M1GEANDO

Analyse de Données

Rapport d'Activité. Parcours débutant.

Sorbonne Université 2025/2026

Préambule

J'ai choisi le parcours débutant par pragmatisme vis-à-vis de mes capacités de code. En terme de statistique je n'étais pas trop perdu, par contre je n'ai pas pu finir les exercices de code – à vrai dire je n'ai pas pu finir la séance 2. La raison est que je ne voyais pas l'intérêt de faire du code sans comprendre ce que je faisais (i.e. en utilisant l'IA) et que la quantité de travail des autres matières ne me laissait pas le temps de m'y consacrer à 100%.

Ainsi, ce rapport ne contient de code que pour la séance 2.

Table des matières

Séance 2. Principes Généraux de la Statistique.

Cette première séance constituait une entrée en matière vis-à-vis de l'utilité des notions de statistique lors d'études et recherches géographiques avec une mise en application des notions statistiques de base en python.

Questions de cours

1. ***Quel est le positionnement de la géographie par rapport aux statistiques ?***

Au cours de son évolution depuis une "science de la terre" vers l'étude du "rapport des sociétés humaines à leurs espaces" (Géoconfluences, 2023), la discipline géographique a pu adopter plusieurs positions vis-à-vis de la science statistique : progressivement, certains éléments de science dite "dure" comme la géologie et la cartographie ont été mis de côté au profit de caractéristiques de science dite "molle" comme les enquêtes sur le terrain, études des rapports sociaux et autres. À l'heure de l'émergence des "humanités numériques", soit la pratique des sciences humaines par l'intermédiaire d'outils informatiques, les apports mathématiques de sciences telles que la science statistique semblent avoir le "vent en poupe".

2. ***Le hasard existe-t-il en géographie ?***

Il n'est pas possible de mesurer un certain nombre de variables (aussi appelées facteurs géographiques) sans prendre en compte une part de hasard. On considère néanmoins la part de hasard dans quelque étude assez minimale de sorte que dans la plupart des cas une certitude globale puisse émerger.

3. ***Quels sont les types d'information géographique ?***

L'information géographique peut être compartimentée en données issues de la géographie humaine (population, variables sociales, économiques, ..) ou physique (relevés météorologiques, minéraux, ..) mais peut aussi relever d'ensembles délimités (clusters, ...).

4. ***Quels sont les besoins de la géographie au niveau de l'analyse de données?***

C'est à partir des relevés de plusieurs organismes (laboratoires, instituts de géologie, topographie, météorologie, sociologie, …) que l'analyse de données en géographie se

structure. C'est après avoir établi une nomenclature (c'est-à-dire avoir ciblé le ou les objets d'étude) et avoir pris en compte une analyse critique des sources (via les « méta-données » : méthodologie du relevé, pertinence de la spatialité et temporalité du relevé, pertinence des résultats, nomenclatures utilisées, ..) que l'on pourra effecuter une analyse de données pertinente.

5. ***Quelles sont les différences entre la statistique descriptive et la statistique explicative ?***

La statistique descriptive met en lumière certaines tendances et propriétés dites « remarquables ». Elle permet d'effectuer des analyses rapides. C'est aussi la première étape d'une l'analyse de données.

La statistique explicative intervient à la suite de la statistique descriptive : elle applique une ou plusieurs fonctions mathématiques (notamment de probabilité) pour extrapoler une tendance générale ou plusieurs scenari possibles à partir de l'indicateur constitué par la statistique descriptive.

6. ***Quelles sont les types de visualisation de données en géographie ? Comment choisir celles-ci ?***

Les données sont manifestées en géographie sous la forme de graphiques (histogrammes, boîtes de dispersion, ...) ou de tableaux de synthèse. Dans tous les cas il est nécessaire de mentionner (entre autres) la méthode de calcul, la nomenclature des données et les intervalles de confiance.

7. ***Quelles sont les méthodes d'analyse de données possibles ?***

Parmi les méthodes d'analyse de données, il y a d'abord la méthode dite descriptive : il s'agit de mettre en évidence les relations ou l'absence de relations entre plusieurs variables, aboutissant le plus souvent à une classification en ascendance hiérarchique (CAH) ou en nuées dynamiques.

Les méthodes explicatives consisteront en une confrontation d'une ou plusieurs variables explicatives avec une variable à expliquer, en d'autres mots il s'agit de faire ressortir différents groupes ou tendances au sein d'un ensemble de données descriptives. Enfin, les méthodes prédictives inscrivent l'ensemble des variables dans une échelle de temps et indiquent les scenari les plus probables.

