LETTRES SORBONNE UNIVERSITÉ

SORBONNE UNIVERSITÉ

UFR GÉOGRAPHIE ET AMÉNAGEMENT - INSTITUT DE GÉOGRAPHIE
MASTER 1 Géographie, Aménagement, Environnement et Développement

Rapport d'activité M1 Analyse de données

Niveau intermédiaire

Source : *https://www.istockphoto.com/fr/photos/python-code*

Par Louka Alet

19 décembre 2025

Dossier réalisé dans le cadre de M1GEANDO - Analyse de données (intermédiaires)

Table des matières

I. Consignes

Votre rapport doit apporter la réponse aux questions posées de manière structurée, c'est-à-dire sous la forme d'un texte de synthèse répondant à l'ensemble des éléments à expliciter avec vos propres mots.

Le code va produire des résultats en fonction des séances vous indiquerez dans votre rapport :

- vos résultats sous la forme d'un tableau ;
- vos graphiques.

Vous commenterez librement sous la forme d'un court paragraphe les résultats obtenus. Le commentaire peut être technique ou scientifique.

Votre rapport doit être structuré par séance d'activités. Il doit contenir des phrases complètes.

Votre rapport doit se conclure par une réflexion personnelle sur les sciences des données et les humanités numériques en fonction des exercices de votre parcours.

II. Séance 4 :

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?

Pour faire le choix entre une distribution statistique avec des variables discrètes ou une distribution statistique avec des variables continues, le critère principal à mettre avant est selon moi la nature de la variable. En fonction de la nature des variables que comportent la base de données, discrètes ou continues, le choix se portera sur une distribution basée sur l'une ou l'autre approche. Les données discrètes sont des données qui peuvent représenter des phénomènes avec des limites distinctes (ex : nombre de personnes par foyer) et les données continues sont des phénomènes qui peuvent prendre une infinité de valeurs (ex : température). Ainsi, une fois le type de données déterminé, découlera le choix de la distribution.

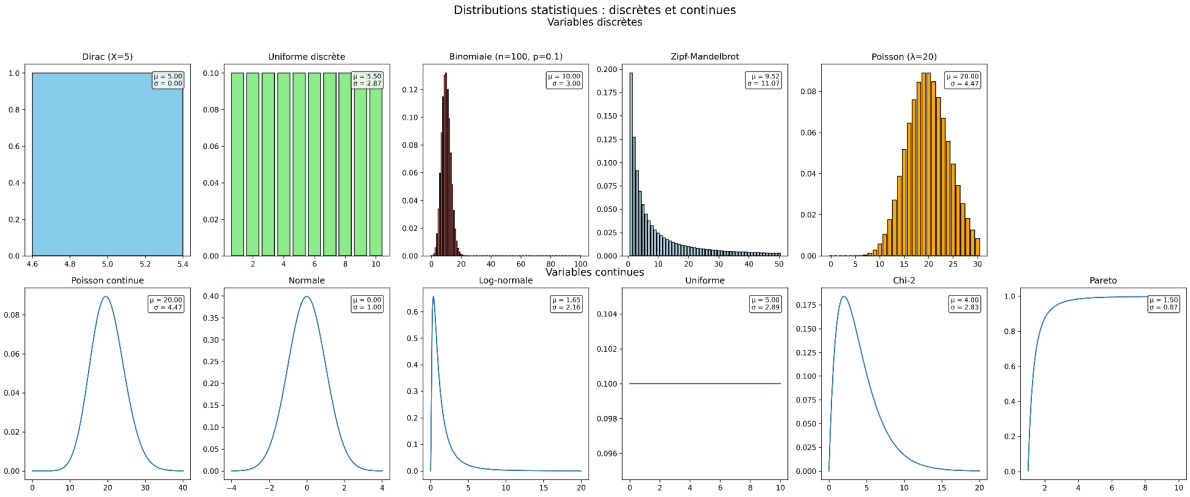
Toutefois, si les données ne sont pas encore récoltées et nous souhaitons déterminer quels types de distribution nous souhaitons mettre en place il serait pertinent de mettre en avant la nature de l'objet étudié ou de l'étude dans lequel il s'inscrit. Il est important de prendre en compte comment valoriser les résultats : une distribution statistique sert avant tout à être interprétée et ses résultats doivent impérativement être cohérents avec la nature de l'étude.

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

Selon moi de nombreuses lois peuvent être utilisées en géographie. De façon non exhaustives voici quelques exemples :

- **La loi du $\chi 2$:** le test de khi², qui est une loi qui sert à déterminer s'il existe une relation entre deux variables qualitatives. En géographie cela est très précieux car permet de tirer des conclusions plus nuancées des hypothèses qui sont confrontées.
- **Loi normale :** cette loi permet de vérifier comme des tests d'hypothèses, estimation du taux de non-conformité, intervalles de confiance etc. Pour moi elle sera donc la base des nombreux usages statistiques afin d'observer la qualité et distribution des données avant de les traiter.
- **Loi de Zipf (ou loi rang-taille) :** la loi de Zipf est aussi appelée loi rang taille car elle permet de classer des variables en fonction de leur taille. Elle est très pertinente en géographie afin de classer des territoires selon leur taille pour les comparer, comme des villes.

