Elise Grimaldi
21224745

Analyse de données
Rapport d'activités

Séance 2

Questions de cours

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

La géographie entretient une relation complexe avec les statistiques :

Elle produit des données massives que seul l'outil statistique permet d'étudier .
Historiquement, elle a souvent sous-estimé leur utilité, mais aucun géographe ne peut aujourd'hui s'en passer.
Les statistiques permettent la réduction de l'incertitude et l'analyse structurée des phénomènes géographiques.

Les statistiques sont donc un outil indispensable pour faire de la géographie une véritable science basée sur l'analyse rigoureuse.

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

Deux visions coexistent :

Le déterminisme (Laplace) : le hasard n'existe pas, tout a une cause.
Le hasard comme cause cachée, explorable grâce à la progression des connaissances.

La géographie utilise ces deux approches :
elle cherche des tendances globales, même si les détails individuels ne sont pas prévisibles. Le hasard existe sous forme de variabilité, que les statistiques permettent de modéliser (loi normale, loi de Pareto, etc.) .

En géographie, le hasard existe localement, mais des tendances globales sont identifiables.

3. Quels sont les types d'information géographique .

Il y a deux grands types d'information géographique : d'une partles données attributaires (caractéristiques des territoires) : la population, les variables sociales, économiques, climatiques… et d'autre part les données géométriques : les formes, contours, surfaces, réseaux et morphologies spatiales. Ces deux éléments sont aussi constitutifs de la base d'un SIG

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

La géographie nécessite la production et la collecte de données (nomenclature, métadonnées) mais aussi l'étude de la structure interne des données (matrice individus, variables) et l'usage d'outils statistiques pour résumer, modéliser, comparer et visualiser les phénomènes complexes.
Le but est d'extraire des connaissances, de détecter des structures et de confronter théorie et réalité.

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

La statistique descriptive, décrit et résume les données en utilisant les moyennes, quantiles, histogrammes, ACP, AFC, ACM… Son objectif est d'ordonner des données et de préparer des comparaisons et des prévisions .

La statistique explicative cherche quant à elle à comprendre une variable Y à partir de variables explicatives X1… Elle inclut : régression, analyse discriminante, régression logistique, ANOVA… et son objectif et d'expliquer ou prédire

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

Les diagrammes sectoriels sont le type de visualisation adapté pour les variables qualitatives. Pour les variables quantitatives on peut utiliser des historiographes, despotes à moustaches, des courbes cumulatives ou encore des polygones de fréquences etc. Enfin pour des données multivariées : des cartes de proximités, des cartes thématiques (dans contexte SIG, évoqué indirectement), ACP, AFC, ACM

Le choix ce fait donc selon le type de variable, la continuité ou non, la nécessité de comparer plusieurs variables ainsi que la volonté de réduire la dimension

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

On retrouve trois grandes familles :

Les méthodes descriptives :

- ACP, AFC, ACM
- Classification (CAH, nuées dynamiques)
- Analyses de proximité .

Les méthodes explicatives :

- Régression simple/multiple
- Régression logistique
- Analyse discriminante
- Analyse de la variance
- Modèles linéaires généraux .

les méthodes de prévision :

- Analyse des séries chronologiques
- Modèles : $X_t = f(X_{t-1}, X_{t-2}, \ldots) + \text{aléa}$

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) Population statistique

Ensemble des unités étudiées
Ex : toutes les villes d'une région

b) Individu statistique

Un élément de la population (unité spatiale)
Ex : une commune, un ménage

c) Caractères statistiques
Caractéristique mesurée sur chaque individu (âge, superficie, revenu…) .

d) Modalités statistiques

Valeurs possibles du caractère, exclusives et exhaustives (ex : homme/femme, 0-77ans etc).

Types de caractères

Qualitatifs nominal / ordinal
Quantitatifs discret / continu

Existe-t-il une hiérarchie entre eux?
Oui, les variables quantitatives permettent plus d'opérations (tests paramétriques, distribution), les qualitatives moins

9. Comment mesurer une amplitude et une densité?

Amplitude d'une classe : (A = b - a)

Densité d'une classe : (d = \dfrac (n_i)(b - a))
—> effectif de la classe divisé par son amplitude .