8. ***Comment définiriez-vous : (a) population statistique ? (b) individu statistique ? (c) caractères statistiques ? (d) modalités statistiques ? Quels sont les types de caractères ? Existe-t-il une hiérarchie entre eux ?***

La population statistique est un ensemble mathématique, en d'autre mots une quantité, un effectif à partir duquel une étude statistique peut être réalisée.

L'individu statistique (ou unité statistique) est un élément unique et précis de ladite population, qu'il soit primaire (un « atome », un élément indivisible, un être humain par exemple) ou secondaire (une « molécule », une agrégation d'unités primaires, par exemple une communauté de communes).

Les caractères statistiques sont l'ensemble de caractéristiques relevées sur l'individu statistique, les modalités statistiques étant les valeurs attribuées à chaque caractère de chaque individu d'une population statistique. Elle peuvent être quantitatives (dénombrables) ou qualitative (indénombrables). Chacune est utilisée dans des contextes différents mais l'utilisation de caractères qualitatifs nominaux induit souvent l'utilisation de modalités quantitatives discrètes (il s'agira alors de compter les effectifs de fréquence de caractères).

9. ***Comment mesurer une amplitude et une densité ?***

Les mesures d'amplitude et de densité s'effectuent au sein d'une "classe statistique" (l'intervalle entre deux valeurs *a* et *b* d'une série statistique). Pour la première, il s'agira de calculer une "étendue" mathématique soit la soustraction de la valeur minimale à la valeur maximale. Le calcul de la densité d'une classe se fera par la division de son effectif par son amplitude.

10. ***À quoi servent les formules de Sturges et de Yule?***

Les formules de Sturges et de Yule permettent d'extrapoler à partir de l'ensemble de valeurs d'un caractère son découpage en classes statistiques idéal.

11. ***Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?***

Un effectif se définit comme le nombre de valeurs recueillies pour une variable dans une population statistique.

La fréquence relative d'une variable se calcule en divisant l'effectif de la variable étudiée par l'effectif total (effectif de l'ensemble des variables). Les fréquences cumulées se calculent par la division de la somme d'un nombre *k* d'effectifs par l'effectif total susmentionné.

Une distribution statistique est l'association de classes statistiques à leur fréquence d'apparition afin d'obtenir une vision générale des probabilités de reproductions ou non de certaines valeur.

Exercice de code

Il s'agissait pour les premières questions (5-8) d'une prise en main du vocabulaire de la bibliothèque pandas avec python. J'ai pu afficher avec la question 5 le *dataframe* issu des données de vote aux élections présidentielles en France en 2022 (Fig.1).

	Code du département	Libellé du département	Inscrits	...	Nom.11	Prénom.11	Voix.11
0	01	Ain	438109	...	DUPONT-AIGNAN	Nicolas	8998.0
1	02	Aisne	373544	...	DUPONT-AIGNAN	Nicolas	5790.0
2	03	Allier	249991	...	DUPONT-AIGNAN	Nicolas	4216.0
3	04	Alpes-de-Haute-Provence	128075	...	DUPONT-AIGNAN	Nicolas	2504.0
4	05	Hautes-Alpes	113519	...	DUPONT-AIGNAN	Nicolas	2142.0
..	...	...	...	...	...	...	...
102	ZP	Polynésie française	205576	...	DUPONT-AIGNAN	Nicolas	1969.0
103	ZS	Saint-Pierre-et-Miquelon	5045	...	DUPONT-AIGNAN	Nicolas	82.0
104	ZW	Wallis et Futuna	9528	...	DUPONT-AIGNAN	Nicolas	244.0
105	ZX	Saint-Martin/Saint-Barthélemy	24414	...	DUPONT-AIGNAN	Nicolas	339.0
106	ZZ	Français établis hors de France	1435746	...	DUPONT-AIGNAN	Nicolas	7074.0

[107 rows x 56 columns]

Fig.1 : dataframe des données de vote aux élections présidentielles en France (2022)

La question 6 formalisait le nombre de lignes et colonnes du tableau (Fig.2).

```
Le tableau contient 107 lignes et 56 colonnes.
```

Fig.2 : Indication des lignes et colonnes du dataframe.

La question 7 permettait d'isoler le type de variable propre à chaque colonne (Fig.3)

	Inscrits
0	438109
1	373544
2	249991
3	128075
4	113519
..	...
102	205576
103	5045
104	9528
105	24414
106	1435746

Fig. 3 : exemple de colonne isolée pour la question 7 (données quantitatives discrètes, type Int)

La question 8 permettait d'afficher le nom de chaque variable en tête des colonnes (Fig.4).

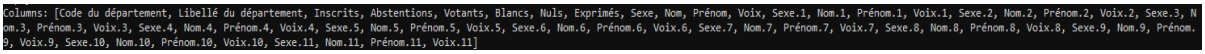

Fig.4: Nom de chaque variable du dataframe

Les questions 9-10 permettent, par l'intermédiaire d'une boucle et en utilisant des données précédemment récupérées dans le code, d'obtenir les effectifs du type de variable souhaité (ici il s'agissait de variables quantitatives discrète ou continues, Fig.5).

