de rang. Cette méthode permet d'étudier si l'ordre naturel des variables est respecté ou non, soit si les classements sont considérés comme "concordants" (+1) ou "discordants" (-1). Après répétition de cette manipulation, si le coefficient Tau de Kendall vaut +1, alors les classements sont identiques. Si le coefficient Tau de Kendall vaut -1, alors les coefficients sont inverses.

Ainsi, ces deux tests permettent tous les deux de visualiser la similitude entre deux classements, mais la méthode de calcul diffère : si le test de Spearman ne calcule qu'une fois la corrélation entre les deux classements étudiés, le test de Kendall cherche à exprimer la relation en changeant la technique, soit par le comptage des paires concordants/discordantes.

5. À *quoi servent les coefficients de Goodman-Kursdal et de Yule ?*

Le coefficient Γ de Goodman-Kursdal est utilisé afin de calculer la proportion de surplus de paires concordantes par rapport aux paires discordantes, et varie entre -1 et +1.

Le coefficient *Q* de Yule suit la même logique que le coefficient Γ de Goodman-Kursdal, mais est appliqués aux tableaux de contingence 2 x 2, lui-même calculé à partir des fréquences observées. En créant une table de contingence pour évaluer la fréquence des évènements, le coefficient de Yule calcule l'association des séries ordinales entre elles. Celui-ci varie entre -1 (l'association est négative totale) et +1 (l'association est positive parfaite).

Mise en œuvre avec Python

1. Dans le dossier src, introduire le dossier data le fichier island-index.csv disponible dans la Seance-06 du GitHub
2. Ouvrir le fichier en utilisant la fonction locale ouvrirUnFichier(). Elle prend en paramètre une chaîne de caractères matérialisant l'adresse et le nom du fichier.
3. Isoler la colonne « Surface (km2) » et lui ajouter dans cette liste :
— Asie / Afrique / Europe : 85 545 323 km2
— Amérique : 37 856 841 km2
— Antarctique : 7 768 030 km2
— Australie : 7 605 049 km2
Attention ! Il faudra forcer le typage en castant les valeurs en float().
Attention ! Ne mettez pas l'unité de mesure dans la liste informatique
4. Ordonner la liste obtenue avec la fonction locale ordreDecroissant() proposée. Elle prend en paramètre une liste.
5. Visualiser la loi rang-taille en créant une image de sortie.
6. L'image obtenue est illisible. Il vous faut convertir les axes en logarithme. Pour ce, utiliser la fonction locale conversionLog() proposée. Elle prend en paramètre une liste.
7. Est-il possible de faire un test sur les rangs ? (mettre votre réponse sous la forme d'un commentaire dans le fichier)

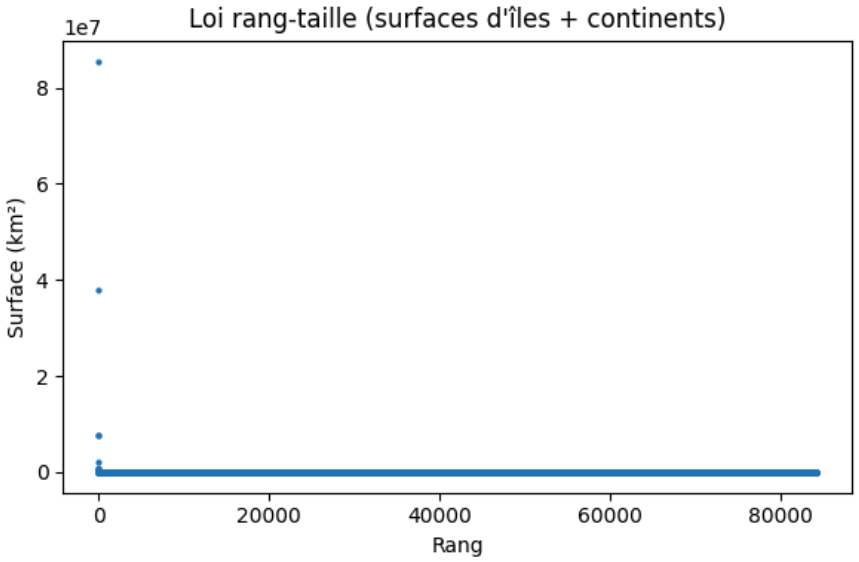
Comme lors des exercices précédents, j'ai eu du mal à créer un container Docker pour cette étape, j'ai codé dans un environnement personnel créé avec Python : (.venv). Ayant également eu des difficultés à importer le dossier comportant les données sur Visual Studio Code, j'ai de nouveau utilisé la commande curl et le "raw file" Github dans le Terminal :

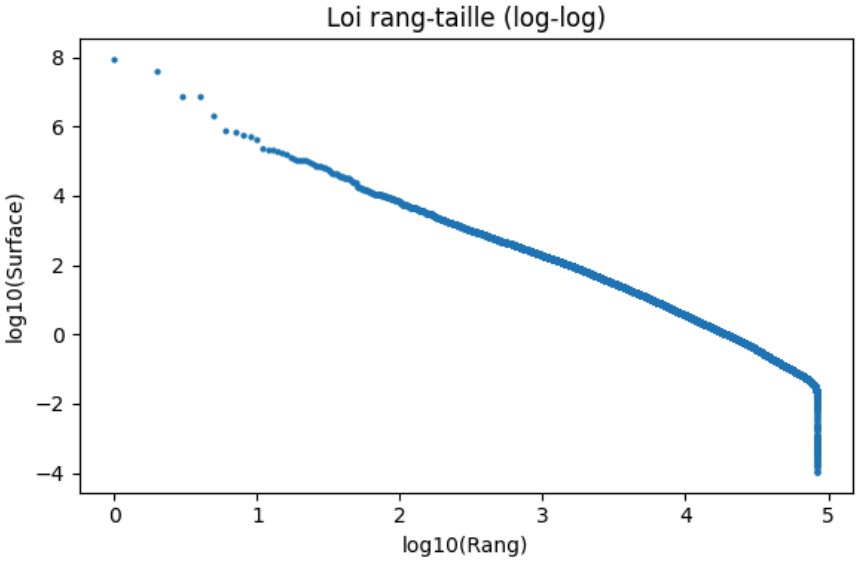
```
claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % source .venv/bin/ac
tivate
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % mkdir -p data

curl -L -o data/island-index.csv \
"https://raw.githubusercontent.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/refs/h
eads/main/Seance-06/Exercice/src/data/island-index.csv"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 9163k  100 9163k    0     0  14.9M      0 --:--:-- --:--:-- --:--:-- 14.9M
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker %
```

