Analyse de données M1
Andréa Alcon-Rebelo
Parcours débutant

Rapport d'activité

- Séance 2

Questions de cours

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

Le positionnement de la géographie par rapport à la statistique peut être qualifié d'ambigu. En effet, la géographie, en tant que science humaine, a pu marginaliser la statistique en la présentant comme une discipline mathématique strictement relative aux sciences dures. Néanmoins, les méthodes de la statistique se révèlent indispensables pour traiter efficacement les données géographiques, de plus en plus massives. L'outil statistique apparaît donc finalement comme une nécessité pour la géographie, qui se présente aujourd'hui plutôt comme une discipline carrefour entre science humaine et science dure. On peut ainsi citer plusieurs géographes comme Pierre Dumolard qui ont participé à la démocratisation de l'usage de la statistique en géographie.

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

La position de la géographie à propos du hasard fait l'objet de débat parmi les géographes, autour de deux notions centrales de la philosophie du hasard : la nécessité et la contingence. Tout l'enjeu est de savoir si la géographie peut être considérée ou non comme une science ; quelle place accorder au hasard dans les études géographiques : est-il prédominant ? Au-delà du débat, on peut bel et bien affirmer que le hasard existe en géographie en raison de la nature même des objets et individus étudiés. En effet, dans le cadre de la géographie humaine, par exemple, on ne peut pas prévoir avec exactitude les faits et gestes des individus sur le terrain étudié. Néanmoins, il est possible de dégager une tendance, une "certitude globale" à l'aide de la statistique. Dès lors, il s'agit d'accepter cette part de hasard qui caractérise les objets d'étude de la géographie et qui font d'elle une science humaine s'appuyant sur des méthodes issues des sciences dures.

3. Quels sont les types d'information géographique .

En géographie, il existe deux types d'information. D'une part les informations liées à la morphologie des ensembles étudiés, à savoir celles caractérisant les données géométriques du S.I.G (polygones, mode vectoriel…). D'autre part, les informations liées à la caractérisation de ces ensembles, à savoir leurs attributs (ex: âge, sexe, professions des individus dans le cadre d'une étude de géographie humaine.).

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

Une fois les données collectées - directement sur le terrain ou auprès d'organismes publics - le géographe doit analyser les données obtenues. Cette étape s'appuie sur les méthodes et le vocable mathématique des probabilités et de la statistique, qui permettent d'établir des lois de probabilité et donc de dégager finalement une tendance pour le phénomène étudié. En matière d'analyse de données, la géographie a donc besoin de l'outil statistique pour "faire parler les données", leur donner du sens; il s'agit concrètement de tirer des résultats à partir d'un ensemble, souvent massif, de données permettant de confirmer ou non les hypothèses formulées au départ. L'outil statistique apparaît comme un outil mathématique de synthèse pour la géographie.

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

La statistique descriptive et la statistique explicative correspondent à deux méthodes différentes d'analyse de données.
La statistique descriptive consiste à appliquer une méthode au sein de laquelle toutes les variables jouent un rôle similaire; on pourrait dire qu'elle se fonde sur les concepts de "d'ensemble", "synthèse", "globalité". Cette méthode consiste à étudier de façon globale les données, à les classer, résumer, afin d'obtenir une vision simplifiée de l'ensemble des données pour saisir l'essentiel du phénomène étudié; en bref, il s'agit de décrire les données. La statistique descriptive constitue ainsi une étape préalable à l'application de méthodes statistiques plus poussées comme celles de la statistique mathématique.
La statistique explicative à quant à elle pour objectif d'établir une relation entre une variable à expliquer et des variables explicatives; une variable est alors amenée à jouer l'un ou l'autre rôle. On pourrait dire qu'elle se fonde sur le concept d'"association". Sur le plan des méthodes, la statistique explicative use de méthodes statistiques plus poussées (comme les régressions par exemple).
On remarque aussi que les méthodes de visualisation diffèrent selon que l'on fasse de la statistique descriptive ou explicative.

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

Il existe trois types principaux de visualisation des données en géographie : les visualisations graphiques (histogramme, représentation sectorielle), les tableaux de synthèse, les analyses factorielles (A.C.P ; A.F.C ; A.C.M; A.F.D.M; A.F.M ect.). On peut également visualiser des données par le biais de méthodes de classification comme la classification ascendante hiérarchique (C.A.H). Le type de visualisation est tout d'abord choisi en fonction de la nature de la variable (quantitative ou qualitative, continue, discrète); il est cependant possible grâce à l'analyse factorielle de données mixtes (A.F.D.M) et l'analyse factorielle multiple (A.F.M) de traiter un ensemble de données de nature hétérogène. Le type de visualisation est également choisi en fonction du nombre de variables.