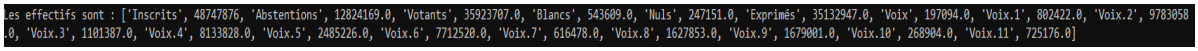

Fig.5 : Nom et effectifs des variables statistiques quantitatives discrète ou continues du dataframe.

La question 11 consistait en une utilisation des formules *matplotlib* de création de graphiques en barre par l'intermédiaire d'une boucle à condition qui permettait de prendre en compte l'entièreté des caractères des variables « Inscrits » et « Votants » en rapport avec la série statistique des départements (Fig.6).

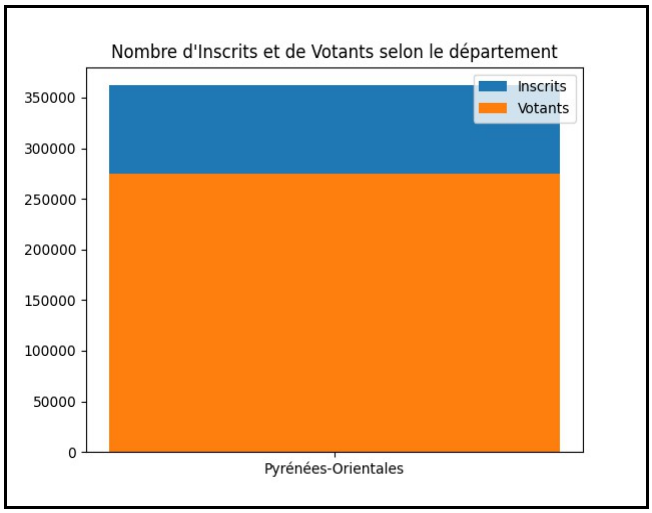

Fig.6 : Exemple de graphique en barre créé avec la question 11.

Ce graphique représente le nombre d'inscrits et de votants (variables quantitatives) par département (variable qualitative) lors des élections présidentielles françaises de 2022. On peut observer qu'il y avait environ 270,000 votants dans les Pyrénées Orientales pour environ 360,000 inscrits, soit une proportion de votants à environ 75%.

La question 12 permet de mettre en graphique circulaire la proportion de votants selon les inscrits et selon le type de vote (Fig.7).

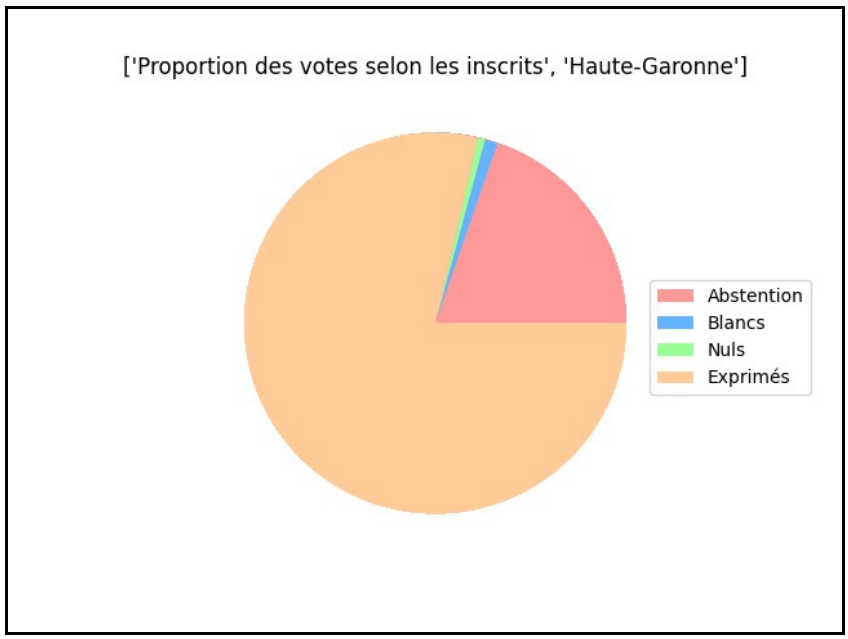

Fig.7 : Exemple de graphique circulaire généré à partir du dataframe

On peut ainsi trouver qu'un peu plus des trois quarts des inscrits ont exprimé un suffrage en Haute Garonne lors des élections présidentielles de 2022.

La question 13 permet d'obtenir à partir d'une formule matplotlib la distribution statistique de la variable des inscrits par département (Fig.8).

