Question de cours séance 2:

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

La géographie utilise les statistiques pour passer d'une simple méthode descriptive à une science capable de dégager des tendances générales.

- Les statistiques permettent de traiter le hasard, d'identifier des régularités dans les phénomènes spatiaux et humains, et de construire des modèles spatiaux ou socio-spatiaux.
- La géographie française, historiquement influencée par Vidal, était plus méthodologique que expérimentale, mais l'analyse statistique permet désormais d'aller vers une approche scientifique et quantitative.
- Les statistiques complètent le raisonnement explicatif sans le remplacer.

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

Deux visions:

Déterminisme: tout événement a une cause identifiable ; le hasard n'existe pas (Laplace).
Hasard comme cause inconnue: le hasard existe, mais pourra être expliqué par la science à l'avenir (théorie du chaos).

En géographie, le hasard peut être:

Bénin: fluctuations sans impact sur la causalité (loi normale).
Sauvage: imprévisible, moins fréquent (lois non normales).
Finalement, le hasard empêche de prévoir chaque action individuelle mais permet de dégager des tendances générales sur un territoire.

3. Quels sont les types d'information géographique.

Données attributaires: caractéristiques mesurables des entités (population, revenus, températures…).
Données géométriques/morphologiques: formes et structures des ensembles géographiques (topographie, limites administratives…).

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

Collecter ou produire des données fiables: (nomenclatures et méta-données).
Décrire et résumer les données: pour comprendre les tendances et les relations.
Visualiser les informations: (cartes, graphiques) pour l'interprétation.
Appliquer des méthodes statistiques: pour tester des hypothèses ou prévoir des phénomènes.

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

Caractéristique	Statistique descriptive	Statistique explicative
Objectif	Résumer et visualiser les données	Relier une variable dépendante à des variables explicatives
Méthodes	Histogrammes, tableaux, moyenne, écart-type, ACP, AFC, ACM	Régression linéaire, régression logistique, analyse discriminante
Usage	Étape préalable, identification de tendances	Prédiction et modélisation, étude de relations de cause à effet

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

Histogramme: variables continues
Diagramme en secteurs (camembert): variables qualitatives
Boîte à moustache: distribution et dispersion des variables quantitatives
Polygone de fréquence: représentation continue des classes
Courbe cumulative: tendances cumulées
Choix: dépend du type de variable et de l'objectif :

- Visualiser la distribution → histogramme
- Visualiser la proportion → diagramme en secteur
- Comparer plusieurs distributions → boîtes à moustache

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

Descriptives: ACP, AFC, ACM, AFD, AFM, nuées dynamiques
Explicatives: régressions, analyse discriminante, modèles linéaires
Prévision: séries chronologiques, modèles reliant passé et présent

8. Définitions

(a) Population statistique: ensemble d'éléments étudiés (ex. habitants d'une ville).
(b) Individu statistique: élément d'une population (ex. un habitant).
(c) Caractère statistique: propriété ou attribut d'un individu (ex. âge, revenu, altitude).
(d) Modalités statistiques: valeurs possibles d'un caractère (ex. couleur des yeux = bleu, vert, marron).

Types de caractères/variables:

- **Qualitatives**: nominales, ordinales
- **Quantitatives**: discrètes, continues (intervalle ou rapport)
  Hiérarchie: toutes sont des variables, mais les quantitatives permettent plus de traitements statistiques et tests paramétriques que les qualitatives.

9. Mesurer amplitude et densité

Amplitude (A) d'une classe : A=b−aA=b−a (valeur maximale – valeur minimale de la classe)
Densité (d) d'une classe : d=nib−ad=b−ani
où nini = effectif de la classe, b−ab−a = amplitude

10. Formules de Sturges et de Yule

Objectif : déterminer le **nombre optimal de classes** pour discrétiser une variable quantitative.
Formules :

- **Sturges:** k≈1+3,322log10nk≈1+3,322log10n
- **Yule:** k≈2,5n4k≈2,54n

Permettent d'éviter un découpage trop fin ou trop grossier des données.