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

Elles servent à déterminer le nombre optimal de classes lors de la discrétisation d'un caractère quantitatif.

Sturges :
(k \approx 1 + 3{,}2222 \times \log_{10}(n))
Yule :
(k \approx 2{,}5 \sqrt[4]{n})

—> Elles évitent un découpage trop fin ou trop grossier.

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

Effectif : nombre d'apparitions d'une modalité dans la population .

Fréquence (fi) :
Calcul : (f_i = \dfrac{n_i}{n})

Proportion de la modalité

Fréquence cumulée

Somme des fréquences des modalités < une valeur donnée :
(F_k = \sum_(i=1)^(k) f_i) .

La distribution statistique correspond à l'ensemble des fréquences observées pour les différentes modalités, elle sert à identifier la loi de probabilité sous-jacente

Analyse de l'exercice de code

Pour cette première approche pratique, nous avons travaillé sur les fondamentaux de la structure des données. La géographie, bien qu'étant une science de terrain, repose aujourd'hui massivement sur des données structurées que nous devons savoir manipuler. L'exercice pratique a consisté en la prise en main de la bibliothèque pandas. Le script main_session2.py a chargé le fichier resultats-elections-presidentielles-2022-1er-tour.csv. L'exécution du code a révélé la structure "brute" de l'information :

- La visualisation des premières lignes via le terminal offre une première "image" textuelle de la donnée. On y voit une matrice où chaque ligne est une commune (l'individu statistique) et chaque colonne une variable (le caractère) nous sommes face à un tableau de plusieurs dizaines de milliers de lignes, correspondant chacune à une commune française. Cette étape, bien que technique, illustre parfaitement la distinction théorique vue en cours entre les individus statistiques (ici les communes) et les caractères (les inscrits, les votes).

- Le code a permis de vérifier comment Python "voit" les données. Les colonnes "Libellé de la commune" sont traitées comme des objets (chaînes de caractères, qualitatives), tandis que les colonnes "Voix" sont des entiers (int64, quantitatives). Cette distinction informatique valide la distinction théorique vue en cours.

- Les graphiques (camemberts et en barre) ont permis de faire une géographie électorale de la France par département. Les diagrammes en barres ont permis de montrer la part d'inscrits contre la part des votants et les « camemberts » (diagramme circulaire) ont mis en lumière la part d'abstentionnistes.

Difficultés rencontrées **:** La principale difficulté a été la gestion de l'encodage des fichiers. Lors des premiers essais, les noms de communes contenant des accents (é, è, ç) apparaissaient sous forme de caractères illisibles. Au début je ne comprenais pas puis je me suis rendue compte qu'il fallait spécifier l'encodage dans la fonction read_csv pour résoudre ce problème et obtenir une lecture correcte.

Séance 3

Questions de cours

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

Le caractère quantitatif est le plus général puisque les paramètres statistiques concernent principalement les variables quantitatives et seulement ponctuellement les variables qualitatives.
En effet, le caractère quantitatif permet de calculer tous les paramètres (moyenne, variance, moments…), il se prête aussi à l'ensemble des opérations statistiques vues (dispersion, forme, position). Les caractères qualitatifs ne permettent quant à eux que des descripteurs limités.

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

Les caractères quantitatifs discrets correspondent aux valeurs dénombrables, séparées (comme le nombre d'enfants), leur moyenne se calcule par une somme

Les caractères quantitatif continus correspondent eux aux valeurs sur un intervalle (longueur, revenu)
Les paramètres deviennent des intégrales

On les distingue d'abord car les les formules ne sont oas les mêmes (somme vs intégrale), aussi parce que les représentations changent (histogrammes continus, classes) et enfin parce que certaines mesures comme les médianes ou les quantiles se calculent différemment pour les deux types.

3. Paramètres de position

- Pourquoi existe-t-il plusieurs types de moyenne?