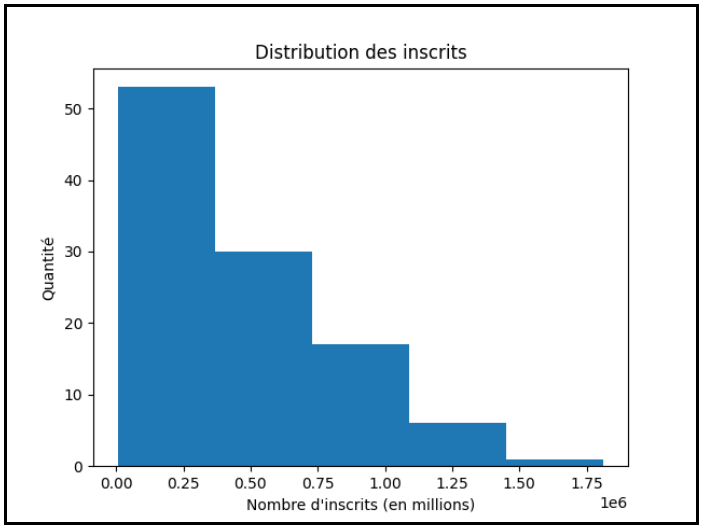

Fig.8 : Histogramme de la distribution des inscrits dans les départements de France en 2022

Ce graphique permet de visualiser que la majorité des départements français recensent moins d'1 million d'inscrits. On peut si on le souhaite modifier le nombre de

divisions de l'histogramme pour obtenir une information plus précise (Fig.9).

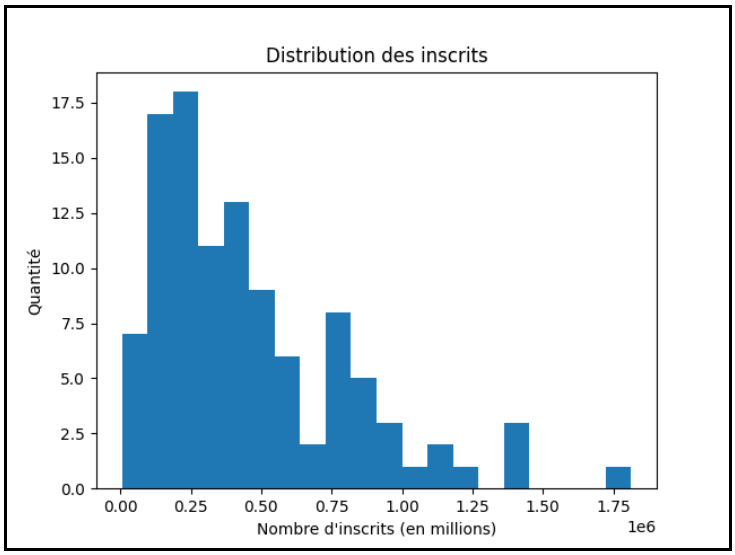

Fig.9 : Histogramme de la distribution des inscrits dans les départements de France en 2022 (20 subdivisions).

Ainsi un histogramme plus précis permet de voir apparaître plusieurs valeurs extrêmes et isolées comme celle du département ayant environ 1,75 millions d'électeurs lorsque tous les autres en ont moins d'1,5 millions.

Vis-à-vis du fichier code, malgré mes tentatives de résoudre le problème je n'ai pas réussi à lancer les trois formules *matplotlib* en même temps : celles-ci interagissent les unes avec les autres résultant en des graphiques erronés et illisibles.

Séance 3. Paramètres statistiques élémentaires.

Il s'agissait dans cette séance de s'intéresser aux paramètres statistiques et d'ainsi découvrir des outils statistiques plus avancés que la médiane, l'étendue ou la moyenne. L'exercice de code python était une mise en pratique de ces paramètres en programmation.

Questions de cours

1. ***Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ? Justifier pourquoi.***

À première vue il semble qu'une étude quantitative soit plus générale ou du moins « généralisable » car elle permet d'effectuer une analyse de données rapide et à grande échelle sans trop de difficulté. Ceci étant dit, l'étude qualitative si recodée peut être traitée de la même façon. Ainsi, afin de pouvoir obtenir des conclusions « générales » vis-à-vis d'une étude statistique, la combinaison des deux types de caractère doit le plus souvent être privilégiée.

2. ***Que sont les caractères quantitatifs discrets et caractères quantitatifs continus ? Pourquoi les distinguer ?***

Les caractères quantitatifs discrets sont généralement des effectifs ou autres nombres exclusivement entiers à l'intérieur d'une certaine intervalle (par exemple le nombre d'enfants par famille). Les caractères quantitatifs continus correspondent à des nombres possiblement décimaux ou en pourcentage au sein d'une certaine intervalle tels qu'un taux de fécondité ou qu'une taille. Il s'agit de les différencier afin de ne pas produire de valeurs incohérentes (par exemple 2,5 personnes) et de bien identifier les sujets d'étude et d'analyse.