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

Pour analyser des données il existe trois principaux ensembles de méthodes : les méthodes descriptives; les méthodes explicatives; les méthodes de prévision. Parmi les méthodes

descriptives, on retrouve les différents types d'analyse factorielle, ainsi que les méthodes de classification. Les méthodes explicatives s'appuient quant à elles sur des méthodes issues de la statistique mathématique (régressions, analyse de la variance, établissement d'un modèle linéaire général…). Enfin, les méthodes de prévision s'appuient sur la construction d'un modèle fondé sur une série chronologique et permettant d'établir un lien entre le passé et le présent.

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?

a) La population statistique est au sens mathématique un ensemble. Plus précisément, il s'agit de l'ensemble des individus statistiques. (ex: le nombre d'habitants d'une ville).
b) Un individu statistique est un élément de la population statistique, il s'agit d'une unité statistique. Autrement dit, il s'agit d'une partie du tout qu'est la population statistique. (ex: un habitant pris parmi l'ensemble des habitants d'une ville.)
c) Les caractères statistiques sont des caractéristiques, des spécificités, attribuables à un individu statistique. Un caractère statistique est donc un attribut de l'individu étudié. (ex: l'âge d'un individu, la couleur de ses cheveux…).
d) Un caractère statistique peut avoir plusieurs modalités. Une modalité d'un caractère correspond à une valeur prise par le caractère. Par exemple : la couleur des yeux (caractère statistique) d'un individu statistique peut être bleue ou marron (modalités statistiques). La modalité statistique prise par le caractère statistique d'un individu peut donc le distinguer ou au contraire l'assimiler aux autres individus statistiques de la population statistique étudiée.

Les caractères (ou types de variables) se divisent en quatre grandes catégories : les caractères quantitatifs ou qualitatifs; les caractères continus ou discrets. Plus précisément, les différents types de variables sont les suivants : les variables qualitatives nominales; les variables qualitatives ordinales; les variables quantitatives discrètes; les variables quantitatives continues.
Il n'existe pas de hiérarchie objective entre les différents types de caractères. La hiérarchie effectuée entre les différents types de variables dépend du type de résultat attendu et des méthodes appliquées; certaines méthodes, comme le calcul d'une moyenne par exemple, ne peuvent être appliquées qu'à des caractères de type quantitatif.

9. Comment mesurer une amplitude et une densité?

Tout d'abord, la mesure d'une amplitude et d'une densité s'effectue sur une classe statistique à la fois, laquelle est envisagée comme un intervalle. Pour calculer l'amplitude d'une classe statistique, il faut soustraire la valeur minimale (a) de la classe à la valeur maximale de celle-ci (b), ce qui donne la formule *b - a ;* le calcul de l'amplitude d'une classe peut être qualifiée d'étendue locale. En effet, l'amplitude peut également être mesurée à une autre échelle, à savoir à partir du nombre total de classes.
Pour calculer la densité d'une classe statistique, il faut diviser l'effectif de la classe statistique par son amplitude. (densité = effectif/amplitude).

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

Les formules de Sturges et de Yule permettent d'obtenir une valeur approximative du nombre de classes statistiques et facilitent ainsi le processus de discrétisation. Une fois le nombre de classes obtenu, il est possible de calculer l'amplitude de ces classes.

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

a) Au sein d'une variable, un effectif est le nombre total d'apparition d'une même valeur (ou modalité). Par exemple : au sein d'une classe de 30 élèves constituant une population statistique, on dénombre au sein de la variable "couleur des yeux" 10 personnes aux yeux bruns et 20 personnes aux yeux bleus, l'effectif de la modalité "yeux bruns" est donc égal à 10.
b) Une fréquence se calcule en divisant l'effectif d'une valeur (ou modalité) par le nombre total de valeurs au sein d'une même variable. Pour reprendre l'exemple du petit a) : pour calculer la fréquence de la modalité "yeux bruns" il faut donc diviser le nombre total d' "yeux bruns" par le nombre total de valeurs au sein de la variable;
Pour calculer une fréquence cumulée, il est tout d'abord nécessaire de calculer l'effectif cumulé, qui correspond à la somme des effectifs de toutes les valeurs inférieures ou égales à une valeur précédemment déterminée. La fréquence cumulée peut être obtenue en additionnant les fréquences inférieures ou égales à une valeur définie.
c) Une distribution statistique est le résultat obtenu une fois le processus de discrétisation achevé. Une distribution statistique correspond donc à un ordonnancement précis et réfléchi des classes statistiques et plus largement des données statistiques, elle révèle comment les valeurs d'un ensemble de données sont réparties.