Mise en œuvre avec Python :



Voici les graphiques que j'ai réalisé, vous pourrez les retrouver sous le nom "graphiques_seance_4" dans le dossier src. Pour plus de facilité j'ai choisi de faire une seule figure, avec sur la ligne du haut les variables discrètes et celles du bas les variables continues. Aussi, les résultats de la moyenne et écart-type sont directement affichés sur les graphiques pour faciliter la lecture. Ces graphiques représentent l'ensemble des lois demandées pour l'exercice, et permettent d'observer visuellement leurs différences.

III. Séance 5

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?

Définition : l'échantillonnage peut-être défini comme la sélection d'individus dans une population mère. L'ensemble de la sélection sera alors nommé l'échantillon (population fille).

Pourquoi échantillonner ? Il est pertinent d'échantillonner car dans l'immense majorité des cas il serait impossible d'enquêter la totalité de la population mère. (eg : il est impossible d'interroger la totalité de la population française pour connaître les intentions de vote lors d'élections, on va faire ces sondages auprès d'échantillons représentatifs, en théorie).

Quelles sont les méthodes d'échantillonnage ? Si l'on souhaite tirer un échantillon représentatif, l'échantillon doit de se faire avec rigueur et des méthodes précises. A ma connaissance il existe deux grandes catégories d'échantillonnage représentatif.

- La catégorie probabiliste qui correspond à une liste sans omission ni répétition de tous les individus constituant la population parente (eg : tirage simple au hasard).
- La catégorie empirique qui correspond à la sélection des individus composant l'échantillon directement sur le terrain (eg : échantillon par quota).

Des cas d'**échantillons non représentatifs** existent aussi dans des contextes spécifiques où l'accès à la population parente est difficile (eg méthode des choix raisonnés).

2. Comment définir un estimateur et une estimation?

Un **estimateur** est une nouvelle variable aléatoire construite à partir des résultats obtenus sur un échantillon aléatoire, et dont la valeur se rapproche du paramètre que l'on cherche à connaître.
L'**estimation** quant à elle est le processus par lequel est produite une estimation d'un paramètre de la population.

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

L'élément différenciant entre l'intervalle de fluctuation et l'intervalle de confiance est que le premier est utilisé lorsque que la proportion p dans la population est connue (ou qu'une hypothèse sur sa valeur est réalisée), alors que l'intervalle de confiance permet justement d'estimer cette proportion inconnue p dans une population.

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

Formellement le biais est défini comme la différence entre l'espérance de l'estimateur et la valeur à estimer.

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

Je ne suis pas sûr d'avoir bien saisi la question, en lisant le cours la seule réponse qui me vient est la **méthode des quotas**. Cette méthode empirique se construisant en définissant des variables de contrôle à partir des caractéristiques connues de la population parente puis sur le terrain sélection des individus à enquêter selon des quotas préétablis pour chaque variable de contrôle.
La **notion de données massives** (ou big data) correspond à un ensemble données tellement important qu'il est impossible à gérer par les outils classiques de gestion statistiques. Le lien que je pourrais faire entre ces deux concepts est que devant la quantité des données massives, il serait pertinent d'utiliser la méthode empirique des quotas pour faciliter la gestion de ces données.

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

Le choix d'un estimateur repose sur de nombreux enjeux.

- Comme mentionné dans une question précédente, les **biais** sont des facteurs à prendre en compte. Certains estimateurs comportent en effet plus de biais que d'autres, ou des biais qui sont différents.
- Ensuite la **précision** d'un estimateur ponctuel, qui se traduit par la variance d'un estimateur. Il est possible d'avoir une analyse plus fine de la précision d'un estimateur en utilisant la méthode des carrés moyens de l'erreur aussi appelée erreur quadratique. Celle-ci se calcule en faisant la somme de la variance de l'estimateur et du carré du biais de l'estimateur.
- La **consistance** de l'estimateur est aussi à prendre en compte. Elle est définie comme la concentration d'un estimateur dans une zone, quand l'estimateur tend vers l'infini.
- Enfin la **convergence** est à prendre en compte : la concentration de la distribution autour d'une valeur lorsque la taille d'un échantillon tend vers l'infini.

Ces quatres enjeux sont à prendre dans le choix de l'estimateur, un **arbitrage** est donc à réaliser entre eux afin d'utiliser l'estimateur qui se prête le plus à la situation.

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

De nombreuses méthodes existent, pour n'en citer que quelques unes :

- Estimation ponctuelle
- Estimation par intervalle
- Méthode des moindres carrés (régression linéaire simple ou multiple, Gauss-Newton…)
- Analyse des résidus

Pour en sélectionner une cela peut **dépendre par exemple du nombre variables, de la disponibilité des données ou encore des hypothèses décidées.**

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

Quelques exemples de tests :

- Test de Student
- Test du khi2

- Test de Mann-Whitney

Un test permet de valider ou d'invalider une hypothèse formulée : on observe donc si les données statistiques observées peuvent-elles être rapprochées de lois de probabilité théoriques.