3. ***Pourquoi existe-t-il plusieurs types de moyenne ? Pourquoi calculer une médiane ? Quand est-il possible de calculer un mode ?***

Il existe plusieurs types de moyennes (arithmétique, quadratique, harmonique, géométrique, mobile, fonctionnelle) afin de traiter différents cas de figure efficacement (ne pas surestimer la moyenne réelle, calculer une vitesse moyenne, calculer des écarts-types, …). Les données rangées en ordre croissant ou décroissant, on calculera plutôt une médiane ou un écart médian (valeur x qui divise la série de données en deux parts égales) afin de

visualiser plus clairement la distribution d'une série sans être influencé par les valeurs « aberrantes ».

Le mode correspond à la fréquence d'apparition la plus élevée parmi les valeurs d'une série. Il n'est pas calculable si les valeurs apparaissent toutes à la même fréquence mais peut être multiple si plusieurs variables apparaissent à la même fréquence maximale.

4. ***Quel est l'intérêt de la médiale et de l'indice de C. Gini ?***

La médiale et l'indice de Gini mettent en valeur la concentration d'une population statistique, la première en séparant l'effectif en deux groupes où la somme de la valeur associée à chaque individu statistique dans chacun des deux groupes est égale, la seconde en visualisant sur un repère orthonormé le partage de la massa totale des valeurs selon l'éloignement ou non de la diagonale principale.

5. ***Pourquoi calculer une variance à la place de l'écart à la moyenne ? Pourquoi la remplacer par l'écart type ? Pourquoi calculer l'étendue ? À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s)? Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?***

L'intérêt d'une variance par rapport à l'écart à la moyenne est celui d'un calcul de la dispersion qui prend en compte l'ensemble des variables (écart à la moyenne, pondération, …). L'écart type peut être privilégié comme mise en forme à la variance puisque sa lecture permet rapidement d'établir le rapprochement ou l'éloignement des données vis-à-vis de la moyenne arithmétique.

L'étendue est rarement prise en compte mais sa facilité de calcul permet d'établir rapidement les marges de la série de donnée.

Le calcul des quantiles permet d'obtenir une visualisation rapide de la distribution de la donnée statistique de part et d'autre de la médiane, on utilisera la plus souvent les second et troisième quartiles (respectivement valeur médiane et valeur à 75% de l'effectif de la série statistique ordonnée en ordre croissant).

Une boîte de dispersion permet de représenter graphiquement une distribution statistique de caractères quantitatifs et d'ainsi identifier médiane et quartiles facilement (on peut aussi y inscrire la moyenne) pour faciliter une comparaison avec d'autres distributions

statistiques. La « boîte » en elle-même représente les valeurs situées entre le premier et le troisième quartile. La médiane est représentée au sein de la boîte par un segment qui la sépare en deux, elle se situe à l'endroit du second quartile. Les deux extrémités de la « boîte à moustache » représentent les valeurs minimale et maximale. On pourra inscrire le premier et le neuvième décile sur les « moustaches ».

6. ***Quelle différence faites-vous entre les moments centraux et les moments absolus ? Pourquoi les utiliser ? Pourquoi vérifier la symétrie d'une distribution et comment faire ?***

Les moments centraux correspondent à la caractérisation d'une distribution telle que l'espérance de la variable (moment d'ordre 1), sa variance (moment d'ordre 2), la symétrie de la distribution (moment d'ordre 3) ou le coefficient d'aplatissement (moment d'ordre 4). Les moments absolus mettent en relation la variance et l'espérance d'une distribution statistique.

Ce sont des outils pratiques afin d'établir par différents indicateurs l'écart d'une distribution statistique vis-à-vis de la moyenne.

La vérification de la symétrie d'une distribution par l'utilisation d'un coefficient d'asymétrie comme celui de Pearson-Fisher (il s'agit de diviser le moment centré 3 par le cube de l'écart-type de la distribution). L'intérêt de cette vérification est d'obtenir, selon le résultat, un indicateur de la distribution des valeurs égale ou inégale de part et d'autre de la médiane.

Séance 4. Distributions statistiques.

Il s'agit dans cette séance de s'intéresser aux lois de distribution statistique et à leur utilité en géographie : calcul de fréquences d'apparition, de probabilité de catastrophes naturelles, ...