❖ Séance 3

<u>Question de cours :</u>

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

Si on entend par généralisation, le processus par lequel des données, jugées représentatives, sont sélectionnées au détriment d'autres, jugées moins représentatives (valeurs aberrantes par exemple), afin d'aboutir à une vision globale d'un phénomène, alors un caractère quantitatif se prête plus à ce type d'opération. En effet, une classe statistique constituée de caractères quantitatifs peut se voir appliquer un grand nombre de paramètres statistiques susceptibles de simplifier la classe statistique à laquelle on s'intéresse et donc d'en tirer une tendance générale. Un caractère qualitatif ne peut pas se voir appliqué un paramètre statistique en

lui-même; par exemple, si l'on demande à un ensemble d'individus leur couleur préférée, on ne pourra pas calculer une moyenne des couleurs citées. La nature même du caractère qualitatif - une valeur non-mathématique, comme par exemple une couleur ou le numéro d'un département - fait qu'il ne peut pas se voir appliqué un grand nombre de paramètres statistiques permettant d'aboutir à une généralité.

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

Tout d'abord, les caractères quantitatifs correspondent à des valeurs de nature numérique qui se prêtent à des opérations mathématiques et donc à l'application de paramètres statistiques. Les caractères quantitatifs discrets correspondent à des valeurs exprimées en nombre absolu, tandis que les caractères quantitatifs continus correspondent à des valeurs exprimées en nombre relatif ou de rapport, qui sont le plus souvent des nombres décimaux ou de pourcentages.
Il est important de les distinguer car un caractère quantitatif ne se voit pas appliquer les mêmes formules de calculs selon qu'il soit discret ou continu. Au-delà de la formule mathématique, l'interprétation différera également selon la nature (discrète ou continue) du critère. Si la nature du critère n'est pas clairement établie avant le calcul des paramètres statistiques, le résultat obtenu risque d'aboutir à un non-sens. De plus, il est utile de les distinguer dans la mesure où valeur discrète et valeur continue offrent chacune des propriétés que l'autre n'offre pas. Le résultat peut s'avérer plus révélateur selon qu'on ait calculé le paramètre statistique à partir d'une valeur absolue ou à partir d'une valeur discrète.

3. **Paramètres de position :**

a) **Pourquoi existe-t-il plusieurs types de moyenne ?**

Il existe plusieurs types de moyenne en fonction de la nature de l'objet étudié. Certains objets nécessitent l'utilisation de formules plus spécifiques. Dans le cas du calcul d'une vitesse moyenne, par exemple, il faut calculer une moyenne harmonique, là où une moyenne quadratique ou géométrique s'appliqueront à des objets de nature polygonale (ou surfacique).

b) **Pourquoi calculer une médiane ?**

La médiane permet d'organiser une population ou une classe statistique en déterminant deux sous-populations de probabilité équiprobable. Elle peut être une étape préalable au calcul d'autres paramètres statistiques appliqués aux deux sous-populations établies, en vue de comparer ces deux sous-populations.
De plus, le calcul de la médiane permet, lorsqu'elle est comparée à la médiale, d'établir une mesure de concentration. Il s'agit donc d'un calcul préalable à celui d'autres paramètres statistiques pertinents.

c) **Quand est-il possible de calculer un mode ?**

Un mode peut être calculé dès lors que l'on connaît la fréquence d'apparition d'une valeur. En effet, un mode peut être considéré comme une moyenne de fréquence; il correspond à la valeur qui est la plus fréquente ou qui a la plus forte densité de probabilité. Dès lors, le calcul du mode s'avère pertinent si l'on émet l'hypothèse qu'une valeur apparaît plus que les autres.

4. **Paramètres de concentration**

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

La médiale permet de partager en deux parties égales la masse d'une variable; elle ne s'applique pas à l'effectif d'une population mais sur la somme totale des valeurs. Par exemple, dans le cas d'une masse salariale, la médiale est la valeur qui divise en deux parties égales la somme totale des salaires. L'intérêt est de déterminer une mesure de concentration; dans l'exemple donnée, il pourrait s'agir de constater si finalement il y a un équilibre ou au contraire un net déséquilibre entre individus. On pourrait avoir d'une part 5 individus représentant 50% de la masse salariale et d'autre part 20 individus représentant les 50 autres pourcents.
Dans cette perspective, l'indice de C.Gini permet de représenter graphiquement les effets de concentration au sein de la population statistique étudiée.