Pour créer un test statistique il faut :

1. Définir les hypothèses
2. Définir les risques d'erreur
3. Choisir le type du test (unilatéral, latéral…)
4. Choisir le test en lui-même

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

L'une des critiques majeures de la statistique inférentielle est que le nombre trop bas des échantillons ne permet pas faire des extrapolations pertinentes. Selon moi cette critique est entendable. Toutefois, je pense que le risque que échantillons tirés ne représentent pas la population mère peut être évité, notamment grâce à l'ensemble des méthodes analysées dans cette séance. De plus, au-delà d'extrapolation sur la population je pense que la statistique inférentielle est un outil très précieux pour observer des tendances au sein de la population mère (observer justement les spécificités des groupes).

Mise en œuvre avec Python :

- **Théorie de l'échantillonnage**

Pour chaque colonne de la population fille, calculer la moyenne obtenue.

Pour	Contre	Sans opinion
391.0	416.0	193.0

Fréquences pour chaque moyenne de la population fille

Pour	Contre	Sans opinion
0.3910	0.4160	0.1930

Fréquences de la population mère

Pour	Contre	Sans opinion
0.3899	0.4169	0.1931

Intervalle de fluctuation de chacune des fréquences de la population fille, à un seuil de 95 %

Pour	Contre	Sans opinion
[0.3705, 0.4115]	[0.3953, 0.4367]	[0.1765, 0.2095]

Dans votre rapport, 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?

Un intervalle de fluctuation permet d'observer si les valeurs de l'échantillon suivent la même tendance que les valeurs de la population mère, avec un seuil de probabilité.. Dans notre cas, les 3 fréquences sont bien comprises dans les 3 intervalles correspondantes. On peut donc en déduire que les fréquences des données de l'échantillon correspondant appartiennent à l'intervalle avec une probabilité d'au moins 95%.

- **Théorie de l'estimation**

Fréquences de l'échantillon isolé

Pour	Contre	Sans opinion
0.3950	0.3960	0.2090

Intervalle de confiance de chacune des fréquences à un seuil de 95 %

Pour	Contre	Sans opinion
[0.3745, 0.4155]	[0.3755, 0.4165]	[0.1920, 0.2260]

Dans votre rapport, vous interpréterez le résultat obtenu et vous le comparerez avec le résultat précédent.

Encore une fois les fréquences de l'échantillon correspondent aux intervalles de confiance respectifs. Un intervalle de confiance n'indique pas la même chose qu'un intervalle de fluctuation : il permet d'estimer un paramètre inconnu au sein d'une population. Dans notre cas, il est donné à 95%. Ainsi il est possible d'affirmer que la probabilité d'avoir construit un intervalle contenant effectivement le paramètre mère inconnu est de 95%.

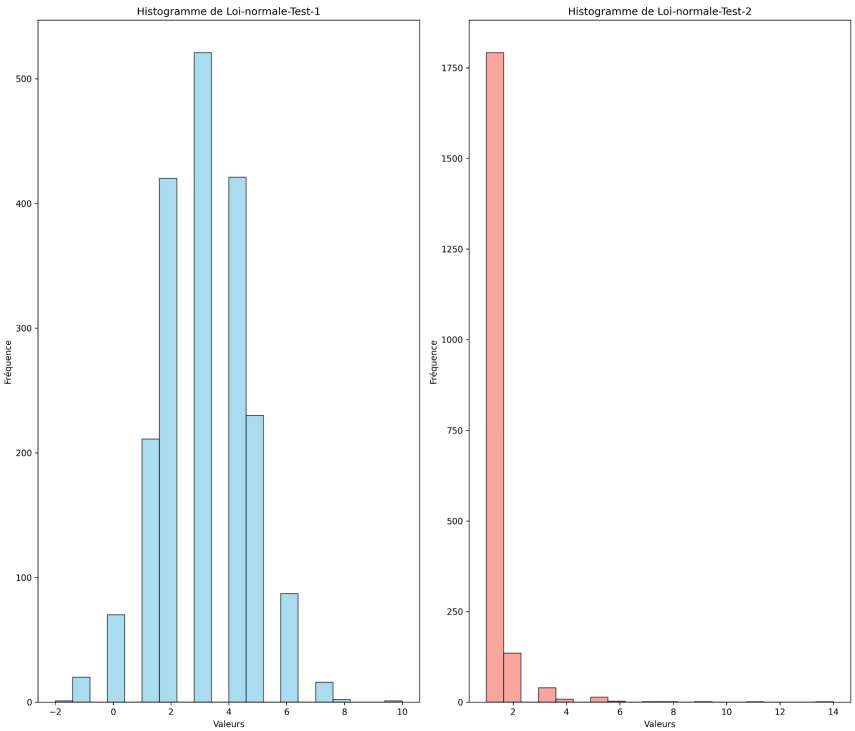
- **Théorie de la décision**

Résultats test shapiro

	Loi 1	Loi 2
statistique	0.96	0.26
p-value	**6.29**	**7.05**

Laquelle est une distribution normale? et pourquoi ?