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 ?***

Une distribution statistique à partir de variables discrètes est une distribution statistique composée d'un ensemble de caractères discrets, c'est-à-dire de valeurs en nombre absolus (des effectifs par exemple). Une distribution statistique à partir de variables continues est, à l'inverse, une distribution statistique composée d'un ensemble de caractères continus, c'est-à-dire de nombre réels pouvant être décimaux (taux, moyennes, températures). Les deux sont des distributions à partir desquels des calculs de statistique univariée à multivariée peuvent être produits (à l'inverse de variables qualitatives).

Afin de mieux choisir entre l'une et l'autre il s'agit de prendre en compte la nature du phénomène étudié, en d'autres mots si les caractères de la distribution peuvent être considérés discrets ou continus. Il s'agit aussi de prendre en compte la forme que prend la distribution lorsque projetée sur un graphique, à partir de laquelle on pourra utiliser une loi statistique spécifique. Il est aussi pratique de connaître les paramètres statistiques de la distribution étudiée (écart type, variance, médiane, espérance, …). Enfin, le choix dépendra aussi de l'applicabilité d'un certain nombre de paramètres vis-à-vis des lois statistiques envisagées pour analyser la distribution statistique.

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

Parmi les lois les plus utilisées en géographie on peut retrouver les lois discrètes de Bernoulli (qui permet d'obtenir une probabilité à partir d'une réponse binaire, soit oui ou non dans un questionnaire, présence ou absence sur le terrain, etc...), binominale (qui permet d'obtenir une probabilité de succès à n tests par la loi de Bernoulli), de Poisson (qui permet de visualiser le comportement d'événements inscrits spatialement ou dans une temporalité spécifique c'est-à-dire si un événement se produit plus fréquemment dans un endroit ou à une heure précise), géométrique (qui permet de visualiser le temps écoulé avant l'apparition

d'un événement spécifique).

On peut aussi retrouver les lois continues normale (qui permet de visualiser la fréquence d'apparition d'événements dans un espace et d'en ainsi obtenir une probabilité), de Cauchy (qui met en avant les valeurs les plus extrêmes d'une distribution statistique et permet d'obtenir une perspective plus large de la distribution statistique), Gamma (dont khi2 ou Fisher-Snedecor qui permettent de visualiser le degré de corrélation de deux variables indépendantes), de Benford (lorsque les fréquences d'une distribution statistique sont décroissantes sur une échelle logarithmique), d'extremum généralisé (Gumbel, Weibull, Fréchet, surtout utile en géophysique pour connaître la probabilité d'événements extrêmes) ainsi que celle de Pareto (qui permet d'obtenir la fréquence d'apparition d'une variable par rapport à une autre variable continue).

Séance 5. Statistiques inférentielles.

Il s'agit dans cette séance d'explorer les enjeux de l'échantillonnage et d'autres outils tels que les tests statistiques pour arriver à une conclusion fiable sur une population statistique donnée.

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 ?***

Il est toujours nécessaire pour l'étude d'une population statistique dont l'ensemble n'est pas interrogeable (c'est l'exemple de poissons dans l'océan, d'arbres en Amazonie, etc, ...) d'en définir et prélever un échantillon, c'est-à-dire une partie (un « sous-ensemble ») plus ou moins grande de ladite population qui, à partir de paramètres et lois de distributions statistiques, pourra être extrapolée pour effectuer des généralisations sur la population entière.

Si on dispose d'une base de sondage (soit d'une idée au préalable des caractéristiques de la population), on peut effectuer un échantillon représentatif de la population (par exemple pour la population française dont on connaît la répartition des activités professionnelles on pourra sélectionner un échantillon de 1,000 personnes dont un pourcentage représentatif sera issu de chaque profession). Sinon, il est nécessaire d'effectuer un échantillonnage aléatoire, un prélèvement soit non-biaisé (pur hasard : tirage) ou biaisé (à partir de critères choisis ou décidés au hasard).

La taille de l'échantillon doit être assez conséquente pour en inférer des généralités sur la population totale (un échantillon de 3 personnes ne donnerait par exemple pas de conclusions généralisables avec confiance à la population française).

2. ***Comment définir un estimateur et une estimation ?***

Un estimateur est la loi statistique choisie pour analyser l'échantillon de la population statistique (loi de Poisson, loi normale, etc.). L'estimation correspond au processus de vérification de la loi statistique.

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

Puisqu'il s'agit en statistique d'obtenir des estimations fiables, d'où des conclusions

peuvent être tirées sur la population mère avec confiance, il faut justement faire intervenir des intervalles de fluctuation et de confiance.

Un intervalle de fluctuation est l'intervalle au sein duquel, selon les estimations, une valeur de l'échantillon doit se trouver avec une fréquence de 95% pour que celui-ci soit représentatif de la population mère.

L'intervalle de confiance est cette valeur, le plus souvent de 95%, nécessaire à l'obtention d'un résultat représentatif de la population mère.