Paramètres de dispersion

a) **Pourquoi calculer une variance à la place de l'écart à la moyenne ? Pourquoi la remplacer par l'écart type ?**

Si l'on considère seulement la moyenne ou l'écart à la moyenne, on ne peut pas rendre compte de la régularité ou non de la série statistique étudiée. La variance - ou carré des écarts à la moyenne - permet, par l'usage des carrés des valeurs, de rendre compte de cette régularité ou autrement dit, du caractère homogène ou non des valeurs d'une série statistique. Elle répond finalement à la question suivante : y-a-t-il eu d'importantes variations de valeurs au sein de la série statistique considérée ? ; là où la moyenne ou l'écart à la moyenne peuvent induire en erreur laissant croire à une certaine homogénéité.
L'écart type peut s'avérer plus simple à utiliser que la variance car il est exprimé dans la même unité que la moyenne.

b) **Pourquoi calculer l'étendue ?**

Prise indépendamment, l'étendue permet de mettre en valeur les valeurs extrêmes d'une série statistique; elle a l'avantage d'être facile à calculer. Lorsqu'elle est appliquée à un deux quartiles, elle devient un paramètre statistique de dispersion

particulièrement pertinent en étant appliquée à différentes portions de valeurs de la série statistique.

c) À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ?

Un quantile permet de partager une série statistique en parties égales. Les quantiles les plus utilisées sont la médiane (quand la série statistique est partagée en deux parties égales) et les quartiles (quand la série statistique est partagée en quatre parties égales).

d) Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

Une boîte de dispersion ou boîte à moustache est une représentation graphique permettant d'avoir un aperçu synthétique des principaux paramètres statistiques calculés à partir d'un caractère quantitatif. Son caractère synthétique constitue son principal intérêt.
L'interprétation d'une boîte de dispersion s'effectue en comparant entre eux visuellement les résultats des différents paramètres statistiques pour un même caractère (comparaison entre la moyenne et la médiane par exemple). On peut également comparer deux boîtes de dispersion entre elles et donc, par ce biais, deux caractères entre eux.

5. Paramètres de forme

a) Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ?

Un moment absolu correspond à la position de la valeur de la moyenne, de la médiane ou du mode au sein d'une distribution statistique. Les moments centrés correspondent à la position de la valeur attendue des écarts d'une variable calculés par rapport à la moyenne, la médiane ou le mode.
Les moments permettent donc de connaître la dispersion d'une variable. Ils permettent de quantifier une dissymétrie identifiée au sein d'une distribution statistique et caractérisent cette dernière à partir de sa dissymétrie ou au contraire de sa symétrie.

b) Pourquoi vérifier la symétrie d'une distribution et comment faire ?

Vérifier la symétrie d'une distribution statistique permet d'avoir une idée de la dispersion globale de celle-ci. Il s'agit de répondre à la question : comment se répartissent globalement les valeurs d'un caractère étudié autour de la moyenne, de la médiane ou du mode ? Cette vérification permet aussi de comparer ces trois paramètres de position entre eux.

Pour vérifier la symétrie d'une distribution, on peut calculer les coefficients de Pearson et de Fisher, qui correspondent respectivement à la mesure de la dissymétrie et à la mesure de l'aplatissement de la distribution statistique. Plus les coefficients sont proches de 0, plus la distribution est symétrique.

❖ Séance 4 :

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 choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues, je prendrais en compte plusieurs critères. Tout d'abord, le choix va certainement dépendre du phénomène géographique étudié (s'agit-il d'un phénomène ou objet géographique dénombrable ? mesurable ? peut-on établir une hiérarchie ?) et des résultats qu'on souhaite en tirer (mettre en avant le caractère hétérogène du phénomène ou au contraire son caractère homogène par exemple.). Ensuite, le choix de l'une ou l'autre distribution peut dépendre du type de paramètres statistiques (moyenne, variance, écart-type etc.) que l'on souhaite appliquer au phénomène; en effet, les formules statistiques peuvent différer en fonction de la nature des variables ou se révéler plus ou moins faciles à appliquer.