Pour cette étape de l'exercice, le code src que j'ai exécuté avec la commande python3 src/main.py est disponible sur Github sous le nom main.py Séance 6. Le résultat obtenu dans le terminal après l'exécution de ce script est le suivant :

```
.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3
src/main.py
Aperçu des données :
          Type          Identifiant  ...   Latitude  Longitude
ILE NORMALE     1          Aruba Island  ... -69.970276  12.509315
ILE NORMALE     3  Baia dos Tigres Island  ...  11.704911 -16.595546
ILE NORMALE     5                   NaN  ...  12.299087  -6.118821
ILE NORMALE     6                   NaN  ...  12.291456  -6.155205
ILE NORMALE     7                   NaN  ...  12.341880 -14.393714

[5 rows x 10 columns]

Colonnes disponibles :
['Type', 'Identifiant', 'Toponyme', 'Code ISO 1', 'Code ISO 2', 'Code ISO 3', 'Trait de côte
(km)', 'Surface (km²)', 'Latitude', 'Longitude']

Images enregistrées à la racine du projet :
  rang_taille_lineaire.png
  rang_taille_loglog.png

Exemple de test sur les rangs (rang vs 1/rang) :
  Spearman r_s = -1.000, p-value = 0
  Kendall tau = -1.000, p-value = 0
```

Ainsi que la visualisation de la loi rang-taille, dont une version dont les axes ont été convertis en logarithmes :





Le bloc “Exemple de test sur les rangs (rang vs 1/rang)” calcule directement les tests du coefficient de Spearman et du coefficient de Kendall sur une relation purement décroissante (rang et 1/rang), ce qui illustre qu’un test sur les rangs est possible et significatif. Avec un résultat de -1 et une p-value nulle, les deux tests illustrent un lien monotone décroissant parfaitement significatif.

8. Dans le dossier src, introduire le dossier data le fichier Le-Monde-HS-Etats-du-monde-2007-2025.csv disponible dans la Seance-06 du GitHub
9. Ouvrir le fichier en utilisant la fonction locale ouvrirUnFichier(). Elle prend en paramètre une chaîne de caractères matérialisant l’adresse et le nom du fichier.

10. Isoler les colonnes « État », « Pop 2007 », « Pop 2025 », « Densité 2007 » et « Densité 2025 »
11. En utilisant la fonction locale ordrePopulation(), ordonner de manière décroissante les listes « Pop 2007 », « Pop 2025 », « Densité 2007 » et « Densité 2025 ». Elle prend en paramètres la liste à ordonner et la liste des États.
12. En utilisant la fonction locale classementPays(), préparer la comparaison des listes sur la population et la densité. Elle prend en paramètres les deux classements (pour le nombre d'habitants et pour la densité) obtenus avec ordrePopulation(). En classant le résultat avec la méthode sort(), vous obtenez une liste avec deux colonnes ordonnées par rapport au classement de 2007.
13. Isoler les deux colonnes sous la forme de liste différents en utilisant une boucle
14. Dans la bibliothèque scipy.stat, utiliser les méthodes spearmanr() et kendalltau() pour calculer le coefficient de corrélation des rangs et la concordance des rangs. Les deux méthodes prennent en paramètres les deux classements que vous avez respectivement calculés pour le nombre d'habitants et pour la densité. Vous commenterez ce résultat dans votre rapport d'activité.

Pour une question de facilité, le dossier csv demandé pour cette partie de l'exercice a également été importé avec la commande curl et le "raw file" Github dans l'espace de travail (.venv) :

(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % mkdir -p data

curl -L -o data/Le-Monde-HS-Etats-du-monde-2007-2025.csv \
"https://raw.githubusercontent.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/refs/heads/main/Seance-06/Exercice/src/data/Le-Monde-HS-Etats-du-monde-2007-2025.csv"
```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 68402  100 68402    0     0   339k      0 --:--:-- --:--:-- --:--:--  340k
```

Le résultat obtenu après exécution de la commande disponible sous le document python main.py Séance 6 est le suivant :

.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 src/main.py
Aperçu des données :

	Code ISO_3	Numéro	Continent rattaché	...	Densité 2023	Densité 2024	Densité 2025
0	afg	1	Asie	...	62.94	64.62	65.24
1	zaf	2	Afrique	...	48.20	49.43	52.37
2	alb	3	Europe	...	96.55	96.55	93.10
3	dza	4	Afrique	...	18.83	19.13	19.63
4	deu	5	Europe	...	233.61	233.33	236.69

[5 rows x 45 columns]

Colonnes disponibles :
['Code ISO_3', 'Numéro', 'Continent rattaché', 'State', 'État', 'Superficie 2007', 'Superficie 2012', 'Pop 2007', 'Pop 2008', 'Pop 2009', 'Pop 2010', 'Pop 2011', 'Pop 2012', 'Pop 2013', 'Pop 2014', 'Pop 2015', 'Pop 2016', 'Pop 2017', 'Pop 2018', 'Pop 2019', 'Pop 2020', 'Pop 2021', 'Pop 2022', 'Pop 2023', 'Pop 2024', 'Pop 2025', 'Densité 2007', 'Densité 2008', 'Densité 2009', 'Densité 2010', 'Densité 2011', 'Densité 2012', 'Densité 2013', 'Densité 2014', 'Densité 2015', 'Densité 2016', 'Densité 2017', 'Densité 2018', 'Densité 2019', 'Densité 2020', 'Densité 2021', 'Densité 2022', 'Densité 2023', 'Densité 2024', 'Densité 2025']

Année 2007 :
Spearman r_s = 0.945, p-value = 1.77e-95
Kendall tau = 0.793, p-value = 7.14e-61

Année 2025 :
Spearman r_s = -0.027, p-value = 0.708
Kendall tau = -0.008, p-value = 0.87

Les résultats des coefficients de Spearman et de Kendall pour les années 2007 et 2025 indiquent deux résultats très différents. Pour l'année 2007, les coefficients de Spearman et de Kendall ainsi que les p value indiquent une forte corrélation positive et statistiquement significative, indiquants que les pays les plus peuplés font partie des plus denses. On note ainsi une forte concordance des hiérarchies. A l'inverse, les coefficients et les p value de l'année 2025 indiquent une corrélation quasi nulle et non statistiquement significative, où le rang par densité et le rang par population ne sont plus liés de manière monotone.

Bonus

Afin de vous faciliter la vie, car les codes n'étaient pas simples à trouver, vous avez utiliser des fonctions que j'ai créées. C'est l'un des bonus qui peut vous rapporter le plus de points. Vous devrez créer des fonctions locales permettant de calculer rapidement une analyse de rangs.
Pour les îles, proposer un algorithme et un code pour comparer le classement obtenu avec les surfaces et un autre classement obtenu avec les traits de côte en utilisant le coefficient de corrélation et la concordance des rangs.
Pour la population mondiale,
1. factoriser le code d'analyse des classements sous la forme
d'une fonction renvoyant les coefficients de corrélation des rangs et la concordances des rangs ;
2. faire un algorithme et un code permettant d'analyser la concordance des rangs de l'ensemble des classements annuels de 2007 à 2025.