Le tableau du cours montre plusieurs moyennes (arithmétique, géométrique, quadratique, harmonique, mobile…) Il en existe plusieurs types afin qu'elles répondent aux situations différentes : la nature de la variable (continue ou discrète), les propriétés voulues et les contextes d'usage (vitesse –> harmonique, produits –> géométrique)

- Pourquoi calculer une médiane ?

On calcule une médiane car contrairement à la moyenne elle n'est pas influencée par les valeurs extrêmes, elle convient aussi à des séries très dissymétriques, engin elle résumer la position centrale même quand la moyenne est trompeuse. On la calcule donc pour obtenir une mesure robuste et insensible aux valeurs aberrantes.

- Quand est-il possible de calculer un mode?

On calcule un mode uniquement lorsque la distribution présente une valeur dominante identifiable. En effet, le mode existe lorsqu'une modalité a l'effectif maximal ou la plus grande densité. Il peut manquer ou être « non unique » (cas des séries pluri-modales) et il dépend du regroupement en classes pour les variables continues

4. Paramètres de concentration

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

La médiale partage la masse totale en deux parties égales (50 % – 50 %) , elle est toujours plus grande que la médiane et elle permet d'évaluer l'inégalité de distribution d'un caractère.

L'indice de Gini mesure la concentration d'un caractère dans la population, il montre si une petite proportion d'individus concentre une grande part de la masse totale
Il sert donc à mesurer l'inégalité (revenus, tailles, surfaces…).

5. Paramètres de dispersion

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

La variance utilise les carrés ce qui donne des propriétés mathématiques utiles que n'a pas la valeur absolue. L'écart type correspond simplement la racine de la variance : il revient à l'unité de l'origine et est donc plus interprétable. Ainsi la variance comprend plus de rigueur mathématique tandis que l'écart type correspond à une interprétation pratique.

- Pourquoi calculer l'étendue?

On calcule l'étendue car elle est simple à obtenir (maximum – minimum) et parce qu'elle donne une première idée de la dispersion. Toutefois sa fiabilité reste faible surtout pour les grands effectifs puisqu'elle ne repose que sur les extrême

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

Les quantiles divisent une série en parties égales, ils permettent d'étudier la répartition interne des valeurs et de construire des indicateurs robustes.

- Pourquoi construire une boîte de dispersion ? Comment l'interpréter?
La boîte de dispersion permet de visualiser rapidement à la fois la médiane, les quartiles, l'étendue,les valeurs extrêmes et aussi de comparer facilement plusieurs distributions

On peut l'interpréter ainsi :
Le rectangle : 50 % des valeurs
Ligne interne : médiane ;
Moustaches : valeurs minimum et maximum

—> Elle résume donc à la fois la position, la dispersion et l'asymétrie

6. Paramètres de forme
- Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser?

Les moments centrés correspondent aux moments calculés par rapport à la moyenne, ils servent à mesurer la variance, l'asymétrie et l'aplatissement
Les moments absolus utilisent la valeur absolue et moins influencé par les valeurs très grandes ou très petites

- Pourquoi les utiliser ?
Pour caractériser la forme de la distribution : symétrie, aplatissement, dissymétrie.

- Pourquoi vérifier la symétrie d'une distribution et comment faire ?
On vérifie la symétrie d'une distribution puisque si une distribution est symétrique elle simplifie l'analyse, la moyenne, la médiane et le mode coïncident dans ce cas et aussi parce que les choix les choix statistiques (tests, modèles) dépendent de la symétrie

On peut utiliser le coefficient d'asymétrie beta 1 (cf formule)

Beta 1 > 0 –> queue à droite

Beta 1 < 0 –> queue à gauche

Beta 1 = 0 –> symétrie .

Ou sinon la comparaison des paramètres :

mode ≈ médiane ≈ moyenne –> distribution symétrique.

Analyse de l'exercice de code

Cette séance visait à résumer l'information contenue dans ces immenses tableaux de données. Il s'agissait de dépasser la simple liste de valeurs pour extraire des indicateurs de position et de dispersion. J'ai appris à distinguer les paramètres de position (moyenne, médiane, mode) des paramètres de dispersion (variance, écart-type, étendue). Un point clé a été la distinction entre la moyenne, très sensible aux valeurs extrêmes, et la médiane, plus robuste, qui partage la population en deux. Nous avons aussi vu l'importance de la forme de la distribution (symétrie, aplatissement) pour choisir les bons indicateurs.