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

Je pense que la loi normale doit être beaucoup utilisée en géographie lorsqu'il s'agit de réaliser des analyses statistiques multivariées. En effet, elle permet de mettre en relation plusieurs facteurs attribuables à une variable; elle permet aussi d'associer des outils statistiques comme la moyenne, la variance et l'écart-type. La représentation de cette loi sous forme de cloches (ogive de de Galton) semble, de plus, faciliter l'interprétation des résultats puisqu'elle permet de représenter les écarts à une moyenne ou "norme" représentée par l'axe des ordonnées. Dès lors, elle pourrait s'avérer très intéressante pour établir un point de comparaison entre une norme et des valeurs extrêmes qui selon leur nombre pourraient être des "valeurs aberrantes" ou bien révéler la survenue d'un événement inhabituel. En géographie physique, par exemple, elle pourrait permettre de révéler des anomalies climatiques (ex: fortes précipitations inhabituelles). En bref, la loi normale apparaît comme un très bon outil de synthèse.
Ensuite, je pense que la loi de Zipf est une loi qui est fréquemment utilisée en géographie en raison des lois rang-taille qu'elle établit. En effet, elle permet d'établir

une sorte de rapport proportionnel entre le nombre d'habitants d'un territoire donné (comme une ville, un département, une région…) et le rang de ce territoire. De plus, si on se base sur l'application initiale qu'en a fait son créateur, c'est-à-dire à la fréquence d'apparition d'un mot dans un texte, la loi pourrait se révéler intéressante pour les études en géographie humaine relevant de la perception d'un phénomène par une population; ce type de loi pourrait accompagnée d'autres méthodes s'appuyant aujourd'hui sur l'intelligence artificielle comme la réalisation de nuages de mots mettant en avant les mots les plus répétés au cours d'entretiens.

- Séance 5 :

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 une méthode qui consiste à prélever au sein d'une population mère une partie seulement des individus statistiques qui vont dès lors former un sous-ensemble représentatif de l'ensemble de la population; c'est la démarche de l'inférence. Plusieurs échantillons peuvent être réalisés au sein d'une même population. On n'utilise pas la population entièrement car un trop grand nombre d'individus statistiques rend l'étude statistique difficile voire impossible en plus de se révéler très onéreuse. L'étude d'un échantillon représentatif constitue donc un gain de temps et d'argent.
Il existe deux types de méthodes d'échantillonnage principales nécessitant de disposer d'une base de sondage : les méthodes aléatoires (faisant appel à un tirage au sort) et les méthodes non aléatoires. Au sein des méthodes aléatoires on retrouve le tirage avec remise et celui sans remise. Pour les méthodes non-aléatoires on retrouve la méthode de l'échantillonnage systématique et celle des quotas. Enfin, il existe aussi les méthodes d'échantillonnage dites de "Monte Carlo" pouvant être appliquées aux petites populations.
La méthode d'échantillonnage dépend de l'objectif poursuivi par le chercheur : s'il se base sur des méthodes d'échantillonnage aléatoires, il obtiendra des résultats globaux sur la population mère, indépendamment de l'existence d'éventuels sous-ensembles au sein de la population mère. S'il choisit d'utiliser des méthodes d'échantillonnages non-aléatoire, alors il peut orienter sa recherche sur un sous-ensemble en particulier de la population mère. Par exemple, le chercheur peut décider d'orienter sa recherche sur la consommation d'alcool chez les 18-25 ans et non pas sur l'ensemble des tranches d'âge présentent au sein de la population mère.

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

Dans le cadre de l'étude statistique d'un échantillon, l'estimateur correspond à une variable aléatoire attribuée à l'échantillon avant même d'observer les données, c'est une fonction des données; l'estimateur est une méthode, un modèle théorique qui permet d'attribuer à l'échantillon un paramètre encore inconnu. L'estimation est quant à elle la valeur numérique obtenue en appliquant l'estimateur. L'estimateur est l'outil et l'estimation est la donnée numérique, le résultat.

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

Les deux intervalles permettent de déterminer la fiabilité de l'étude statistique menée. La différence entre les deux est la suivante :

- Dans le cas d'un intervalle de fluctuation, le paramètre à tester est connu, il s'agit de savoir si la fréquence observée appartient ou non à l'intervalle de confiance. On part du paramètre et on "teste" l'échantillon à partir de lui.
- Dans le cas de l'intervalle de confiance, on part à l'inverse de l'échantillon car le paramètre est encore inconnu. Il permet de répondre à la question : quelle est la valeur la plus probable que peut prendre la moyenne de la population mère dans un échantillon de taille n ?