Aucune des deux distribution n'indique une distribution normale. Pour être considérée comme normale, la distribution doit avoir une p-value < 0,05 hors ici les deux sont significativement supérieures. J'ai refait de plusieurs façons différentes les tests de loi normale, sur les deux lois et je tombe à chaque fois sur les mêmes p-value. Hors la p-value est un pourcentage et doit donc être comprise entre 0 et 100 strictement, ce qui n'est pas le cas dans mes valeurs. Au vu de ce dernier commentaire et de la question posée, il est donc évident qu'il y a un problème à un endroit mais je ne sais pas dire s'il concerne mon code, le fichier ou autre. Toutefois, pour tenter d'identifier lequel des deux jeux de données suit une loi normale j'ai choisi de faire deux représentations graphiques des fichiers. Grâce à la visualisation graphique, il alors possible de dire que c'est le fichier 1 qui suit une loi normale.



IV. Séance 6

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 une série dont les valeurs peuvent être ordonnées. L'exemple le plus facile à comprendre est une liste de nombre, mais une statistique ordinale peut très bien aussi prendre la forme d'une série de caractères (comme la dangerosité par exemple, ou l'importance).

Elle s'oppose à la statistique catégorielle nominale (série dont les valeurs ne peuvent pas être hiérarchisées). La statistique ordinale utilise donc des variables tant qualitatives que quantitatives.

Cette statistique est très utilisée en géographie afin de classer des entités. Par exemple en géographie physique, grâce au taux d'ensoleillement il est possible d'utiliser la statistique ordinale pour hiérarchiser les territoires les plus exposés au soleil par an.

2. Quel ordre est à privilégier dans les classifications?

L'ordre à privilégier dans les classifications est l'ordre croissant, aussi appelé ordre naturel. Cela facilite la lecture et permet dans le même temps de repérer les potentielles valeurs dites "aberrantes" (excessivement éloignées de la répartition attendue).

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

La corrélation des rangs consiste à vérifier si deux classements qui auraient été au sein d'une même population (mais à partir de deux variables différentes) suivent le même ordre. Donc, elle met en avant les similitudes et différences. Elle se fait grâce à des méthodes comme les coefficients de Spearman ou de Kendall. La concordance de classement cherche quant à elle comparer les paires d'objets pour observer les concordances et discordances entre tous les objets.

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

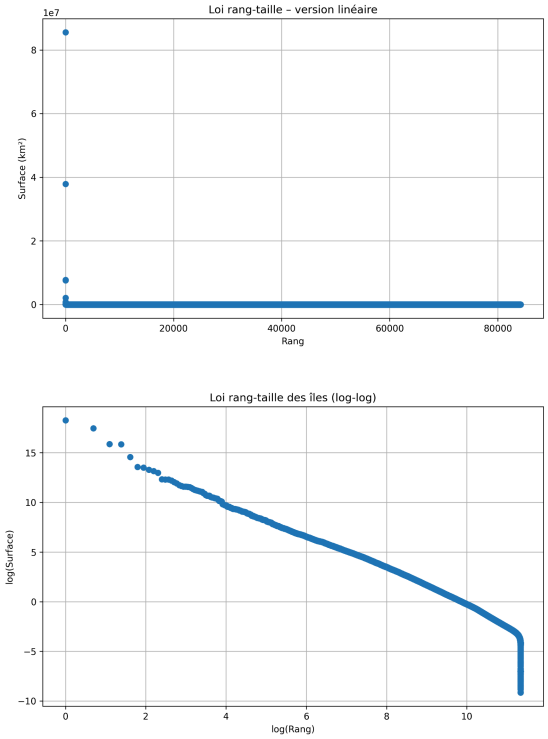
Le premier compare les rangs objet par objet afin de mesurer la corrélation, et le second va observer pour chaque paire d'objet à quel point l'ordre est conservé entre les deux classements. Ainsi le premier met juste en avant un coefficient de corrélation alors que le second donne plus de détails sur les différences de positions entre les deux classements.

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

Ces coefficients permettent d'obtenir la proportion de paires concordantes par rapport aux paires discordantes. Ces coefficients, compris entre -1 et 1 mettent donc en avant le lien entre les paires : si la proportion est nulle il n'y a pas de lien, si elle est positive il y a concordance, et si elle est négative il y a discordance.

Mise en œuvre avec Python :

Image loi rang-taille (avant puis après conversion avec log)



Coefficient de corrélation des rangs et la concordance des rangs

	Corrélation de Spearman	Concordance de Kendall
statistique	**0.11**	**0.08**
p-value	0.15	0.11

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.