Pour analyser les résultats électoraux, mon code a généré une série de statistiques pour chaque candidat. Ce qui frappe à la lecture des résultats, c'est l'écart important entre la moyenne et la médiane pour la plupart des candidats. Par exemple, pour les "petits" candidats, la médiane est souvent très faible, voire nulle dans de nombreuses petites communes, alors que leur moyenne est relevée par quelques scores plus importants dans les grandes villes. Cela traduit une distribution très asymétrique des votes. Le script « main_session3.py » a généré des statistiques descriptives et surtout des boîtes à moustaches (boxplots) pour les résultats électoraux. C'est ici que l'analyse visuelle prend tout son sens et a été particulièrement éclairante. Ces graphiques permettent de voir d'un coup d'œil la dispersion des votes. On remarque immédiatement une structure particulière. La « boîte » centrale (qui contient 50% des communes, l'écart interquartile) est surmontée d'une très longue moustache et d'une multitude de points isolés vers le haut. Ces points sont les outliers (valeurs aberrantes). En géographie électorale, ils matérialisent les "bastions". Cela signifie que la distribution des votes n'est pas homogène : le candidat réalise des scores moyens partout, mais explose ses records dans certaines communes spécifiques. L'asymétrie est flagrante vers le haut.

C'est la preuve visuelle d'une polarisation spatiale du vote.
En parallèle, j'ai travaillé sur le fichier des îles (island-index.csv). J'ai procédé à une discrétisation des surfaces en créant des classes (de 0 à 10 km², de 10 à 25 km², etc.). Le tableau de répartition obtenu montre une écrasante majorité de très petites îles et une poignée d'îles gigantesques : en effet la classe des petites surfaces (0-10 km²) écrase toutes les autres en effectif, alors que les classes supérieures sont quasi vides. Cela illustre parfaitement la notion de concentration et d'inégalité (que l'on pourrait mesurer avec l'indice de Gini). En géographie, les « petits » objets sont la règle, les géants sont l'exception.

Problèmes rencontrés : J'ai été confrontée à des erreurs lors du calcul des moyennes à cause de cellules vides (NaN) dans les données. J'ai dû modifier mon approche pour nettoyer les colonnes (dropna) avant de lancer les calculs, réalisant que la statistique descriptive ne tolère pas les « trous » dans la donnée.

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?

Le choix d'une loi statistique, donc d'une distribution, dépend en premier lieu de la nature du phénomène étudié, ce qui permet de choisir entre « loi discrète et loi continue ». Viennent ensuite la forme de la distribution empirique (visuellement ou statistiquement testable), les caractéristiques de la série : espérance, médiane, variance, asymétrie etc. et le nombre de paramètres de la loi, certaines lois s'adaptant davantage selon leur flexibilité .
On choisit donc une loi/distribution discrète lorsque :

- les valeurs possibles sont dénombrables, souvent limitées à des entiers
- il s'agit de comptages: nombre d'événements, de succès/échecs, d'individus

(cf lois discrètes : Binomiale, Bernoulli, Poisson, Hypergéométrique…) .

Parallèlement, on choisit une loi/distribution continue lorsque :

- la variable peut prendre toutes les valeurs d'un intervalle, non dénombrables
- il s'agit de mesures continues : temps, distance, altitude, température… (cf les lois continues : normale, log-normale, exponentielle, uniforme continue…) .

Le critère majeur est donc la nature du phénomène et la structure des valeurs observée, le tout appuyé par la forme empirique de la distribution et les paramètres statistiques.

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

Certaines lois sont particulièrement importantes la géographie, en premier lieu la loi de Poisson qui est décrite comme « indispensable pour les événements rares » et apparaît lorsque l'on compte des occurrences dans une surface ou un intervalle. Son utilité en géographie est de modéliser des événements ponctuels dans l'espace ou le temps comme par exemple le nombre d'accidents, de séismes, d'occurrences d'un phénomène localisé.