L’algorithme et le code créé pour cette question bonus est disponible sur Github. Le résultat obtenu dans le terminal après avoir exécuté le code est le suivant :

.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 src/main.py

=== ÎLES : comparaison du classement par surface et par trait de côte ===
Nombre d'îles utilisées : 84219
Spearman r_s = 0.971, p-value = 0
Kendall tau = 0.854, p-value = 0

=== POPULATION MONDIALE : analyse des classements 2007–2025 ===
Année 2007 : Spearman r_s = 0.945 (p=1.77e-95), Kendall tau = 0.793 (p=7.14e-61)
Année 2008 : Spearman r_s = 0.945 (p=2.77e-95), Kendall tau = 0.792 (p=8.55e-61)
Année 2009 : Spearman r_s = 0.945 (p=1.43e-95), Kendall tau = 0.793 (p=7.14e-61)
Année 2010 : Spearman r_s = 0.948 (p=9.23e-98), Kendall tau = 0.793 (p=7.14e-61)
Année 2011 : Spearman r_s = 0.947 (p=3.27e-97), Kendall tau = 0.793 (p=6.64e-61)
Année 2012 : Spearman r_s = 0.946 (p=1.08e-96), Kendall tau = 0.791 (p=1.27e-60)
Année 2013 : Spearman r_s = 0.945 (p=6.73e-96), Kendall tau = 0.788 (p=3.38e-60)
Année 2014 : Spearman r_s = 0.530 (p=1.7e-15), Kendall tau = 0.359 (p=9.1e-14)
Année 2015 : Spearman r_s = -0.051 (p=0.48), Kendall tau = -0.025 (p=0.599)
Année 2016 : Spearman r_s = -0.038 (p=0.595), Kendall tau = -0.018 (p=0.707)
Année 2017 : Spearman r_s = -0.036 (p=0.613), Kendall tau = -0.016 (p=0.746)
Année 2018 : Spearman r_s = -0.036 (p=0.622), Kendall tau = -0.015 (p=0.753)
Année 2019 : Spearman r_s = -0.038 (p=0.595), Kendall tau = -0.017 (p=0.723)
Année 2020 : Spearman r_s = -0.025 (p=0.727), Kendall tau = -0.008 (p=0.872)
Année 2021 : Spearman r_s = -0.033 (p=0.647), Kendall tau = -0.014 (p=0.776)
Année 2022 : Spearman r_s = -0.030 (p=0.682), Kendall tau = -0.011 (p=0.827)
Année 2023 : Spearman r_s = -0.030 (p=0.676), Kendall tau = -0.010 (p=0.839)
Année 2024 : Spearman r_s = -0.010 (p=0.888), Kendall tau = 0.002 (p=0.962)
Année 2025 : Spearman r_s = -0.027 (p=0.708), Kendall tau = -0.008 (p=0.87)

Séance 7 - Chapitre 6 : Régression et corrélation statistique de deux variables

Questions de cours

1. *Quel est l'intérêt de passer des statistiques univariées aux statistiques bivariées ?*

Les statistiques bivariées permettent l'analyse simultanée de deux variables, sous l'hypothèse qu'il existe une relation logique entre les deux variables étudiées. Ainsi, les statistiques bivariées quantifient et évaluent la relation entre deux variables, permettant d'analyser des facteurs ou phénomènes nécessitant plus d'une valeur pour être étudiés.

2. Quelles différences opérez-vous entre corrélation et correspondances? Qu'est-ce qu'un rapport de corrélation ?

La corrélation fait référence à la liaison existant entre les deux variables étudiées, où une corrélation non nulle remet en cause l'hypothèse d'indépendance de celles-ci. À l'inverse, la correspondance s'applique au traitement de deux caractères dont les classes sont connues, permettant de hiérarchiser les informations données et d'analyser les relations entre modalités de variables qualitatives.

Le rapport de corrélation fait référence à la proportion de la variance de Y expliquée par X, obtenue à partir de la décomposition et de l'analyse de la variance.

3. Quelles différences faites-vous entre les valeurs marginales et les valeurs conditionnelles? Pourquoi distinguer les deux?

Les valeurs marginales sont des variables aléatoires X et Y, obtenues en marge du tableau étudié, dont les distributions individuelles respectives sont analysées afin d'obtenir des lois de probabilité de ces deux mêmes valeurs.

Les valeurs conditionnelles sont des valeurs dont la probabilité est calculée en connaissance d'un autre évènement ayant lieu. Cette probabilité représente la valeur attendue d'une variable aléatoire dans des conditions imposées par un événement extérieur.

La distinction entre les distributions marginales et les distributions conditionnelles permet d'analyser d'une part les distributions individuelles des deux variables aléatoires sélectionnées, et d'autre part de calculer la valeur attendue de ces variables dans les cas d'un évènement autre.

4. Quelles différences faites-vous entre variance et covariance?

La variance mesure la dispersion d'une variable autour de sa moyenne. La décomposition de la variance fait référence à la somme de la moyenne pondérée des

variances conditionnelles ainsi que de la variance pondérée des moyennes conditionnelles, et permet de traduire la dispersion de la distribution étudiée.

La covariance, elle, mesure les variations simultanées des variables aléatoires X et Y. Si celle-ci est nulle, cela indique une absence de liaison linéaire entre les variables aléatoires étudiées.

5. Pourquoi mesurer la corrélation ou l'indépendance?

En mesurant la corrélation ou l'indépendance de deux variables aléatoires X et Y, il est rendu possible de calculer si les deux variables étudiées entretiennent une relation logique, ou si, à l'inverse, n'influent en aucun cas l'une sur l'autre. C'est en calculant la corrélation ou l'indépendance de ces deux variables qu'il est possible d'analyser des phénomènes où plusieurs facteurs de façon simultanée, et d'en déduire si ces facteurs sont corrélés ou indépendants les uns des autres. Ainsi, il est possible de détecter les liens potentiels entre les facteurs (pourquoi ce facteur affecte-t-il la variable X ou Y ?) ou de construire des modèles permettant d'analyser deux variables simultanément (comme la régression linéaire, par exemple).

6. Quel est le principe de la méthode des moindres carrés? À quoi sert-elle?

La méthode des moindres carrés cherche à minimiser la somme des carrés des écarts observés entre Y observé, et Y estimé dans le modèle. Si la somme est faible, alors l'ajustement de la droite de régression des valeurs X et Y étudiées est bon. À l'inverse, si la somme est grande, alors l'ajustement est mauvais.

7. Expliquez en un court paragraphe ce qu'est la théorie de la corrélation (simple) ?

La théorie de la corrélation est une mesure ayant pour but d'exprimer la relation entre deux variables. Une corrélation simple fait référence à une relation composée de seulement deux variables. On analyse celles-ci en fonction de leur croissance ou décroissance simultanée : la corrélation est dite "positive" ou "directe" si Y croît en même temps que X, mais elle peut également être dite "négative" ou "indirecte" si Y décroît en même temps que X. Si, à l'inverse, les variables X et Y n'évoluent pas dans le même sens simultanément, alors on dit de celles-ci qu'elles ne sont pas corrélées. La théorie de la corrélation est mesurée par un coefficient compris entre -1 et 1, utilisé pour quantifier le sens et la force de la relation linéaire.