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

Dans la théorie de l'estimation, le biais permet de mesurer l'erreur, l'aspect "biaisé" d'un estimateur, c'est-à-dire l'écart entre le résultat que donne l'estimateur et la valeur réelle du paramètre estimé. On l'appelle également "erreur 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 est appelée statistique exhaustive. Elle a un point commun avec les *big data* qui est celui de vouloir traiter ou stocker un volume massif de données.

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

Le choix de l'estimateur est important car l'objectif est de choisir l'estimateur qui pourra donner une estimation ponctuelle la plus proche possible du paramètre connu ou que l'on cherche à connaître. Les estimateurs n'ont pas tous des biais de même

nature. Par exemple, la moyenne est un estimateur sans biais et convergent ce qui signifie qu'il est très peu probable qu'elle s'écarte de sa cible. Le premier enjeu est donc de choisir un estimateur ayant un biais le plus faible possible en fonction du résultat recherché. Ensuite, il faut choisir un estimateur avec lequel on perd le moins d'information possible afin que l'échantillon conserve son caractère représentatif. Finalement, c'est donc le caractère plus ou moins précis et donc fiable de l'estimateur qui est en jeu et par extension le caractère représentatif de l'échantillon.

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

Il existe plusieurs méthodes d'estimation d'un paramètre. On peut estimer un paramètre à partir d'un intervalle (intervalle de confiance, intervalle de pari, méthode du *bootstrap*) ou encore à l'aide de distributions se basant sur la loi normale ou celle de Poisson. D'autres méthodes existent comme la méthode des moindres carrés ou celle du maximum de vraisemblance. Le choix de la méthode dépend de la nature du paramètre à estimer mais aussi de la nature de la statistique (exhaustive ou s'appuyant sur l'échantillonnage). Certaines méthodes peuvent également être plus ou moins efficaces en fonction du nombre de variables aléatoires à étudier. Par exemple, la méthode des moindres carrés se révèle plus pertinente à utiliser lorsque plusieurs variables aléatoires doivent être étudiées.

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

Il existe deux types principaux de tests statistiques aussi appelés tests d'hypothèse : les tests paramétriques qui s'appuient sur le moyenne, l'écart type, le type de distribution, et les tests non-paramétriques s'appuyant sur l'effectif, la médiane, entre autres. Un test statistique vise à établir un jugement sur l'échantillon, c'est-à-dire qu'il permet de vérifier si une hypothèse statistique formulée à propos d'une population est vraie ou fausse; il peut s'agir de savoir si un événement quelconque a eu un impact sur les observations considérées par exemple. Parmi les différents types de test statistiques il existe également : les tests robustes, les tests unilatéral et bilatéral, le test de signification, les tests d'ajustement, le tests de comparaison, les tests d'indépendance.
Pour créer un test, plusieurs étapes sont nécessaires. Il faut tout d'abord poser une hypothèse de travail et formuler clairement une hypothèse nulle et une hypothèse alternative. Il faut également prendre en compte la nature du paramètre considéré pour déduire une loi de distribution. Il est ensuite nécessaire de choisir un seuil d'erreur de probabilité, préciser les conditions d'application du test et le tester en s'appuyant sur la collecte des données d'un ou plusieurs échantillons. Il faut également prendre en considération les caractéristiques propres à la population.

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

La statistique inférentielle fait l'objet de plusieurs critiques portant notamment sur sa dépendance à des échantillons; en effet, si ces derniers sont considérés comme biaisés et donc pas suffisamment représentatifs de la population totale, la statistique inférentielle se révèle alors être une mauvaise méthode. Finalement ce qui est majoritairement reproché à la statistique inférentielle c'est le fait de ne pas prendre en considération la totalité des données à la manière de la statistique exhaustive. En ce qui concerne cette première critique, je dirais que je suis d'accord avec elle aux premiers abords mais que dans les faits, il s'avère très difficile d'un point de vue financier et organisationnel de réaliser des études exhaustives. Dès lors, je pense qu'il faudrait insister sur le fait qu'il est nécessaire de s'appuyer sur plusieurs échantillons et non pas un seul pour augmenter les chances de "représentativité". Ensuite, en ce qui concerne la question des tests, dont la mise en place est parfois critiquée, je pense qu'il est encore une fois nécessaire d'utiliser plusieurs tests et non pas seulement un. Finalement le plus grand risque en statistique inférentielle c'est "l'absolutisation" d'un échantillon, d'un paramètre, d'un test, d'une interprétation. Je pense que tant que la statistique inférentielle est pratiquée avec un recul critique et dans une logique de pluralité des méthodes utilisées, les risques de non-représentativité peuvent être limités. C'est donc finalement la rigueur du scientifique à propos de l'application des méthodes et de leur choix qui est le facteur le plus important.