En outre, la loi normale (Gauss) est elle décrite comme « la plus fréquente » et constitue souvent la distribution limite de nombreux phénomènes Elle permet de nombreuses variables naturelles et sociales approximées par une normale par exemple la distribution des hauteurs, les températures, les revenus ou les rendements.

La loi log-normale est mentionnée comme essentielle pour des variables multiplicatives et asymétriques (loi de Galton-Gibrat), elle est utile pour tout ce qui est taille des villes, surface des parcelles, revenus, intensité de certains flux.

Les lois rang-taille (Zipf et Zipf–Mandelbrot) sont utilisée en géographie pour les distributions rang-taille notamment pour les tailles de villes . Elle permet de modéliser la hiérarchie urbaine et analyser la structure polarisée d'un territoire.

Enfin, La loi exponentielle est utilisée pour les processus liés au temps d'attente ou aux risques, décrite comme adaptée aux phénomènes de fiabilité et de survie .Donc utile pour les durées d'événements naturels ou techniques ou la modélisation de probabilité de défaillance, de temps entre deux occurrences.

Analyse de l'exercice de code

L'objectif était ici de modéliser le hasard et de comprendre les formes théoriques que peuvent prendre nos données. Plutôt que de traiter des données réelles, python permet de simuler différentes lois de probabilité et visualiser leurs fonctions de densité. Le choix d'une loi de probabilité ne se fait pas au hasard : il dépend de la nature du phénomène (discret ou continu).
Le script main_session4.py a servi de laboratoire pour visualiser ces concepts abstraits.
Le graphique en bâton *poisson_discrete_pmf.png* montre bien la nature discrète de la loi de Poisson : on ne peut avoir que 1, 2 ou 3 événements, pas 2,5. On note l'asymétrie caractéristique vers la gauche pour les petites valeurs de lambda.
La Loi Normale vs La Loi de Pareto :

Cf *normale_pdf.png* la courbe en cloche produite par le code est d'une symétrie parfaite. Elle incarne l'équilibre autour de la moyenne. Visuellement, elle nous dit que les valeurs extrêmes sont très improbables.

À l'inverse *pareto_pdf.png* , le graphique de la loi de Pareto que j'ai généré montre une courbe qui part de très haut et décroît lentement (courbe en J). On peut l'interpréter comme la signature visuelle de l'inégalité. Contrairement à la courbe normale, la « queue » de distribution est épaisse : les événements extrêmes (villes géantes, revenus immenses) ne sont pas si improbables que ça. C'est fondamental pour comprendre les risques ou la hiérarchie urbaine.

Cette confrontation visuelle m'a permis de mieux saisir pourquoi certaines lois sont plus adaptées à la géographie humaine (inégalités, hiérarchies) et d'autres à la biométrie ou aux erreurs de mesure (loi Normale).

Séance 5

Questions de cours

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

L'échantillonnage peut se définir comme la procédure consistant à sélectionner un sous-ensemble d'individus appelé « échantillon », permettant d'estimer des caractéristiques de la population complète. Le sondage est lui-même défini comme « un ensemble de méthodes de collecte, exploitation et analyse d'informations données par un échantillon ».

La population entière n'est pas utilisée puisque cela pose des problèmes liés au recensement exhaustif (parfois impossible et très coûteux),il existe aussi des difficultés de couverture (c'est à dure la possibilité que la base de sondage oublie des individus) et plus généralement au temps et aux moyens nécessaires pour collecter des données massives.

Deux familles se distinguent pour les méthodes d'échantillonnage :

les méthodes probabilistes (aléatoires) :

- aléatoire simple
- aléatoire systématique
- stratifié
- en grappes

Ces méthodes reposent sur la probabilité et permettent des inférences valides.

les méthodes non probabilistes :

- quota
- boule de neige

- itinéraire

Ces méthodes ne reposent pas sur un tirage aléatoire et exposent à des biais.

On trouve entre autre le sondage aléatoire simple, systématique, stratifié, par grappe, etc, le sondage par quotas, boule de neige, itinéraire

Comment choisir ?
On choisit selon la nature de la base de sondage(qualité de couverture), l'objectif de l'étude (description, estimation, inférence), les ressources disponibles, l'exigence de précision (les méthodes probabilistes sont préférées dès que l'objectif est inférentiel).