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

Un biais en théorie de l'estimation correspond à la différence de l'estimation vis-à-vis de la valeur attendue. Si l'estimateur est « biaisé » on parlera « d'erreurs d'estimation ».

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

Une statistique travaillant sur la population totale à partir d'une généralisation des caractéristiques observées sur un échantillon est une statistique « inférentielle ». Elle est contrainte par la non-possibilité d'obtenir des données à grande échelle sur une population mais cette contrainte est aujourd'hui remise en question par la collecte de données massives (*big data*) par certaines entreprises de la « tech » mais aussi par des gouvernements (c'est l'enjeu d'une surveillance de masse de la population). Cette nouvelle capacité à collecter des données est néanmoins restreinte dans un contexte (ce sont des données exclusives à l'identité « numérique » des individus).

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

Le choix d'un estimateur dit « robuste » qui repose sur des statistiques exhaustives (dont la proportion d'erreur au sein des lois statistiques ont été réduites au maximum) est crucial puisqu'en dépend la fiabilité des conclusions inférées sur la population mère totale.

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

Parmi les méthodes d'estimation d'un paramètre statistique il y a la méthode des moindres carrés (qui permet avec un modèle théorique d'affiner l'estimation faite à partir de

données relevées) et du maximum de vraisemblance (qui permet de généraliser les valeurs d'un échantillon à partir d'une fonction de vraisemblance qui donne la probabilité d'appartenance des valeurs à la population totale). La première est plus utile lorsque la distribution statistique résultant de l'échantillon fournit plusieurs variables aléatoires, la seconde au contraire permet d'extrapoler des généralités à partir de peu de valeurs.

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

Un test statistique est la mise à l'épreuve d'une hypothèse statistique vis-à-vis d'un échantillon. Ainsi plusieurs manières existent d'effectuer des tests statistiques.

D'abord, on peut effectuer des tests de « conformité » ou « d'ajustement » qui comparent les résultats tirés d'un échantillon (distribution expérimentale) à un modèle théorique (distribution théorique). On peut aussi effectuer des tests « d'homogénéité » qui comparent les résultats tirés de plusieurs échantillons au sein d'une même population mère. Il existe aussi les tests « d'adéquation » (qui vérifient l'adéquation de l'échantillon à la population mère selon une loi de probabilité appropriée) et ceux de « l'indépendance » de deux caractères. Enfin, on peut comparer deux variables dites « appariées » (qui influent l'une sur l'autre) au sein d'un même échantillon.

Ces tests servent à vérifier la fiabilité des valeurs obtenues au sein d'un échantillon et d'ainsi affirmer ou infirmer des hypothèses vis-à-vis de probabilités statistiques.

On peut créer un test à partir d'une hypothèse statistique. Après l'extrapolation d'une loi statistique, d'un seuil d'erreur, et la définition de la taille de l'échantillon. Une fois les relevés effectués, on choisit le test correspondant à l'hypothèse statistique à vérifier. Deux « zones » de rejet et non-rejet peuvent être établies à partir de l'hypothèse statistique, on y comparera la « variable de décision » calculée à partir des résultats de l'échantillon. Le test statistique d'une hypothèse nulle (que l'échantillon obtient les mêmes valeurs qu'un échantillon théorique) est rejeté si ladite variable est incluse dans la zone de rejet, celui d'une hypothèse alternative (que l'échantillon obtient des valeurs différentes) est rejeté si ladite variable est incluse dans la zone de non-rejet (mais alors le test n'est pas très significatif et sa conclusion peut être imputée aux fluctuations de l'échantillonnage).

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

Les critiques de la statistique inférentielle soulignent qu'un grand échantillon rend

chaque conclusion significative (c'est vrai mais ce n'est pas un point nécessairement négatif) lorsqu'un petit échantillon rend toute conclusion non-significative (ce n'est pas nécessairement vrai si l'échantillon est représentatif ou ajusté rigoureusement).

D'autres soulignent que l'hypothèse nulle (qui souligne que l'échantillon obtient les valeurs qu'il peut théoriquement obtenir), si elle n'est pas rejetée lors de tests statistiques, est trop souvent interprétée comme juste. Généralement la critique est que les tests statistiques sont mal conçus et mal interprétés. C'est peut être vrai mais cela ne leur enlève pas leur utilité si utilisés rigoureusement et dans des cas qui en ont besoin.

Séance 6. Statistique d'ordre des variables quantitatives.

Il s'agit dans cette séance de s'intéresser aux statistiques ordonnées ou classements, notamment en ordre croissant ou ordre naturel.