8. En quoi consiste le piège de l'autocorrélation?

L'autocorrélation mesure la corrélation d'une variable avec une copie différée de celle-ci, quantifiant la similitude entre les observations d'une variable aléatoire à différents moments. Ainsi, l'autocorrélation cherche à corréler une variable X avec une valeur x provenant d'un moment antérieur. Cependant, cela peut poser un problème dans les analyses statistiques classiques, comme la régression linéaire (expliquée ci-dessous), puisqu'on y

suppose l'indépendance des observations entre elles. Ainsi, les tests statistiques de la variable autocorrélée sont biaisés, et les estimateurs inefficaces.

9. Expliquez en un court paragraphe ce qu'est une régression linéaire?

Une régression linéaire est un calcul statistique permettant d'estimer la valeur d'une variable Y aléatoire en fonction d'une autre variable X. En mettant ces variables aléatoires en relation, on obtient une droite, qui modélise la relation entre la variable dépendante Y en fonction de la valeur de la variable explicative X. La régression linéaire est représentée sous la forme d'une droite, qui permet d'expliquer l'impact d'une variable statistique sur une autre variable. Ce calcul analyse la relation d'une variable X sur une variable Y, et permet de contrôler, prévoir, et de prendre des décisions quant à la relation existant entre les deux variables.

10. Quelle est la différence entre coefficient de corrélation et coefficient de détermination ?

Le cœfficient de corrélation est utilisé afin de mesurer le degré de liaison linéaire entre les variables aléatoires X et Y. Représenté par un nuage de points, le cœfficient de corrélation r indique si la relation étudiée est parfaitement linéaire ($r = +1$) ou manque d'une liaison linéaire ($r = 0$).

Le cœfficient de détermination indique la qualité de l'ajustement linéaire, et évalue l'efficacité du modèle de régression linéaire dans l'analyse de la relation entre les variables aléatoires Y et X. Ce coefficient provient du rapport entre la variance expliquée de la variable Y et la variance totale de la variable Y, et propose un pourcentage de la variance expliquée par rapport à la variance totale.

11. Pourquoi faut-il tester les deux droites de régression?

Il est nécessaire de tester les deux droites de régression (une de la variable aléatoire Y en fonction de X, et une de la variable aléatoire de X en fonction de Y) permettant d'analyser s'il existe des différences significatives entre les deux séries de résultats étudiées dans une régression linéaire. L'analyse de ces deux régressions peut permettre de déduire des conclusions cohérentes quant à la relation entre X et Y.

Mise en œuvre avec Python

Existe-t-il un lien entre le produit intérieur brut (P.I.B.) et la consommation énergétique?
Le fichier de données que vous allez analyser est issu des données de la Banque mondiale (https://donnees.banquemondiale.org/). Il regroupe deux jeux de données :
— le P.I.B. de chaque territoire (étatique ou non) de 1962 à 2024 en dollars courants (c'est-à-dire sans prendre en compte l'inflation);
— la consommation énergétique en kilogrammes équivalent pétrole de 1962 à 2024.
J'ai opéré les principales opérations de nettoyage, et fais en sorte que vous ayez le moins de

difficultés possibles à obtenir le résultat recherché.
1. Il existe un décalage entre les données du P.I.B. et de la consommation énergétique. Il faut de fait sélectionner en utilisant Pandas les colonnes PIB_2022 et Utilisation_d_energie_2022 dans le fichier pib-vs-energie.csv.
2. Malheureusement, plusieurs données sont censurées. Il vous faut créer un algorithme qui ne sélectionnera que les couples complets, c'est-à-dire ayant une valeur pour le P.I.B. et la consommation énergétique. Dit autrement, vous devez exclure les données manquantes (aucune valeur pour l'un et l'autre), et les données partielles (une donnée chez l'une, mais pas l'autre).
3. On considère que la variable explicative est la consommation énergétique et la variable à expliquer est le P.I.B. Calculer une régression linéaire simple entre les deux colonnes avec la méthode scipy.stats.linregress(x, y) prenant en arguments x, la variable à expliquer, et y, la variable explicative.
4. Calculer la corrélation simple entre les deux colonnes. Vous pouvez utiliser indifféremment les bibliothèques Pandas ou Scipy.
5. Faire un graphique de synthèse permettant de visualiser la droite de régression obtenue.
6. Dans votre rapport, vous commenterez votre résultat sous la forme d'un ou deux paragraphes.

Comme pour les autres séances, j'ai réalisé mon script python dans mon espace de travail .venv, et les données ont été importées sur Visual Studio Code à l'aide de la commande -curl :

```
claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 -m venv .venv
source .venv/bin/activate   # macOS / Linux

(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % curl -L \

"https://raw.githubusercontent.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/refs/heads/main/Seance-07/Exercice/src/data/pib-vs-energie.csv" \
  -o pib-vs-energie.csv
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  393k  100  393k    0     0  1399k      0 --:--:-- --:--:-- --:--:-- 1406k
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker %
```

Le code utilisé dans le dossier src pour effectuer les calculs demandés dans la consigne est disponible sur mon Github sous le nom main.py Séance 7. Les résultats de son exécution sont les suivants :

```
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 ./src/main.py
Dimensions du fichier : (217, 132)
Colonnes disponibles :
```