2. Comment définir un estimateur et une estimation ?

Un estimateur est une variable calculée à partir de l'échantillon, destinée à approcher un paramètre inconnu de la population
Une estimation est une valeur numérique obtenue en appliquant l'estimateur aux données observées.

3. Intervalle de fluctuation vs intervalle de confiance

Intervalle de fluctuation :
Dans le cadre d'un sondage, c'est un intervalle qui décrit la fluctuation normale des résultats possibles d'un échantillon lorsqu'on connaît la proportion réelle.

Intervalle de confiance :
Ici c'est l'inverse c'est un intervalle estimé à partir de l'échantillon pour encadrer un paramètre inconnu de la population.

Le cours compare explicitement : « Les intervalles de fluctuation peuvent être considérés comme une méthode d'estimation… équivalente à l'intervalle de confiance autour d'une fréquence. »

Ainsi, avec l'intervalle de fluctuation on connaît la population –> on prédit un échantillon et avec l'intervalle de confiance on observe un échantillon –> on estime la population.

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

Un biais est la différence systématique entre la valeur moyenne de l'estimateur et la vraie valeur du paramètre. Le cours mentionne des biais de sondage, notamment le biais de couverture, c'est à dire une base qui « n'inclut pas certains éléments » de la population .

5. Comment appelle-t-on une statistique portant sur toute la population ? Faites le lien avec la notion de données massives 1 ?

Une statistique portée sur l'ensemble de la population est appelée statistique exhaustive (recensement). Les données massives sont un cas particulier où l'intégralité (ou une immense part) de la population est disponible, rendant parfois l'échantillonnage moins nécessaire.

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

Le choix d'un estimateur repose sur les propriétés mathématiques (méthode des moments, maximum de vraisemblance), la complexité de mise en œuvre, la qualité de l'ajustement : choix d'une loi pertinente (« choix d'une loi adaptée » selon les critères de forme, paramètres, nature des données) .

Il y a différents enjeux qui incluent l'importance d'assurer la cohérence avec la loi supposée, de minimiser le biais et de minimiser la variance,

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

On retrouve la méthode des moments, la méthode du maximum de vraisemblance et la méthode des rangs. Elles sont sélectionnées selon la loi supposée (discrète, continue), les propriétés désirées (simplicité, absence de biais, efficacité), la disponibilité des données, la facilité de calcul.

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

Le cours évoque les tests d'adéquation : comparer une distribution observée à une loi théorique, les tests d'hypothèse via intervalles de fluctuation ainsi que la méthode GES pour redresser les résultats d'enquête (lié à l'estimation).
Ces tests servent à vérifier si les observations sont compatibles avec un modèle ou une hypothèse (ex : normalité, distribution Poisson…).

On peut créer un test de cette façon :

1. Choisir l'hypothèse nulle
2. Choisir une statistique de test adaptée à la loi supposée
3. Déterminer une règle de décision (seuil, intervalle de fluctuation)
4. Conclure en fonction de la position de l'estimation par rapport à l'intervalle de fluctuation ou à l'intervalle de confiance

9. Que penser des critiques de la statistique inférentielle ?

Le cours évoque plusieurs limites :

- Les biais de couverture en sondage (la base est imparfaite)
- Les méthodes non probabilistes exposées aux biais
- La nécessité de redressement (GES) pour éviter des estimations erronées

Ces limites illustrent les critiques classiques indiquant que les résultats dépendent de la qualité de l'échantillon, que les méthodes supposent souvent des lois théoriques

(normale, Poisson…) et que la mauvaise application des intervalles ou des tests peut conduire à des interprétations erronées.