11. Effectif, fréquence et distribution statistique

Effectif (ni): nombre d'occurrences d'une valeur ou d'une classe
Fréquence relative (fi): fi=ninfi=nni, proportion d'individus pour une modalité
Fréquence cumulée: somme des fréquences (ou effectifs) jusqu'à une valeur ou une classe kk
Distribution statistique: représentation des valeurs (ou classes) d'un caractère et de leur fréquence, permettant d'identifier des tendances et d'associer une loi de probabilité.

Question de cours: Séance 3

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

Le caractère **qualitatif** est le plus général parce qu'il peut décrire toute information non numérique, comme une catégorie, une classe ou une modalité descriptive. Le caractère quantitatif n'est qu'un type particulier de caractère qualitatif dans lequel les modalités sont mesurables numériquement. Autrement dit, tout caractère quantitatif reste un caractère qualitatif, mais l'inverse n'est pas vrai. C'est pourquoi, du point de vue statistique, le qualitatif est plus englobant et plus général.

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

Un **caractère quantitatif discret** est une variable numérique qui ne peut prendre que certaines valeurs spécifiques, souvent entières (par exemple : nombre d'enfants, nombre de logements).
Un **caractère quantitatif continu** peut prendre n'importe quelle valeur dans un intervalle, sans interruption (exemple : revenu, superficie, altitude).

On les distingue car ils ne se traitent pas de la même manière statistiquement : les variables continues nécessitent souvent un regroupement en classes et recourent à la notion de densité, tandis que les variables discrètes utilisent des fréquences simples. Cette distinction conditionne donc les méthodes de calcul des paramètres et leur interprétation.

3. Paramètres de position

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

Il existe plusieurs types de moyenne car chacune répond à une situation particulière. La moyenne arithmétique est adaptée à la plupart des distributions mais elle est sensible aux valeurs extrêmes. La moyenne géométrique est adaptée aux phénomènes multiplicatifs (croissance, indices), la moyenne harmonique aux rapports (vitesses, ratios), et la moyenne quadratique à certains contextes géométriques. Chaque moyenne possède des propriétés mathématiques différentes et convient à des usages spécifiques.

B) Pourquoi calculer une médiane ?

La médiane est utile car elle n'est pas influencée par les valeurs extrêmes, contrairement à la moyenne arithmétique. Elle représente la valeur centrale réelle d'une série, ce qui en fait un indicateur particulièrement pertinent pour les distributions asymétriques ou fortement dispersées.

C) Quand est-il possible de calculer un mode ?

Le mode peut être calculé dès qu'une ou plusieurs modalités apparaissent avec une fréquence maximale. Il existe si une valeur est clairement « dominante ». Toutefois, il peut ne pas exister (si toutes les fréquences sont identiques) ou au contraire être multiple (distribution bimodale ou multimodale).

4. Paramètres de concentration : intérêt de la médiale et de l'indice de Gini

La **médiale** découpe la population en deux groupes ayant chacun 50 % de la masse totale (par exemple 50 % de la masse salariale). Elle permet d'évaluer la concentration des valeurs, car elle tient compte non pas seulement des effectifs mais de l'importance globale des valeurs observées.

L'**indice de Gini**, associé à la courbe de Lorenz, mesure le degré d'inégalité ou de concentration d'une distribution. Plus il est élevé, plus la masse totale est concentrée entre peu d'individus. Ensemble, médiale et Gini offrent une vision approfondie de la façon dont une variable (comme un revenu, une surface ou une production) est distribuée.

5. Paramètres de dispersion

a) Pourquoi calculer une variance plutôt que l'écart simple à la moyenne ? Pourquoi la remplacer par l'écart-type ?