Les résultats montrent une faible corrélation de Spearman (+0.11) ainsi qu'une faible concordance de Kendall (+0.08). Comme détaillé plus haut, plus ces statistiques sont proches de 0 plus la corrélation sera nulle. De plus, si elles sont supérieures à 0 cela signifie qu'il y a corrélation positive. Dans notre cas, cela signifie que la hiérarchie de classement à partir de la variable *nombre d'habitants* est très similaire à la

hiérarchie de la variable *densité*. Ainsi il est possible de conclure que le classement des pays par nombre d'habitants est similaire que celui par densité : plus le pays est peuplé, plus il sera haut dans le classement et plus il est dense, plus il sera haut dans le classement.

V. Séance 7

Questions de cours :

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

Les statistiques univariées étudient une seule variable. Les statistiques bivariées étudient deux variables simultanément. L'étude de deux variables permet donc de faire des liens entre ces deux objets afin de voir s'il y a corrélation et si l'un influe sur l'autre. Cette approche est selon moi impérative pour mieux cerner les nuances de notre monde.

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

Selon moi la différence entre corrélation et correspondance est majeure. La correspondance est la similitude entre deux variables. Par exemple, on observe le taux d'ensoleillement annuel par territoire et la production agricole de ces territoires : ici la correspondance seront les territoires à la fois les plus ensoleillés et avec la production agricole la plus élevée. Tandis que la corrélation est le taux de l'influence d'une variable sur l'autre. Ici ce serait déterminer la valeur numérique de l'influence de l'ensoleillement sur la production. La corrélation permet donc de déterminer un rapport permettant de tirer des conclusions sur l'ensemble des territoires alors que la correspondance met en avant les lieux où cette corrélation est présente : l'usage est donc distinct.

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

Les valeurs conditionnelles sont les valeurs de l'influence d'une loi sur une autre (eg : X sur Y ; ou Y sur X). Les valeurs marginales sont elles les distributions individuelles des variables étudiées, sans mesure leur influence de l'une sur l'autre. Il est intéressant de distinguer les deux afin de comprendre la variation de ces deux valeurs. En effet, si la variable X a un variation très marginale, même si la variable Y avait une influence considérable sur X, la variation de X serait fiable.

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

La variance est la méthode pour observer la dispersion d'une variable. Alors que la covariance est une méthode qui permet d'attribuer la part de la variation d'une variable B induite par la variation de la variable A. Bien que ces deux méthodes aient une dénomination proche, leur usage n'a donc pas grand chose à voir.

5. Pourquoi mesurer la corrélation ou l'indépendance?

La corrélation permet de mesurer si le mouvement d'une variable influe sur une autre. C'est donc une méthode indispensable pour évaluer si une action n'induira pas d'externalités négatives (eg : hausse d'impôts et lien avec les départs fiscaux). Inversement, si l'on souhaite avoir une action positive, cela permet aussi de déterminer dans quelle mesure l'action A aura bien une influence sur l'action B (eg : augmentation du budget dans l'éducation et hausse du niveau des élèves).

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

La méthode des moindres carrés permet de tracer la ligne de correspondance d'un nuage de point. Son principe est de faire d'abord apparaître un nuage de point, puis il faudra calculer la position des points de cette droite en cherchant les points qui minimisent les résidus (écarts entre la droite et les points observés) grâce à de l'algèbre, et effectuer trouver les éléments suivants :

1. moyenne des valeurs X et Y
2. pente
3. ordonnée à l'origine
4. constante (b)

Une fois ces étapes réalisées il sera possible de tracer la meilleure droite de régression linéaire. Cette méthode permet de s'économiser les calculs de la différence des carrés à partir des nombreuses droite tracée au hasard qui devraient être faits pour trouver la meilleur droite. De plus, cette méthode est également plus précise.

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

La théorie de la corrélation est une méthode permettant de faire un lien entre deux variables. Plus précisément, grâce à la théorie de la corrélation il sera possible d'observer le degré de relation entre deux variables, offrant une indication de la manière dont elles varient ensemble. Sur les deux variables, l'une sera la variable explicative, et l'autre la variable expliquée. C'est la variation de la variable explicative qui se répercute sur la variation de la variable expliquée. Pour mesurer son influence il faudra observer deux caractéristiques de la variation : sa direction et son intensité. La variable expliquée pourra : varier positivement ou négativement par rapport à la variable explicative ou inversement, et varier intensément ou peu (voir pas). Ces deux caractéristiques sont rassemblées sous le nom de coefficient de corrélation qui est compris entre -1 et 1. Pour modéliser visuellement ces relations il est possible d'utiliser des diagrammes de dispersion. Il sera ainsi possible d'observer si les points s'alignent sur une droite, une courbe, ou sont dispersés de manière aléatoire. Enfin, il est impératif de noter que corrélation ne signifie en aucun cas causalité.

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

Il y autocorrélation est quand une variable est corrélée à cette même variable mais présente dans un autre état (par exemple passé ou futur). Il peut également avoir une autocorrélation quand deux variables très similaires sont analysées. Plus deux variables sont proches plus leurs liens seront importants : il y a donc un piège car l'explication du lien paraîtra logique mathématiquement mais ne le sera pas méthodologiquement. C'est notamment le cas lorsque des données trop proches spatialement sont corrélées.