- Séance 6 :

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 ?**

La statistique ordinale aussi appelée statistique d'ordre est un type de statistique très utilisée en géographie humaine qui utilise les variables qualitatives. Elle permet d'établir un classement, une hiérarchie entre les données. Elle s'oppose à la statistique nominale au sein de laquelle aucune hiérarchie ne peut être effectuée. La statistique ordinale peut matérialiser une hiérarchie spatiale à condition qu'elle s'appuie sur un ou plusieurs critères attribuables aux territoires étudiées. Par exemple, en géographie urbaine, les villes sont souvent des indicateurs de fécondité et d'activité; dès lors, il est possible d'effectuer un classement des villes à partir d'un échantillon, en fonction du taux de fécondité par exemple.

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

L'ordre qui doit être privilégié est l'ordre croissant aussi appelé ordre naturel. En effet, il permet de déterminer quelles sont les valeurs aberrantes d'une série d'observations, ainsi que d'étudier la loi de la plus grande valeur de la série.

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

Une corrélation des rangs consiste à comparer deux classements entre eux à partir du calcul de coefficients de corrélation sur leurs rangs, on les compare à partir de leurs valeurs. La concordance de classements constitue en quelque sorte l'étape suivant la corrélation des rangs. En effet, elle consiste à généraliser le coefficient de corrélation des rangs établi par Kendall; elle a un aspect plus "globalisant" et permet de déterminer le degré de concordance, d'accord, entre les différents classements. On passe de l'échelle des rangs à celle plus globale des classements en quelque sorte.

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

Ces deux tests permettent de comparer deux classements et de déterminer s'ils sont identiques ou non. Cependant, ils ne procèdent pas de la même façon. En effet, le test de Spearman s'appuie la différence entre les rangs d'un classement à partir d'un coefficient de corrélation linéaire. Les rangs sont considérés comme ayant des permutations équiprobables. À l'inverse, le test de Kendall s'appuie sur les concepts de concordance et de discordance en formant des couples de rangs et en les comparant à partir de leur caractère discordant ou concordant.

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

Le coefficient de Goodman-Kursdal permet d'établir une proportion entre paires concordantes et paires discordantes, ainsi que définies par les créateurs du coefficient. Le coefficient de Yule permet quant à lui d'étudier des fréquences à partir de tables de contingences et évalue donc la fréquence d'un événement "statistique" au sein des classements.

<u>Réflexion sur les humanités numériques :</u>

Avant d'évoquer les apports de l'informatique aux humanités, je pense qu'il est intéressant de distinguer au sein des humanités les sciences humaines, des lettres et de l'art ou encore l'économie de la sociologie, tout simplement parce qu'il n'est pas question des mêmes besoins. Néanmoins, le point commun unissant ces disciplines du point de vue du numérique - et pas seulement pour elles, mais aussi pour les sciences dures - est le fait que l'informatique a permis de rendre la connaissance plus accessible, il constitue en cela un formidable instrument de