Les écarts simples à la moyenne s'annulent toujours (leur somme vaut zéro) et ne permettent donc pas de mesurer la dispersion. La variance élimine ce problème en élevant les écarts au carré et permet ainsi de mesurer la dispersion globale autour de la moyenne. L'écart-type est ensuite utilisé car il exprime la dispersion dans la **même unité** que la moyenne, ce qui facilite son interprétation.

b) Pourquoi calculer l'étendue ?
L'étendue, qui est la différence entre les valeurs extrêmes, constitue un indicateur simple et rapide de dispersion. Même s'il ne repose que sur deux valeurs, il donne immédiatement une idée de l'amplitude totale des données.
c) À quoi sert-il de créer un quantile ? Quels quantiles sont les plus utilisés ?
Les quantiles servent à découper une distribution en parties égales, ce qui est particulièrement utile pour analyser des distributions asymétriques ou pour résumer l'information sans recourir à la moyenne.
Les quantiles les plus utilisées sont les **quartiles** (Q1, médiane = Q2, Q3), puis les **déciles** (notamment D1 et D9), et les **centiles** dans les grandes populations.

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

La boîte à moustaches permet de représenter visuellement plusieurs paramètres essentiels : la médiane, les quartiles, l'étendue, et éventuellement les valeurs extrêmes. Elle sert à comparer facilement plusieurs distributions et à identifier rapidement asymétrie, dispersion centrale et observations anormales.
On l'interprète en observant la largeur de la boîte (dispersion des 50 % centraux), la position de la médiane, et la longueur des moustaches.

6. Paramètres de forme
a) Différence entre moments centrés et moments absolus. Pourquoi les utiliser ?
Les **moments centrés** sont calculés autour de la moyenne $(x_i - \langle x \rangle)^r$; ils servent à caractériser la variance (moment d'ordre 2), l'asymétrie (ordre 3) et l'aplatissement (ordre 4).
Les **moments absolus**, basés sur $|x_i - a|^r$, sont moins sensibles aux valeurs extrêmes et permettent d'évaluer la dispersion autour d'un point de référence, souvent la médiane.
Ces paramètres permettent de décrire la *forme* d'une distribution au-delà de la simple moyenne et de la variance.

b) Pourquoi vérifier la symétrie d'une distribution et comment faire ?
Il faut s'assurer que la symétrie est essentielle pour choisir les indicateurs adaptés. Une distribution symétrique permet d'utiliser la moyenne comme indicateur central pertinent, tandis qu'une distribution asymétrique justifie l'usage de la médiane.
On vérifie la symétrie en comparant **moyenne, médiane et mode** (égaux en cas de symétrie), en calculant le coefficient d'asymétrie ($\beta_1 = 0 \rightarrow$ symétrie), ou en observant un histogramme ou un boxplot.

Question de cours: 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 ?
Le choix entre une distribution discrète ou une distribution continue repose avant tout sur la **nature du phénomène étudié**. Une variable discrète correspond à un comptage : elle ne peut prendre que des valeurs isolées, souvent entières, comme un nombre d'individus, de logements ou d'occurrences d'un événement. À l'inverse, une variable continue peut

prendre n’importe quelle valeur dans un intervalle : il s’agit de mesures comme la température, la distance, le revenu ou une concentration.
Le premier critère est donc la **nature des valeurs** : sont-elles dénombrables ou mesurées ?
Le second critère concerne la **forme empirique de la distribution** observée dans les données : une distribution asymétrique, très étalée ou fortement concentrée n’appellera pas les mêmes lois qu’une distribution centrée ou régulière.
Un troisième critère est lié aux **caractéristiques statistiques observées** telles que la moyenne, la médiane, la variance, l’asymétrie et l'aplatissement, qui permettent d’orienter le choix vers des lois adaptées.
Enfin, le choix dépend également du **nombre de paramètres** de la loi : certaines lois très flexibles, car dotées de plusieurs paramètres, permettent de modéliser des phénomènes complexes, tandis que d’autres, plus simples, conviennent à des distributions régulières et symétriques.
Ainsi, le choix entre une distribution discrète et une distribution continue n’est pas seulement formel : il conditionne directement la qualité du modèle statistique et son adéquation à la réalité géographique étudiée.