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

La régression linéaire est une technique permettant d'établir un modèle mathématique entre une variable dépendante et une ou plusieurs variables indépendantes. Elle est utilisée pour contrôler, prévoir puis décider. Pour déterminer une régression linéaire il est nécessaire de déterminer la pente et l'ordonnée à l'origine, qui définissent respectivement l'intensité et la direction de cette relation. Les paramètres du

modèle sont estimés en minimisant la somme des carrés des écarts entre les valeurs observées et celles prédites par la droite de régression. La régression linéaire va plus loin que la théorie de la corrélation car grâce à elle on détermine une équation mathématique qui décrit comment une variable dépendante varie en fonction d'une variable indépendante : elle permet donc de faire des prédictions et d'estimer l'impact d'une variable sur l'autre.

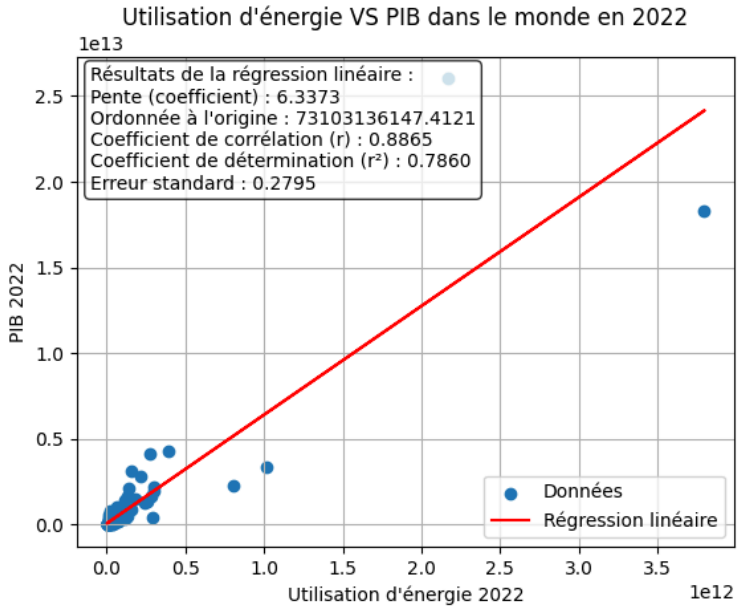
10. Quelle est la différence entre coefficient de corrélation et coefficient de détermination?

Le coefficient de corrélation met en avant le degré de variation d'une variable explicative lorsque varie une variable expliquée. Le coefficient de détermination met lui en avant la contribution de la variable expliquée à la variation de la variable explicative.

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

Comme mentionné plus haut, corrélation ne signifie pas causalité. Ainsi, proposer de mettre une variable explicative plutôt qu'expliquée dépend surtout du contexte des hypothèses décidées. C'est pourquoi tester deux droites de régression permet de varier de prisme et vérifier l'influence de chaque variable. De plus cela permet de minimiser les possibles erreurs vu que les deux droites optimisent des variables différentes.

Mise en œuvre avec Python :



Ce graphique présente la consommation d'énergie et le PIB, dans le monde, en 2022 On observe une corrélation positive et intense entre la consommation annuelle d'énergie et la taille du PIB. Le coefficient de corrélation, qui indique le degré de variation d'une variable explicative lorsque varie la variable expliquée, vaut +0,8865. Cela signifie qu'une variation de +1% de la consommation d'énergie est corrélée à une variation de +0,88% du PIB. Ensuite, le coefficient de détermination, qui lui indique une

contribution de la variable expliquée à la variation de la variable explicative, vaut +0,786. Cela signifie que 78,6% de la variance du PIB est expliquée par la variance de la consommation d'énergie.

Transposé au réel, cela montre une corrélation entre la consommation d'énergie et l'évolution de son PIB. Cela n'est pas étonnant car pour produire des biens ou fournir des services il sera impératif de consommer de l'énergie. Cette relation a déjà été démontrée par de nombreux chercheurs, comme J-M Jancovici qui a réalisé cette démonstration pour alerter les pouvoirs publics sur la diminution de la disponibilité des ressources énergétiques, ce qui pourrait amener nos économies à de fortes contractions du PIB dans les décennies à venir.

VI. Séance 8

Questions de cours :

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

Selon moi la corrélation entre deux variables qualitatives a un sens : les variables statistiques quantitatives ne permettent pas de tout expliquer. En sciences sociales l'analyse de données qualitatives redonne un côté humain que les variables quantitatives ne permettent pas d'expliquer à elles seules. Un bon exemple pour moi aujourd'hui est l'économie, qui est une science sociale, mais qui se repose fortement sur des données statistiques qualitatives : cela ne permet donc pas de capter de nombreuses informations pourtant extrêmement précieuses pour tirer des conclusions plus fines. Enfin, l'utilisation de variables qualitatives nécessite une approche scientifique et peut être analysée statistiquement grâce à différentes méthodes. Par exemple, une fois un tri à plat réalisé, les liens possibles entre deux variables qualitatives seront observables mathématiquement. Au-delà d'avoir un sens je dirais donc que la corrélation entre deux variables qualitatives est nécessaire.