['Nom_du_territoire', 'Code_ISO_du_territoire', 'PIB_1960', 'PIB_1961', 'PIB_1962', 'PIB_1963', 'PIB_1964', 'PIB_1965', 'PIB_1966', 'PIB_1967', 'PIB_1968', 'PIB_1969', 'PIB_1970', 'PIB_1971', 'PIB_1972', 'PIB_1973', 'PIB_1974', 'PIB_1975', 'PIB_1976', 'PIB_1977', 'PIB_1978', 'PIB_1979', 'PIB_1980', 'PIB_1981', 'PIB_1982', 'PIB_1983', 'PIB_1984', 'PIB_1985', 'PIB_1986', 'PIB_1987', 'PIB_1988', 'PIB_1989', 'PIB_1990', 'PIB_1991', 'PIB_1992', 'PIB_1993', 'PIB_1994', 'PIB_1995', 'PIB_1996', 'PIB_1997', 'PIB_1998', 'PIB_1999', 'PIB_2000', 'PIB_2001', 'PIB_2002', 'PIB_2003', 'PIB_2004', 'PIB_2005', 'PIB_2006', 'PIB_2007', 'PIB_2008', 'PIB_2009', 'PIB_2010', 'PIB_2011', 'PIB_2012', 'PIB_2013', 'PIB_2014', 'PIB_2015', 'PIB_2016', 'PIB_2017', 'PIB_2018', 'PIB_2019', 'PIB_2020', 'PIB_2021', 'PIB_2022', 'PIB_2023', 'PIB_2024', 'Utilisation_d_energie_1960', 'Utilisation_d_energie_1961', 'Utilisation_d_energie_1962', 'Utilisation_d_energie_1963', 'Utilisation_d_energie_1964', 'Utilisation_d_energie_1965', 'Utilisation_d_energie_1966', 'Utilisation_d_energie_1967', 'Utilisation_d_energie_1968', 'Utilisation_d_energie_1969', 'Utilisation_d_energie_1970', 'Utilisation_d_energie_1971', 'Utilisation_d_energie_1972', 'Utilisation_d_energie_1973', 'Utilisation_d_energie_1974', 'Utilisation_d_energie_1975', 'Utilisation_d_energie_1976', 'Utilisation_d_energie_1977', 'Utilisation_d_energie_1978', 'Utilisation_d_energie_1979', 'Utilisation_d_energie_1980', 'Utilisation_d_energie_1981', 'Utilisation_d_energie_1982', 'Utilisation_d_energie_1983', 'Utilisation_d_energie_1984', 'Utilisation_d_energie_1985', 'Utilisation_d_energie_1986', 'Utilisation_d_energie_1987', 'Utilisation_d_energie_1988', 'Utilisation_d_energie_1989', 'Utilisation_d_energie_1990', 'Utilisation_d_energie_1991', 'Utilisation_d_energie_1992', 'Utilisation_d_energie_1993', 'Utilisation_d_energie_1994', 'Utilisation_d_energie_1995', 'Utilisation_d_energie_1996', 'Utilisation_d_energie_1997', 'Utilisation_d_energie_1998', 'Utilisation_d_energie_1999', 'Utilisation_d_energie_2000', 'Utilisation_d_energie_2001', 'Utilisation_d_energie_2002', 'Utilisation_d_energie_2003', 'Utilisation_d_energie_2004', 'Utilisation_d_energie_2005', 'Utilisation_d_energie_2006', 'Utilisation_d_energie_2007', 'Utilisation_d_energie_2008', 'Utilisation_d_energie_2009', 'Utilisation_d_energie_2010', 'Utilisation_d_energie_2011', 'Utilisation_d_energie_2012', 'Utilisation_d_energie_2013', 'Utilisation_d_energie_2014', 'Utilisation_d_energie_2015', 'Utilisation_d_energie_2016', 'Utilisation_d_energie_2017', 'Utilisation_d_energie_2018', 'Utilisation_d_energie_2019', 'Utilisation_d_energie_2020', 'Utilisation_d_energie_2021', 'Utilisation_d_energie_2022', 'Utilisation_d_energie_2023', 'Utilisation_d_energie_2024']

Aperçu des données longues :

	Nom_du_territoire	Code_ISO_du_territoire	Annee	PIB	Energie
30	France	fra	1990	1.257649e+12	223744434890.612152
31	France	fra	1991	1.258962e+12	236485406515.720276
32	France	fra	1992	1.389663e+12	232406014139.68103
33	France	fra	1993	1.314383e+12	236249808923.287231
34	France	fra	1994	1.385823e+12	227280548390.181793

Statistiques descriptives PIB :

count	3.100000e+01
mean	2.087941e+12

std 6.220855e+11
min 1.257649e+12
25% 1.468154e+12
50% 2.192146e+12
75% 2.669412e+12
max 2.926803e+12
Name: PIB, dtype: float64

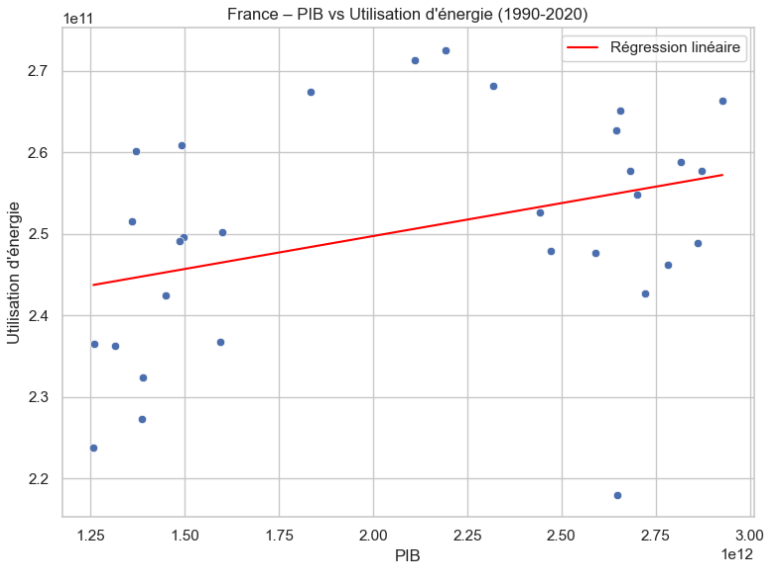
Statistiques descriptives Utilisation d'énergie :
count 3.100000e+01
unique 3.100000e+01
top 2.237444e+11
freq 1.000000e+00
Name: Energie, dtype: float64

Aperçu série temporelle (PIB/Energie par année) :
Année PIB Energie

1990 1.257649e+12 223744434890.612152
1991 1.258962e+12 236485406515.720276
1992 1.389663e+12 232406014139.68103
1993 1.314383e+12 236249808923.287231
1994 1.385823e+12 227280548390.181793

Covariance PIB–Energie : 3.1286e+21
Corrélation de Pearson r : 0.3577 (p = 0.04818)
Coefficient de détermination R^2 : 0.1280

Régression linéaire Energie = a + b * PIB
a (intercept) = 2.3356e+11
b (pente) = 8.0846e-03
r = 0.3577
R^2 = 0.1280



Les calculs obtenus ainsi que le graphique représentant la régression linéaire du PIB et de l'utilisation d'énergie nous indiquent que la relation entre les deux variables est une liaison positive mais faible. Avec les résultats produits, on étudie la force de la régression linéaire avec les coefficients suivants :

- Le coefficient de corrélation de Pearson étant égale à r = 0,36, on déduit une corrélation linéaire et positive entre le PIB et l'utilisation de l'énergie.
- Le coefficient R^2, soit le coefficient de détermination est égal à environ 0,13, indiquant que seulement 13% de la variance de la consommation d'énergie est expliquée par le PIB sur la période étudiée. Cela indique qu'une grande partie des variations dépend d'autres facteurs.

Le graphique produit par le script python nous indique une tendance similaire : le nuage de points démontre une tendance générale à la hausse de l'utilisation d'énergie lorsque le PIB augmente. On constate également que la droite de la régression linéaire est ascendante, indiquant ainsi une relation positive. Cependant, les points sont répartis avec une dispersion importante autour et éloignés de la droite de régression, traduisant une relation de corrélation n'étant pas parfaite.