Analyse de l’exercice de code

Nous sommes passés de la description à l’inférence, la statistique inférentielle permet de comprendre comment tirer des conclusions sur une population entière à partir d'un petit échantillon. En effet, j'ai appris qu'il est souvent impossible d'observer une population entière (recensement exhaustif trop coûteux). On procède donc par échantillonnage. La théorie de l'estimation permet de faire le chemin inverse : partir de l'échantillon pour deviner la population, en construisant un intervalle de confiance. J'ai retenu que la qualité de cette estimation dépend crucialement de l'absence de biais (comme le biais de couverture)
Le script main_session5.py n'a pas produit d'images, mais des résultats chiffrés essentiels dans la console, simulant une enquête d’opinion. J'ai travaillé sur un fichier de sondages simulés (Echantillonnage-100-Echantillons.csv). En calculant les moyennes d'opinions ("Pour", "Contre", "Sans opinion") sur un échantillon de 100 personnes, j'ai obtenu des fréquences observées, pour une opinion à 50%, l'intervalle est large (environ 40% - 60%). Le point crucial a été le calcul de l'intervalle de fluctuation. Avec un échantillon de cette taille, la marge d'erreur calculée par le script est significative. Cela rappelle concrètement qu'un chiffre de sondage n'est jamais une vérité absolue, mais une estimation comprise dans une fourchette. L'application d'un facteur de correction pour population finie (fpc) a permis d'affiner légèrement cet intervalle, une nuance technique importante lorsque l'on sonde une petite population.

Enfin, j'ai réalisé des tests de normalité (Shapiro-Wilk) sur deux séries de données distinctes. Le programme a permis de trancher de manière objective : pour le premier fichier, la p-value était supérieure au seuil de 0,05, nous permettant de considérer la distribution comme normale. Pour le second, la p-value très faible a conduit au rejet de l'hypothèse de normalité et m'a obligée à envisager d'autres modèles de distribution. Cet exercice montre l'importance de vérifier ses hypothèses avant d'appliquer des tests statistiques classiques. On ne peut pas appliquer aveuglément des méthodes

statistiques (comme celles basées sur la moyenne/écart-type) sans vérifier d'abord si la distribution est normale. Le code agit ici pour nous faire comprendre la rigueur scientifique.

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?

La statistique ordinale est explicitement définie comme « le cœur de la géographie humaine » et repose sur l'utilisation de classements d'objets géographiques (villes, régions, phénomènes, etc). Elle porte sur des variables ordinales, c'est à dire des données qualitatives hiérarchisées, où l'on peut établir un ordre croissant ou décroissant. Elle s'oppose à la statistique nominale qui traite des catégories sans ordre (par exemple les catégories socio professionnelles)
Elle utilise les catégorielles ordonnées comme type de variables : rangs, postions, niveaux
La statistique ordinale matérialise directement les hiérarchies spatiales puisque un certain nombre de classements montrent quelle entité a monté ou descendu dans le classement. En géographie elle peut servir au classement des villes selon la population, à la hiérarchie urbaine, l'intensité des risques, les niveaux de richesse…

2. Quel ordre est à privilégier dans les classifications?

Par convention et logique, on privilégie l'ordre naturel croissant.

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

La corrélation des rangs mesure la force et le sens de l'association entre deux classements. Elle regarde si deux classements varient dans le même sens (relation monotone) (ex: classement PIB et classement espérance de vie)
La concordance des classements donne une vision plus globale. Elle concerne la comparaison de plusieurs classements simultanés, elle mesure si plusieurs critères produisent une hiérarchie cohérente.
La corrélation des rangs fait donc une comparaison deux à deux là ou la concordance fait une comparaison de plusieurs classements à la fois

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

Le test de Spearman est fondé sur la corrélation entre les rangs via le coefficient qui lui est attribué (cf cours)Il permet d'évaluer si deux classements sont identiques, inverses ou indépendants. Il n'exige pas de rang ex aequo (correction nécessaire sinon).

Le test de Kendal est lui basé sur le nombre de paires concordantes et discordantes : plus robuste aux ex aequo. Il se généralise à plus de deux classements (coefficient W).