2. Pourquoi pratiquer le test d'indépendance du $\chi 2$?

Ce test permet de rejeter l'hypothèse d'indépendance entre deux variables qualitatives. Comme pour les autres tests il permet de connaître les biais et liens entre des variables afin de mieux maîtriser la méthodologie de l'étude.

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 sert à déterminer s'il existe des différences importantes entres des moyennes. Elle permet de décomposer la variabilité totale des données en :

- une part expliquée par la variable explicative
- et une part expliquée par des fluctuations aléatoires

Une fois ces deux déterminées il sera possible d'utiliser le test de fisher pour tester si toutes les moyennes sont égales et rejeter ou non l'hypothèse.

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

Un rapport de corrélation est l'étude de la variation d'une variable qualitative par rapport à une variable quantitative. Ce rapport est compris entre 0 et 1 : s'il vaut 1 alors il existe presque une dépendance en moyenne, en revanche, si il vaut 0 alors il n'existe presque pas de dépendance en moyenne. La différence est que ce rapport porte sur la moyenne et sur la liaison entre une variable quantitative et une variable qualitative

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

L'analyse factorielle est une méthode statistique permettant de trier des jeux de données complexes afin de les rendre plus lisibles. Pour ce faire l'objectif est de transformer le jeu données en différents tableaux dit simples qui sont le produit de facteurs simples (à partir desquels il sera possible de produire des graphiques. Ces différents tableaux, aussi appelés matrice, seront mis facteur.

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

L'analyse factorielle des correspondances (AFC) se fait systématiquement sur deux variables qualitatives. Elle consiste à transformer un tableau difficilement analysable en différentes matrices (tableaux) dites simples. Cette méthode permet de hiérarchiser l'information contenue dans un tableau de données. Ainsi il sera plus simple d'observer les irrégularités (valeurs aberrantes) ou mettre en évidence des combinaisons entre les variables. Pour vérifier les liens entre les variables, le test du khi2 sera utilisé.

Mise en œuvre avec Python :

Tableau de contingence, avec calcul des marges

	Femmes	Hommes	**Total**
Agriculteurs exploitants	94	273	367
Artisans, commerçants et chefs d'entreprise	661	1295	1956
Cadres et professions intellectuelles supérieures	2889	3797	6686
Professions intermédiaires	3918	3511	7429
Employés	5770	1816	7586
Ouvriers	1193	4638	5831
Chômeurs n'ayant jamais travaillé	167	166	333
Inactifs	13566	10645	24211
Non classés	60	63	123
Total	28318	26204	**54522**

Test d'indépendance du χ2

	Indépendance du χ2
statistique	4812.42
p-value	0.0

Le test du chi carré est une méthode statistique fondamentale qui évalue s'il existe une association significative entre deux variables catégorielles. Pour l'utiliser il faut faire deux hypothèses :

- H0 = il n'existe aucune relation entre les variables
- H1 = il existe une relation significative entre les deux variables

Le seuil pour déterminer s' il faut accepter ou rejeter H0 est 0,05. Donc, si la p-value est supérieure à cette valeur alors H0 sera acceptée.

Dans notre cas, p-value = 0,00. Ainsi H0 est rejetée, et nous pouvons conclure qu'il existe une relation significative entre les deux variables. Ainsi dans le cas de notre étude nous observons bien un lien entre la catégorie socioprofessionnelles et le sexe biologique.

Coefficient de Pearson

	Liaison φ2 de Pearson
Statistique	0.87
P-value	0.002

Le coefficient de corrélation de Pearson indique la force de la relation linéaire entre deux valeurs. Il prend une valeur comprise entre -1 et +1 :

- +1 : Corrélation positive parfaite (les deux valeurs augmentent simultanément)
- -1 : Corrélation négative parfaite (l'une augmente, l'autre diminue)

Dans notre exercice, le coefficient de Pearson obtenu est +0,87 : il est positif et se rapproche fortement. Il possible de conclure de qu'il y a une relation linéaire assez forte entre la catégorie socioprofessionnelles et le sexe biologique, bien que cette association ne soit pas parfaite (car pas égale à +1).

VII. Bonus

Questions de cours :

Mise en œuvre avec Python :

VIII. Commentaires personnels sur le module *Analyse de données*

Je suis conscient qu'à cause de l'organisation terrible du service informatique de la fac, les séances de présentation de github, python , et du module Analyse de données plus largement ont dû être accélérées. Je pense que ce sont vraiment des séances cruciales qui permettent aux étudiants de monter en douceur dans le train du code et des statistiques. Au vu de la situation, je pense que cette entrée en matière a été assez chaotique pour nombreux d'entre nous. Voici quelques retours qui pourraient selon moi permettre de la faciliter.