Bonus

Vous avez écrit un algorithme permettant de traiter deux colonnes se rapportant à la même année. Écrivez un algorithme permettant de généraliser votre résultat à toutes les années de 1962 à 2022. N'oubliez pas d'organiser correctement vos fichiers de sortie

L'algorithme permettant de généraliser le résultat à toutes les années de 1962 à 2022 est disponible sur mon Github, sous le nom de main.py Séance 7 Bonus. Après exécution de la commande, les résultats sont les suivants :

```
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 ./src/main.py
Dimensions du fichier : (217, 132)
Colonnes disponibles :
['Nom_du_territoire', 'Code_ISO_du_territoire', 'PIB_1960', 'PIB_1961', 'PIB_1962', 'PIB_1963', 'PIB_1964', 'PIB_1965', 'PIB_1966', 'PIB_1967', 'PIB_1968', 'PIB_1969', 'PIB_1970', 'PIB_1971', 'PIB_1972', 'PIB_1973', 'PIB_1974', 'PIB_1975', 'PIB_1976', 'PIB_1977', 'PIB_1978', 'PIB_1979', 'PIB_1980', 'PIB_1981', 'PIB_1982', 'PIB_1983', 'PIB_1984', 'PIB_1985', 'PIB_1986', 'PIB_1987', 'PIB_1988', 'PIB_1989', 'PIB_1990', 'PIB_1991', 'PIB_1992', 'PIB_1993', 'PIB_1994', 'PIB_1995', 'PIB_1996', 'PIB_1997', 'PIB_1998', 'PIB_1999', 'PIB_2000', 'PIB_2001', 'PIB_2002', 'PIB_2003', 'PIB_2004', 'PIB_2005', 'PIB_2006', 'PIB_2007', 'PIB_2008', 'PIB_2009', 'PIB_2010', 'PIB_2011', 'PIB_2012', 'PIB_2013', 'PIB_2014', 'PIB_2015', 'PIB_2016', 'PIB_2017', 'PIB_2018', 'PIB_2019', 'PIB_2020', 'PIB_2021', 'PIB_2022', 'PIB_2023', 'PIB_2024', 'Utilisation_d_energie_1960', 'Utilisation_d_energie_1961', 'Utilisation_d_energie_1962', 'Utilisation_d_energie_1963', 'Utilisation_d_energie_1964', 'Utilisation_d_energie_1965', 'Utilisation_d_energie_1966', 'Utilisation_d_energie_1967', 'Utilisation_d_energie_1968', 'Utilisation_d_energie_1969', 'Utilisation_d_energie_1970', 'Utilisation_d_energie_1971', 'Utilisation_d_energie_1972', 'Utilisation_d_energie_1973', 'Utilisation_d_energie_1974', 'Utilisation_d_energie_1975', 'Utilisation_d_energie_1976', 'Utilisation_d_energie_1977', 'Utilisation_d_energie_1978', 'Utilisation_d_energie_1979', 'Utilisation_d_energie_1980', 'Utilisation_d_energie_1981', 'Utilisation_d_energie_1982', 'Utilisation_d_energie_1983', 'Utilisation_d_energie_1984', 'Utilisation_d_energie_1985', 'Utilisation_d_energie_1986', 'Utilisation_d_energie_1987', 'Utilisation_d_energie_1988', 'Utilisation_d_energie_1989', 'Utilisation_d_energie_1990', 'Utilisation_d_energie_1991', 'Utilisation_d_energie_1992', 'Utilisation_d_energie_1993', 'Utilisation_d_energie_1994', 'Utilisation_d_energie_1995', 'Utilisation_d_energie_1996', 'Utilisation_d_energie_1997', 'Utilisation_d_energie_1998', 'Utilisation_d_energie_1999', 'Utilisation_d_energie_2000', 'Utilisation_d_energie_2001', 'Utilisation_d_energie_2002', 'Utilisation_d_energie_2003', 'Utilisation_d_energie_2004', 'Utilisation_d_energie_2005', 'Utilisation_d_energie_2006', 'Utilisation_d_energie_2007', 'Utilisation_d_energie_2008', 'Utilisation_d_energie_2009', 'Utilisation_d_energie_2010', 'Utilisation_d_energie_2011', 'Utilisation_d_energie_2012', 'Utilisation_d_energie_2013', 'Utilisation_d_energie_2014', 'Utilisation_d_energie_2015', 'Utilisation_d_energie_2016', 'Utilisation_d_energie_2017', 'Utilisation_d_energie_2018', 'Utilisation_d_energie_2019', 'Utilisation_d_energie_2020', 'Utilisation_d_energie_2021', 'Utilisation_d_energie_2022', 'Utilisation_d_energie_2023', 'Utilisation_d_energie_2024']

Aperçu des données longues :
   Nom_du_territoire Code_ISO_du_territoire  Annee         PIB           Energie
30            France                   fra   1990  1.257649e+12  223744434890.612152
31            France                   fra   1991  1.258962e+12  236485406515.720276
32            France                   fra   1992  1.389663e+12   232406014139.68103
33            France                   fra   1993  1.314383e+12  236249808923.287231
34            France                   fra   1994  1.385823e+12  227280548390.181793
```

Statistiques descriptives PIB :
count 3.100000e+01
mean 2.087941e+12
std 6.220855e+11
min 1.257649e+12
25% 1.468154e+12
50% 2.192146e+12
75% 2.669412e+12
max 2.926803e+12
Name: PIB, dtype: float64

Statistiques descriptives Utilisation d'énergie :
count 3.100000e+01
unique 3.100000e+01
top 2.237444e+11
freq 1.000000e+00
Name: Energie, dtype: float64

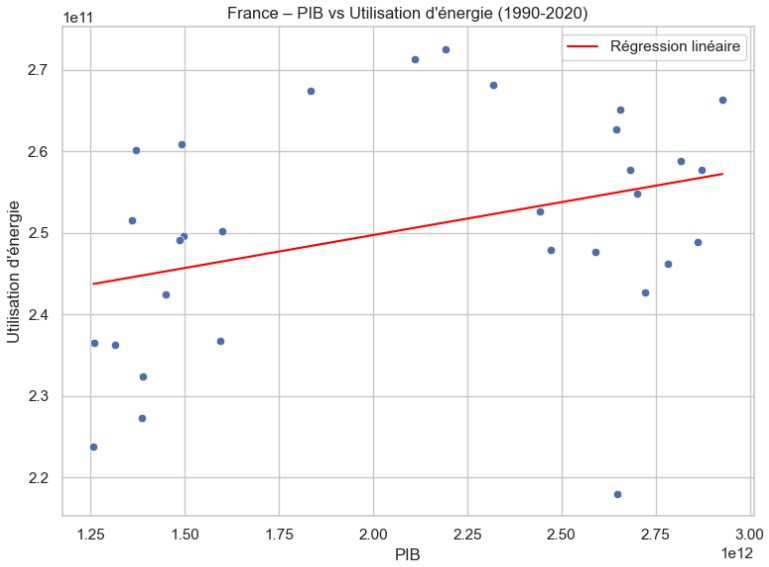
Aperçu série temporelle (PIB/Energie par année) :
Année PIB Energie