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

En géographie, les lois statistiques les plus utilisées sont celles qui permettent de décrire les phénomènes spatiaux, démographiques, économiques ou environnementaux, souvent marqués par des répartitions inégales, des effets d’échelle et des dynamiques complexes.
Parmi les **lois discrètes**, la loi de Poisson occupe une place importante. Elle est très utilisée pour modéliser des phénomènes rares ou des comptages dans l’espace, comme le nombre d’accidents, de commerces ou d’événements naturels dans une unité spatiale. On utilise aussi fréquemment la loi binomiale, notamment pour les phénomènes de présence/absence ou les sondages.
Certaines lois discrètes, comme la loi de Zipf ou la loi de Zipf-Mandelbrot, sont centrales en géographie urbaine. Elles décrivent les distributions rang–taille des villes ainsi que les hiérarchies urbaines. Ces lois permettent de comprendre comment les populations ou les fonctions urbaines se répartissent dans un système de villes.
Parmi les **lois continues**, la loi normale est très courante, notamment parce que de nombreux phénomènes géographiques résultent de l’addition de processus indépendants, conformément au théorème central limite. Elle s’applique à des variables centrées, symétriques et relativement régulières.
Cependant, la loi log-normale et la loi de Pareto sont probablement encore plus importantes en géographie, car beaucoup de phénomènes spatiaux ou socio-économiques sont asymétriques et présentent de fortes inégalités. La loi log-normale est adaptée aux tailles de villes, aux revenus, aux surfaces ou à des phénomènes de croissance multiplicative. La loi de Pareto est très utilisée pour modéliser les distributions à queue lourde, comme la richesse, les surfaces de lacs ou les tailles de villes.
Enfin, les **lois des valeurs extrêmes** (Gumbel, Fréchet, Weibull) sont essentielles en géographie physique et environnementale pour analyser les crues, sécheresses, tempêtes ou pics de pollution.

Ainsi, la géographie mobilise un ensemble varié de lois, allant des lois de comptage aux lois continues asymétriques, en passant par les lois extrêmes et les distributions hiérarchiques. Leur usage reflète la diversité et la complexité des phénomènes géographiques.

Questions de cours: Séance 5

1. Comment définir l'échantillonnage ? Pourquoi ne pas utiliser la population en entier ?

Définition:

L'échantillonnage est l'opération statistique consistant à **prélever au hasard un sous-ensemble** d'une population mère, appelé échantillon.
L'objectif est de **tirer des conclusions** sur la population totale à partir de ce sous-ensemble, selon les principes de l'inférence statistique.
Un échantillon doit être **représentatif**, c'est-à-dire qu'il doit permettre une généralisation des résultats à l'ensemble de la population, tout en tenant compte de l'erreur d'échantillonnage.

Pourquoi ne pas utiliser la population entière ?

Dans le cours il est mentionné deux raisons majeures :

- **Impossibilité matérielle et logistique:**
  Exemple : impossible de mesurer tous les poissons d'un océan ou de sonder 40 millions de votants.
- **Coût trop élevé:**
  Un recensement exhaustif est souvent trop cher et trop long.
  Exemple : pour une enquête électorale, interroger 1 000 personnes représentatives suffit pour une erreur de 5 %, et 10 000 personnes pour une erreur de 0,3 %.

On étudie un **petit ensemble** pour en inférer des informations sur un **grand ensemble**, car c'est économiquement et techniquement plus réaliste.

2. Quelles sont les méthodes d'échantillonnage ? Comment les choisir ?

Ton cours distingue deux grandes familles :

A) Les méthodes aléatoires

Elles reposent sur un **tirage au sort**, supposent l'existence d'une **base de sondage**, et permettent des traitements probabilistes rigoureux.

1. Tirage aléatoire simple (S.A.S.)

- Chaque individu a la **même probabilité** d'être sélectionné.
- Deux variantes :
  - **Avec remise** (modèle simple pour les calculs, irréaliste en pratique)
  - **Sans remise** (plus réaliste, nécessite de prendre en compte le taux de sondage nNNn)

Exemple :
Tirer 10 étudiants au hasard dans une liste de 500 étudiants pour connaître le temps passé sur un projet.