J'ai pu rencontrer deux difficultés majeures : mon organisation pour apprendre, et utiliser python avec les statistiques pour la géographie.

- La première est l'organisation pour travailler tout du long du semestre. j'ai en effet beaucoup travaillé au début et à la fin mais que très peu au milieu. Je ne sais pas si cela m'est propre, mais si ce sentiment est généralisé il pourrait être pertinent de rajouter des rendus ou diviser le dossier en plusieurs rendus.
- La seconde, ma difficulté à passer de simple codage python à python adapté aux statistiques vient selon moi du fait que les ressources sur github sont trop orientées sur les concepts mathématiques et très peu sur python. Or, pour la plupart des étudiants nous n'avons pas de compétences particulières dans ces deux domaines. Il serait donc pertinent que des ressources pédagogiques sur le code en lui-même soient publiées car j'ai trouvé pour ma part que leur présence ou mise en avant sur le github était vraiment marginale par rapport aux ressources mathématiques. Dans mon cas, grâce aux vidéos youtube qui étaient recommandées j'ai pu maîtriser et comprendre quelques bases de python, mais une fois que j'ai souhaité réaliser les séances notées cela a été beaucoup compliqué étant donné que ce n'étaient pas vraiment les mêmes exercices. Je suis en groupe intermédiaire mais j'ai du commencer par les séances 2 et 3 car je ne comprenais pas du tout comment réaliser la 4. Ce qui me posait des difficultés était par exemple la réalisation de graphiques, mais dont la consigne n'était objectivement pas très claire pour les séances du groupe intermédiaire. De plus, la réalisation de graphiques n'était pas du tout expliquée dans les vidéos youtube. Etant donné que c'est un point crucial des compétences que nous devons développer en python, je pense qu'il serait pertinent que la méthode pour réaliser des graphiques soit expliquée dans le TD 4 comme elle l'est dans le TD 2.

Pour faciliter la montée en compétences des futurs étudiants, il serait pertinent selon moi de prendre plus de temps pour expliquer aux étudiants le fonctionnement du github, notamment les ressources pédagogiques qui s'y trouvent. Je trouve qu'il y a trop de documentation pour ce que nous en faisons, ou la documentation et son organisation devraient nous être présentées. Par exemple, je suis tombé par hasard sur les vidéos youtube qui étaient recommandées, mais quand j'ai souhaité délibérément les retrouver, impossible de remettre la main dessus…

En toute transparence, j'ai utilisé de façon assez récurrente des intelligences artificielles (le chat et chat-gpt) quand je bloquais trop longtemps sur des séances ou que je comprenais pas d'où venait les

erreurs. J'ai souhaité m'investir dans ce module car je suis conscient de l'importance de python et de l'informatique dans les sciences humaines. J'ai donc farfouillé dans le github dans les nombreuses ressources afin d'essayer comprendre au mieux le fonctionnement de ce langage informatique. Toutefois je n'en ressors pas nécessairement satisfait car malgré mon investissement intense (autour de 100h je pense) je ne me pas du tout automne en codage.

La raison qui justifie cela est que selon moi il y n'a pas de ressources pédagogiques expliquant clairement les premières étapes pour apprendre à coder. Si cela était mis en avant je pense que ça permettrait à tous les étudiants de gagner un temps fou pour comprendre très rapidement quelque chose de très simple, dont j'ai la sensation que nous avons mis un certain temps à comprendre (exemple : on doit coder dans main.py et exécuter dans le terminal). Un petit guide de quelques pages avec des exemples précis (à l'instar des pdf de statistiques qui eux sont très fournis et pratiques !) serait vraiment adapté. De plus, si nous avions à disposition des petits exercices (non notés) qui nous permettent de nous entraîner sur les basiques de python, et les points qui sont abordés dans les séances cela permettrait aux étudiants de pouvoir s'exercer et de ne pas utiliser un concept seulement lors d'une séance (et de l'oublier juste après). Dans mon cas j'ai cherché des exercices sur youtube ou des sites d'universités. Toutefois de nombreux exercices pythons n'avaient strictement rien à voir avec l'application aux statistiques et la géographie, d'où ma frustration et mon sentiment de ne pas avoir tant progressé que ça. Ainsi, si vous proposiez des ressources pour s'entraîner adaptées à la géographie et aux concepts utilisés (sans forcément les corriger ni même les noter) cela serait vraiment utile pour les étudiants.

En ce qui concerne les maths, je n'ai pas grand chose à redire, les fichiers sont vraiment auto-portants et clairs. Certains concepts sont compliqués à appréhender, mais pas impossible notamment en faisant ses propres recherches. Enfin la communication sur discord est finalement très pratique et permet vraiment d'échanger entre étudiants.

J'ai fait de nombreux retours pas forcément positifs mais pour ma part je garde un bon souvenir de cette aventure python. Je pense faire en janvier un mooc pour continuer de monter en compétences pour tirer profit de ce que j'ai pu apprendre ce semestre et tenter d'éclaircir les nombreux points qui sont encore flous.

Merci pour votre investissement intense tout au long de ce semestre !