communication et de diffusion du savoir, à condition de ne peut pas se perdre dans la masse informationnelle.
Au sein des sciences humaines que sont l'histoire, l'archéologie et surtout la géographie, on ne peut nier l'apport de l'informatique au niveau des différents chaînons de la production du savoir, notamment au travers du développement des SIG ou d'instruments comme le LIDAR, lequel a permis de mettre au jour des vestiges archéologiques. Néanmoins, je peine à voir comment l'outil informatique - mis à part pour rendre la diffusion du savoir plus aisée - pourrait être exploité, du point de vue du traitement de données, dans une matière comme la philosophie qui est pourtant bel et bien une science humaine. Si je caricature, je ne vois pas en quoi dénombrer le nombre de mots d'un texte philosophique va faire avancer la réflexion dans cette discipline. L'outil informatique pourrait certes permettre d'établir le nombre de fois où un mot revient dans un texte, éventuellement établir une classification par thèmes, mais son utilité peut se révéler limitée voire inexistante selon la discipline des humanités considérée. En quoi l'outil informatique dans le cas de la philosophie, pour reprendre mon exemple, ou encore de la littérature ou des arts pourrait-il mener à "davantage de rigueur" ? En fait, l'outil informatique se révèle réellement utile dès lors que l'on a véritablement affaire à des données supposant un traitement comme par exemple le nombre de morts qu'a engendré un évènement historique, et encore, un nombre de morts, une statistique, ne suffisent pas à saisir le contexte dans lequel a eu lieu un évènement ou ses retombées. Les mots et les phrases d'un poème, les couleurs et les lignes d'un tableau, les raisonnements logiques et rigoureux d'un philosophe, ne peuvent être réduites à des données traitables et classifiables. On ne parviendra pas à mieux interpréter un poème ou un texte philosophique simplement en dégageant des thèmes ou des mots "clefs". De ce point de vue là, je ne pense donc pas que "l'informatique permet un meilleur apprentissage (de l'ensemble) des humanités". Par contre, du point de vue du langage informatique, il est vrai que dès lors qu'un texte issu de l'une de ces disciplines est posté sur internet, par exemple, il devient de la donnée nécessitant un stockage (*data center* etc.); néanmoins, ce fait ne réduit pas la pensée, la réflexion sur un sujet, à de la "pensée" computationnelle.
Je pense donc que le numérique est avant tout un outil au service des humanités au sens large dont l'utilisation et la pertinence diffèrent grandement d'une discipline à l'autre. L'outil informatique constitue un allié certain pour les humanités à condition qu'il soit utilisé correctement et qu'il conserve sa place d'outil et/ou d'aide à la diffusion des savoirs. Je pense qu'il serait dommageable pour la production de toutes connaissances et de pensée que les humanités, mais également les sciences dures, soient réduites à des "jeux de données", et se trouvent happées par un langage informatique qui leur est, en réalité, hétérogène. Le principal risque lorsque l'on songe au lien entre humanités et numérique est celui de vouloir réduire le langage et le fonctionnement de l'un au langage et au fonctionnement de l'autre. Deux écueils sont à éviter : tomber dans une forme de dogmatisme technocrate naïf ou rejeter en bloc l'apport de l'outil informatique, qui pertinemment utilisé peut faire

avancer la connaissance, pour certaines disciplines des humanités, et faciliter la diffusion des savoirs, pour toutes.

Retour d'expérience personnel sur le cours d'analyse de données :

Comme vous l'aurez constaté, je n'ai rendu qu'une partie du travail (principalement les questions et une partie du code). Venant d'une filière ayant une approche très littéraire des sciences humaines, le cours m'a paru assez complexe, j'ai donc mis un certain temps pour comprendre vos cours de statistiques et pour répondre aux questions, que j'ai privilégiées car elles m'apparaissaient comme étant la partie la plus accessible du cours, en plus d'avoir un poids important dans la notation du parcours débutant.
Je dois reconnaître que même en ayant lu vos cours sur python et en ayant regardé quelques tutos, la pratique du code n'était pas aisée, bien que j'ai réussi quelques manipulations et que j'ai réussi à utiliser GitHub, outil que je trouve très intéressant.
Le degré de précision des cours de statistiques - surtout quand on "reprend" les maths - m'a semblé parfois un peu déconnecté des besoins réels que nous avons dans nos parcours respectifs, ce qui n'enlève pas son caractère intéressant et instructif en lui-même. Dans mon cas, il a été question dans un cours de mon parcours GAED de réaliser une ACP et un test du Khi2 et j'ai trouvé dommage que la réalisation d'un test de Khi2 avec python ne soit "réservée" qu'à une séance du parcours intermédiaire ou confirmé (j'ai un doute); je me suis demandée pourquoi il arrivait "aussi tard" sachant qu'il ne nécessite que très peu de lignes de codes, comme vous l'avez expliqué lors d'un cours. Ce n'est bien entendu qu'un détail, mais j'insiste sur le test du Khi2 car il se trouve que l'ensemble des parcours GAED en ont eu également besoin pour l'autre cours de statistique (celui de Mme.Huguenin). Le principe de la classe inversée ne m'a pas spécialement rebuté, mais peut-être serait-il intéressant de consacrer un temps d'une séance en présentiel à des éléments clefs qui nous sont communs à tous comme par exemple ce fameux test du Khi2. Dans cette perspective, j'ai trouvé ça un peu dommage qu'il n'y ait pas plus de "passerelles" entre votre cours et celui de Mme.Huguenin, bien qu'il ne soit pas question des mêmes outils informatiques.
En conclusion, ce cours m'a paru très complexe, parfois un peu démoralisant, mais m'a permis de démystifier un peu les statistiques et le codage.