1990 1.257649e+12 223744434890.612152
1991 1.258962e+12 236485406515.720276
1992 1.389663e+12 232406014139.68103
1993 1.314383e+12 236249808923.287231
1994 1.385823e+12 227280548390.181793

Covariance PIB–Energie : 3.1286e+21
Corrélation de Pearson r : 0.3577 (p = 0.04818)
Coefficient de détermination R² : 0.1280

Régression linéaire Energie = a + b * PIB
 a (intercept) = 2.3356e+11
 b (pente) = 8.0846e-03
 r = 0.3577
 R² = 0.1280

Ainsi que le même graphique de régression linéaire :



Séance 8 - Chapitre 7 : Étude de deux variables qualitatives

Questions de cours

1. La corrélation entre deux variables qualitatives a-t-elle un sens? Expliquez votre réponse.

La corrélation entre deux variables qualitatives a un sens, puisque la quantité de variables qualitatives existantes est bien plus importante que la quantité de variables quantitatives. Ainsi, la corrélation entre deux variables qualitatives permet de mesurer l'intensité de la liaison entre les deux variables étudiées.

2. Pourquoi pratiquer le test d'indépendance du χ2 ?

Le test d'indépendance du X^2 (chi carré) permet de mettre en évidence une liaison entre les deux variables étudiées. Par l'hypothèse nulle, on teste si les deux variables étudiées X et Y sont indépendantes ou associées. Sous cette même hypothèse nulle, on suppose que les distributions conditionnelles de l'une des variables sont les mêmes pour toutes les modalités de l'autre. Lorsque la statistique du chi carré est trop grande (la plupart du temps, p value < 0.05), on rejette l'hypothèse nulle et conclut la liaison existante entre les deux variables qualitatives étudiées.

3. Expliquez dans un court paragraphe ce qu'est l'analyse de la variance à simple entrée.

L'analyse de la variance à simple entrée est une manipulation permettant de généraliser le test de comparaison des moyennes de plusieurs échantillons indépendants. Cette théorie nous permet d'étudier la variabilité d'un objet par rapport à un facteur donné, tout en contrôlant systématiquement ce même facteur, afin d'en dissocier les parts revenant à ce facteur. En étudiant la dépendance d'une variable quantitative à une ou deux variables qualitatives, l'analyse de la variance à simple entrée permet de contrôler un de ces facteurs, que son intervention dépende de sa nature ou de son intensité. Ainsi, l'analyse de la variance à simple entrée permet de contrôler un seul facteur dans l'analyse de son impact sur une autre variable étudiée.

4. Qu'est-ce qu'un rapport de corrélation ? Quelle différences avec la correspondance?

Un rapport de corrélation est un outil statistique permettant d'étudier la liaison entre une variable quantitative Y et une variable qualitative X dans un ensemble d'individus *n*. En calculant le rapport de corrélation, on obtient les variations entre les différentes modalités, soit la variation expliquée par le facteur contrôlé dans l'analyse de la liaison entre les variable Y et X données. Avec un résultat entre 0 (les variables ne sont pas liées) et 1 (les variables sont parfaitement liées), cette manipulation statistique mesure ainsi l'intensité de la relation.

La correspondance, elle, se concentre sur la hiérarchisation des informations données, permettant d'analyser les relations entre modalités de variables qualitatives. La différence entre le rapport de corrélation et la correspondance se trouve ainsi dans le processus même de ces deux outils statistiques : bien que les deux cherchent à analyser la liaison entre deux variables (une qualitative et une quantitative pour le rapport de corrélation, deux variables qualitatives pour la correspondance), le rapport de corrélation mesure l'intensité de la relation, alors que la correspondance se concentre sur les modalités factorielles afin de visualiser les proximités des associations entre les variables étudiées.

5. Qu'est-ce qu'une analyse factorielle ?

Une analyse factorielle permet de décomposer un tableau étant considéré comme difficile à analyser afin de produire un nombre de facteurs latents plus petits, ayant pour but de résumer l'information. Le but d'une analyse factorielle est de déterminer et mettre en avant les structures de facteurs corrélées dans un tableau.

6. Expliquez en un court paragraphe ce qu'est l'analyse factorielle des correspondances.

L'analyse factorielle des correspondances (A.F.C.) est une méthode statistique ayant pour objectif de transformer un tableau de nombre en un graphique, appelé *mapping*, plus lisible afin de visualiser la liaison entre deux variables étudiées. Par l'utilisation de cette méthode statistique, l'objectif revient à exprimer graphiquement la différence entre la situation observée et la situation théorique, afin de visualiser celle-ci de façon optimale et de tirer des conclusions quant à la nature de la liaison entre les deux variables étudiées. Une fois la manipulation terminée, le tracé de différents vecteurs sur le *mapping* permet de visualiser si la liaison entre les valeurs étudiées est une situation de conjonction (le produit scalaire est strictement positif, impliquant une affinité entre les deux modalités), une situation d'opposition (le produit scalaire est strictement négatif, impliquant une répulsivité entre les deux modalités), ou une situation de quadrature (le produit scalaire est nul, les deux modalités sont dans une situation d'équilibre). En projetant sur un plan les modalités des lignes et des colonnes du tableau étudié, les proximités trouvées sur ce plan permettent d'en déduire la similarités des valeurs, et d'associer les modalités entre elles.

Mise en œuvre avec Python

Vous allez utiliser les données 2024 de l'Institut national de la statistique et des études économiques (I.N.S.E.E.) concernant la relation entre catégorie socioprofessionnelles et le sexe biologique https://www.insee.fr/fr/statistiques/2381478#figure1_radio2. J'ai simplifié le tableau initial. Si vous avez compris le cours, vous avez compris qu'il s'agit d'un tableau de contingence avec deux variables qualitatives. Normalement, vous le calculerez un tableau croisé dynamique avec la méthode de la bibliothèque Pandas, crosstab(). Toutefois, il arrive souvent, comme dans le cas présent, que le tableau de contingence n'ait pas être calculé. De fait, on ne peut pas utiliser les méthodes de Pandas directement, notamment pour calculer les marges.

1. Calculer les marges des lignes et des colonnes en vous servant des fonctions locales sommeDesColonnes() et sommeDesLignes().
2. Faire une condition vérifiant si le total des marges des lignes et le total des marges des colonnes est identique.
3. Faire un test d'indépendance du χ2 à partir de la bibliothèque scipy.stats et de sa méthode chi2_contingency(). Existe-t-il une liaison?
4. Calculer l'intensité de liaison φ2 de Pearson à partir des résultats précédents.
5. Dans votre rapport, faire un court paragraphe expliquant vos résultats.