B) Les méthodes non aléatoires

1. Échantillonnage systématique

- On choisit un pas k=Nnk=nN
- On tire un premier individu au hasard entre 1 et k
- On prend ensuite : d,d+k,d+2k,…d,d+k,d+2k,…

Exemple:
Dans une usine, on inspecte **1 produit toutes les 50 unités produites**.

2. Méthode des quotas

Échantillon construit pour respecter les proportions connues de la population : âge, sexe, CSP…

Exemple:
Si une ville compte 60 % d'adultes et 40 % de seniors, l'échantillon doit respecter ces proportions.

3. Méthodes Monte-Carlo

Basées sur un générateur pseudo-aléatoire pour simuler des tirages et estimer une quantité difficile à calculer analytiquement.

Exemple:

- Pour estimer l'intégrale: ∫01f(x) dx,∫01f(x)dx,
- On prend des valeurs aléatoires XiXi uniformes, et on calcule:
  1n∑i=1nf(Xi).n1i=1∑nf(Xi)

Comment les choisir ?

- Si une **base de sondage** est disponible → tirage aléatoire simple.
- Si la base de sondage est inaccessible → quotas ou systématique.
- Si les modèles sont **intraitables analytiquement** → Monte-Carlo.
- Si les répétitions coûtent peu → favoriser les méthodes aléatoires.

3. Comment définir un estimateur et une estimation ?

Estimateur

Un estimateur est une **statistique**, c'est-à-dire une fonction des variables aléatoires d'un échantillon : θ^=y(X1,…,Xn)θ^=y(X1,…,Xn)
Il sert donc à approcher un paramètre inconnu.

Estimation

L'estimation est la **valeur numérique** de l'estimateur issue de l'échantillon observé : θn=y(x1,…,xn)θn=y(x1,…,xn)

Exemple:
Pour la moyenne d'une population :

- Estimateur :
  X¯=1n∑i=1nXiX¯=n1i=1∑nXi

- Estimation pour les observations (3,5,4,6) :
  $x^{-}=3+5+4+64=4.5x^{-}=43+5+4+6=4.5$

4. Comment distinguer l'intervalle de fluctuation et l'intervalle de confiance ?

Intervalle de fluctuation

- Le paramètre pp est **connu**.
- On étudie la variabilité d'une **fréquence observée** FnFn.
- Utilisé pour la **prise de décision** lors d'un tirage.

[p−zCp(1−p)n, p+zCp(1−p)n][p−zCnp(1−p),p+zCnp(1−p)]

Exemple :

Proportion théorique de boules blanches : p=0.3p=0.3.
Pour n=50n=50, l'intervalle est [0.173;0.427][0.173;0.427].

Intervalle de confiance

- Le paramètre est **inconnu**.
- On cherche à encadrer la **vraie valeur** du paramètre.
- Exemple pour la moyenne : μ^±tα/2σ^nμ^±tα/2nσ^

Exemple:

Moyenne observée = 12 ; σ estimé = 3 ; n = 36.
Au seuil 95 % :
12±1.96×36=12±0.9812±1.96×63=12±0.98

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

Le biais d'un estimateur θ^θ^ est la différence entre son espérance et la vraie valeur du paramètre :
Biais=E(θ^)−θBiais=E(θ^)−θ

Sans biais

E(θ^)=θE(θ^)=θ

Exemples

- La moyenne est **sans biais** : $E(X^{-})=\mu E(X^{-})=\mu$
- La variance empirique est **biaisée** :
  E(s^2)=n−1nσ2E(s^2)=nn−1σ2

La correction du biais :
s^∗2=1n−1∑i=1n(Xi−X−)2s^∗2=n−11i=1∑n(Xi−X−)2

6. Comment appelle-t-on une statistique travaillant sur la population totale ? Lien avec les données massives ?

Lorsqu'une statistique travaille sur **toute la population**, il s'agit d'un **recensement** ou **statistique exhaustive**.