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

Le coefficient de Goodman–Kruskal permet de mesurer l'association entre deux ordres ou deux variables ordinales et indiquer si les classements sont concordants (+1), inverses (–1) ou indépendants (0).Il peut être nul même sans indépendance totale (cas SC = 0).
Le coefficient Q de Yule est un cas particulier de Goodman–Kruskal pour une table de contingence 2×2, il mesure la force d'association entre deux variables dichotomiques :
Q = +1 : association positive parfaite
Q = –1 : association négative parfaite

Q = 0 : absence d’association

Analyse de l’exercice de code

Cette dernière séance a touché au cœur de la géographie quantitative classique : l'étude des classements. La statistique ordinale s'intéresse aux rangs plutôt qu'aux valeurs brutes. Nous avons étudié la Loi Rang-Taille (Zipf), qui postule une relation inverse entre le rang d'une ville et sa taille. Nous avons aussi vu comment mesurer la concordance entre deux classements grâce aux coefficients de Spearman et Kendall

L'exercice le plus formateur a été la vérification de la loi rang-taille sur les superficies des îles. Le script main_session6.py a produit deux visualisations capitales pour vérifier la loi de Zipf sur les surfaces insulaires.
Le premier graphique *rang_taille_lineaire.png* permet de comprendre le piège de l'échelle linéaire. En effet, sous forme d’un L mais presque illisible on ne lit que quelques points tout en haut (les très grandes îles) et tout le reste est écrasé sur l'axe des abscisses. Il montre l'inégalité, mais ne permet pas d'analyser la structure de la hiérarchie.

Le graphique *rang_taille_loglog.png* permet quant à lui la révélation de l'échelle Log-Lo. En passant les axes en logarithmique, la courbe se transforme en une droite décroissante presque parfaite. L’interprétation majeure réside dans la validation empirique de la loi de Zipf. Le fait que les points s'alignent prouve qu'il existe un ordre mathématique caché dans la nature. La distribution des surfaces n'est pas chaotique, elle est fractale. Ce graphique est sans doute le résultat le plus marquant car il rend visible une loi théorique complexe.

Dans un second temps, j'ai comparé les classements des États du monde selon leur population et leur densité. En calculant les corrélations de rangs de Spearman et Kendall, le code m’a permis d’obtenir des coefficients positifs mais modérés (autour de 0,5) qui montre une corrélation positive mais modérée. Cela s'interprète aisément

géographiquement : les pays les plus peuplés (comme la Chine ou l'Inde) sont souvent denses, ce qui tire la corrélation vers le haut. Cependant, des contre-exemples majeurs comme la Russie (immense et peuplée, mais très peu dense) ou Monaco (minuscule et dense) viennent brouiller cette relation. L'analyse des rangs permet ainsi de nuancer l'idée intuitive selon laquelle « plus c'est grand, plus c'est dense ».

Difficultés techniques : La manipulation des rangs en Python a été complexe. Contrairement à un tableur où l'on trie visuellement, il a fallu coder la logique de tri (sort) tout en gardant la trace des noms des pays. J'ai dû faire preuve de rigueur algorithmique pour ne pas mélanger les étiquettes et les valeurs.

Bilan personnel et difficultés rencontrées

Ce parcours d'analyse de données a été pour moi une découverte exigeante. Si la partie théorique prolongeait, à mon sens, plus facilement les cours de géographie, l'ajout d'exercice de code en Python a représenté un véritable défi. Le passage par la séance 2 a été le plus dur, j'ai bloqué pendant un long moment puis les autres séances me sont parues plus abordables.
Ma principale difficulté a résidé dans la préparation des données. J'ai réalisé que les fichiers bruts ne sont jamais directement exploitables. J'ai souvent rencontré des erreurs dues à des types de variables incorrects (par exemple, des chiffres lus comme du texte à cause des guillemets ou des séparateurs décimaux). J'ai dû apprendre à "nettoyer" mes données, notamment en gérant les valeurs manquantes (NaN) qui faussaient mes calculs de moyenne au début.
Cependant, j'ai trouvé une réelle satisfaction dans la visualisation des graphiques. Réussir à produire un graphique qui confirme une intuition théorique (comme la droite de la loi rang-taille) est très gratifiant. Cela m'a permis de comprendre que la statistique n'est pas qu'une affaire de chiffres, mais un outil puissant pour objectiver des réalités

géographiques, confirmant ainsi que la géographie moderne gagne à s'appuyer sur la rigueur des sciences des données.