Comme pour toutes les séances, j'ai créé un espace de travail .venv avec Python, et ai importé les données nécessaires avec la commande curl et le "raw file" disponible sur le Github :

claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % source ./venv/bin/activate
(venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % curl -L \

"https://raw.githubusercontent.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/refs/heads/main/Seance-08/Exercice/src/data/Socioprofessionnelle-vs-sexe.csv" \
  -o Socioprofessionnelle-vs-sexe.csv
  % Total   % Received % Xferd Average Speed   Time   Time   Time Current
                                 Dload Upload   Total   Spent   Left Speed
100   359 100   359   0   0   1752   0 --:--:-- --:--:-- --:--:-- 1759
(venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker %

Le code utilisé pour répondre à la consigne est disponible sur mon GitHub sous le nom de main.py Séance 8. Après exécution, les résultats sont les suivants :

(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 src/main.py
Tableau de contingence (effectifs) :

	Femmes	Hommes
Agriculteurs exploitants	94.0	273.0
Artisans, commerçants et chefs d'entreprise	661.0	1295.0
Cadres et professions intellectuelles supérieures	2889.0	3797.0
Professions intermédiaires	3918.0	3511.0
Employés	5770.0	1816.0
Ouvriers	1193.0	4638.0
Chômeurs n'ayant jamais travaillé	167.0	166.0
Inactifs	13566.0	10645.0
Non classés	60.0	63.0

Marges de lignes :
Agriculteurs exploitants : 367

Artisans, commerçants et chefs d’entreprise : 1956
Cadres et professions intellectuelles supérieures : 6686
Professions intermédiaires : 7429
Employés : 7586
Ouvriers : 5831
Chômeurs n'ayant jamais travaillé : 333
Inactifs : 24211
Non classés : 123

Marges de colonnes :
Femmes : 28318
Hommes : 26204

Total des marges de lignes : 54522
Total des marges de colonnes : 54522

Vérification : le total des marges de lignes est égal au total des marges de colonnes.

Test du chi2 d'indépendance :
Chi2 = 4812.4194
ddl = 8
p-value = 0.0000e+00
Conclusion : On rejette l'hypothèse d'indépendance : il existe une liaison entre catégorie socioprofessionnelle et sexe.

Intensité de liaison phi^2 de Pearson : 0.0883

Résumé des résultats sauvegardé dans : resultats_chi2_socioprofessionnelle_sexe.csv

Les résultats obtenus suite à l’exécution du script python montrent qu’il existe une liaison entre les catégories socioprofesionnelles et le sexe biologique. Le test du Chi2 montre une p value pratiquement nulle (p = 0.0000) qui, pour un seuil de 5%, porte à rejeter l’hypothèse nulle d’indépendance. Comme l’indique la conclusion produite dans les résultats : il existe une liaison entre catégorie socioprofessionnelle et sexe biologique.

L’intensité de la liaison, calculée par le phi^2 de Pearson, indique un résultat de 0.0883. Celui-ci étant compris entre 0 et 1, la valeur obtenue démontre une relation d’une force modérée, indiquant qu’il existe bien une association entre les valeurs. Ainsi, on peut conclure que la catégorie socioprofessionnelle dépend bien du sexe, mais ne représente pas une association “parfaite”.

Bonus

Les points bonus sont attribués par rapport aux chapitres d’approfondissement. Avec le fichier Echantillonnage-100-Echantillons.csv, faire une analyse ANOVA. Avec le tableau de données, calculer une A.F.C. et faisant le commentaire.

Le code utilisé pour répondre à la consigne est disponible sur mon GitHub sous le nom de main.py Séance 8 Bonus. Après exécution, les résultats sont les suivants :

```
(.venv) claraboriebioules@MacBook-Air-de-Clara-2 projet-python-docker % python3 src/main.py
Aperçu des données :
   Pour  Contre  Sans opinion
0   395     396          209
1   379     432          189
2   384     426          190
3   395     407          198
4   389     413          198

Statistiques descriptives par modalité :
Pour : count    100.000000
mean     390.520000
std       10.944746
min      367.000000
25%      384.000000
50%      392.000000
75%      397.000000
max      420.000000
Name: Pour, dtype: float64
Contre : count    100.000000
mean     416.060000
std       11.148058
min      387.000000
25%      409.000000
50%      416.000000
75%      423.000000
max      442.000000
Name: Contre, dtype: float64
Sans opinion : count    100.00000
mean     193.42000
std        8.50452
min      174.00000
25%      188.00000
50%      193.00000
75%      198.00000
max      222.00000
Name: Sans opinion, dtype: float64

ANOVA une voie (Pour vs Contre vs Sans opinion) :
  F = 14075.7097
```

p-value = 5.9800e-295

Conclusion : On rejette H0 : il existe au moins une différence significative entre les moyennes des trois positions.

Tableau de contingence global (somme des 100 échantillons) :

	Pour	Contre	Sans opinion
Total	39052	41606	19342

Analyse factorielle des correspondances (AFC) :

Valeurs propres (inerties) :

	axe	valeur_propre	pourcentage_inertie
0	1	5.113362e-04	5.273756e+01
1	2	4.582502e-04	4.726244e+01
2	3	7.917243e-33	8.165588e-28

Coordonnées des modalités (colonnes) sur les deux premiers axes :

	Dim1	Dim2
Pour	-0.024520	0.013280
Contre	0.025518	0.007719
Sans opinion	-0.005384	-0.043416

Fichiers exportés :

- resultats_anova_echantillons.csv
- resultats_afc_valeurs_propres.csv
- resultats_afc_coordonnees.csv

Les résultats ci-dessus proposent une analyse ANOVA des données proposées. Celle-ci démontre des moyennes très différentes, avec des écarts-type allant à des dizaines d'unités. La statistique F et la p-value proposées par l'ANOVA montre des résultats très faibles, ce qui pousse à conclure que les moyennes des effectifs "Pour", "Contre", "Sans opinion" ne sont pas égales, étant donné qu'il existe des différences très marquées entre les trois dans les 100 échantillons proposés. Ainsi, la conclusion proposée après l'exécution du code écrit : "On rejette H0 : il existe au moins une différence significative entre les moyennes des trois positions."

L'analyse factorielle des correspondances réalisée en fonction du tableau des données de l'ANOVA montre des totaux très différents selon les trois modalités "Pour", "Contre", et "Sans opinion". Les deux premières valeurs montrent cependant leur importance dans la variabilité, puisqu'elles captent environ 52,7% et 47,3% de l'inertie. Ainsi, l'AFC démontre un axe principal séparant les modalités "Pour" et "Contre", tandis que "Sans opinion" se distingue sur un second axe, reflétant les profils de réponses relativement différents pour chaque position.