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 (par opposition à une statistique nominale) est une statistique qui s'intéresse à des distributions « rangées » en ordre croissant ou décroissant (pour les variables quantitatives du plus petit au plus grand ou l'inverse). Elle peut être observée au sein de variables qualitatives (degré de satisfaction, …) mais aussi quantitatives (classement d'une distribution statistique par ordre croissant ou décroissant). L'intérêt de cette statistique en géographie se situe dans sa matérialisation dans l'espace des emplacements les plus valorisés (plus haut degré sur une échelle, plus grand nombre de valeurs ou valeurs les plus hautes, …) et à l'inverse les moins valorisés.

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

Il convient de privilégier l'ordre croissant dans le traitement des données puisque c'est à partir de celui-ci que plusieurs fonctions statistiques peuvent être appliquées (fonction de répartition de Weibull, de Tukey, …).

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

La corrélation des rangs correspond à la similarité entre deux distributions statistiques ordonnées, elle se mesure par un coefficient résultant de tests (Spearman, Kendall). On dit que les classements sont « concordants » si ceux-ci sont ordonnés de façon croissante (« ordre naturel »), ils sont « discordants » dans le cas contraire (ordonnés de façon décroissante, désordonnés).

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

Le test de Spearman vise à rapprocher différents classements dont les variables sont rattachées à un même objet d'étude, et d'ainsi vérifier si malgré les différentes caractéristiques des classements leurs variables observent une corrélation ou non.

Le test de Kendall vise de son côté à déterminer si deux ou plus classements observent les mêmes trajectoires (+1) ou s'il sont inverses (-1).

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

Le coefficient de Goodman-Kursdal souligne en proportion la différence entre le nombre de paires de classements concordants et le nombre de paires de classements discordants. Il varie entre -1 et +1, sa valeur 0 indique que les classements sont rigoureusement indépendants.

Le coefficient de Yule est un dérivé du précédent et offre la même grille de lecture, il sert à partir d'une matrice entre le nombre de paires de classements concordants et le nombre de paires de classements discordants. Sa variation de -1 à +1 indique une fois encore une association plus ou moins marquée entre les différents classements.

Réflexion sur les humanités numériques et retour vis-à-vis du cours.

Les humanités numériques constituent sans aucun doute l'une des évolutions majeures des sciences humaines au XXI^{ème} siècle et, à titre personnel, je rejoins votre diagnostique qu'il est nécessaire de s'y éduquer de manière à ne pas être laissé de côté dans un monde en changement face à des géants de la tech américains, chinois ou indiens en devenir (la France a notoirement pris du retard en n'investissant pas dans le numérique dès les années 1990s).

Ainsi, la connaissance d'outils et de leur fonctionnement tels que le bloc de commande est selon moi un enjeu de “souveraineté intellectuelle” pour la population : savoir ce qui est à l'origine des appareils qui aujourd'hui régissent nos vies est certainement essentiel pour adopter un regard critique vis-à-vis des narratifs qui nous sont proposés. La connaissance d'outils “open source” et de la façon dont les traiter tels qu'InfoClimat est source d'émancipation pour une population qui souhaite s'informer de façon indépendante. C'est aussi l'enjeu de savoir manipuler les statistiques pour, à l'heure d'une utilisation très accrue des sondages dans le débat public, pouvoir adopter une position critique.

Néanmoins, je trouve dommage que plus de temps n'ait pas été passé dans ce cours à développer un véritable apprentissage de l'outil informatique que représente python. À mon sens, de manière à ce que chaque étudiant comprenne les mécanismes de python et des bibliothèques (pandas, matplotlib, …) il aurait fallu passer plus de temps sur chaque séance et que les exercices de code soient étalés sur plusieurs séances, que vous nous accompagniez dans la réalisation du code. La façon dont nous étions “livré à nous-même” n'a pas été fructifiante pédagogiquement à mon sens si ce n'est que le peu de programmation que j'ai pu faire dans ce cours, puisque j'y ai passé un certain temps pour en comprendre les rouages, vont me rester.

Ainsi, mon propos est que le cours et les humanités numériques sont nécessaires mais, dans le contexte où la plupart des étudiants du master GAED n'avaient pas fait de mathématique depuis la terminale voire la seconde et au vu de la quantité de travail qui était demandée, je doute qu'elle ait eu une plus-value pédagogique pour une grande majorité des étudiants. Ainsi, je pense que vous devriez simplifier (peut-être à l'excès) vos séances et notamment votre cours qui, rédigé très “mathématiquement”, est aussi très opaque à certains endroits.

Je reconnais néanmoins l'effort titanesque que vous avez fourni dans l'organisation de ce cours et je vous en suis reconnaissant, seulement je pense que vos attentes étaient bien au-delà de nos capacités, en tout cas des miennes. J'essaierai éventuellement de me remettre au codage car c'est une ressource très utile, comme vous l'avez souligné dans le monde du travail, à un rythme adapté.