Lien avec les données massives

Le Big Data permet parfois d'approcher un recensement car les volumes sont très importants.
Cependant, le cours rappelle que la population totale est **presque toujours inaccessible**, d'où la nécessité de l'échantillonnage.

7. Quels sont les enjeux autour du choix d'un estimateur ? Quelles sont les méthodes d'estimation ? Comment en sélectionner une ?

Enjeux :

Un bon estimateur doit être :

- **sans biais**
- **efficient** (variance minimale)
- **convergent**
- **robuste** si possibilité de valeurs aberrantes
- **exhaustif** (ne pas perdre d'information)

La borne de **Cramér–Rao** rappelle qu'un estimateur sans biais ne peut pas avoir une variance inférieure à : 1In(θ)In(θ)1 ou In(θ)In(θ) est l'information de Fisher.

Méthodes d'estimation (selon le cours) :

- statistique exhaustive
- méthode du maximum de vraisemblance
- méthode des moments
- méthodes robustes :
    - médianes
    - moyennes tronquées
    - M-estimateurs (Huber, bicarré)
    - median polish (pour plans à deux voies)

Comment choisir ?

- **Si modèle théorique bien défini :** maximum de vraisemblance.
- **Si présence d'aberrants :** méthodes robustes.
- **Si besoin de minimiser la variance :** un estimateur sans biais de variance minimale (théorème de Lehmann-Scheffé).
- **Si données massives :** choisir un estimateur peu sensible aux dégradations de l'information.

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

Bien que le cours n'énumère pas les tests classiques, il décrit leur **fonctionnement général**.

À quoi servent les tests ?

- Prendre une décision en contrôlant un **risque d'erreur**.
- Comparer une statistique observée à une **loi théorique**.
- Vérifier si une fréquence appartient à un **intervalle de fluctuation**.

Comment créer un test ?

1. **Choisir l'hypothèse** à tester (ex : proportion p).

2. Identifier l'**estimateur** associé.
3. Déterminer sa **loi d'échantillonnage**.
4. Fixer un **seuil de décision** (par ex. 5 %, z = 1,96).
5. Construire la **règle :**
    - rejeter l'hypothèse si la statistique sort de l'intervalle critique
    - l'accepter sinon

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

Ton cours mentionne plusieurs limites importantes :

- **Fluctuation d'échantillonnage** : un échantillon peut être non représentatif.
- **Hypothèses du modèle** (normalité, indépendance, etc.) parfois irréalistes.
- **Sensibilité aux valeurs aberrantes** → nécessité d'estimateurs robustes.
- **Biais possible** (échantillon biaisé, variance empirique biaisée).
- **Perte d'information** si la statistique n'est pas exhaustive.

Malgré ces critiques, la statistique inférentielle reste indispensable : **la population totale est rarement accessible**, et les estimateurs convergents garantissent des décisions rationnelles.

Questions cours: séance 6

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

Une **statistique ordinale** est une statistique qui repose sur le **classement** des données selon un **ordre naturel**, généralement croissant. Elle consiste à **ordonner une série finie d'observations** X1,…,XnX1,…,Xn pour obtenir :
X(1)≤X(2)≤⋯≤X(n)X(1)≤X(2)≤⋯≤X(n)
où X(1)X(1) est la plus petite valeur, X(n)X(n) la plus grande, et chaque X(k)X(k) représente la **statistique d'ordre** de rang kk.
Cette façon d'étudier les données est au cœur de la géographie, car elle permet d'identifier :

- les **valeurs extrêmes**,
- les **valeurs aberrantes**,
- la **répartition hiérarchique** des phénomènes.

Dans le cours, cette statistique est particulièrement mise en avant dans l'étude des **classements géographiques**, omniprésents en géographie humaine et physique.

Opposition avec une autre statistique catégorielle:
La statistique ordinale s'oppose aux **statistiques nominales** (même si le terme n'est pas explicitement utilisé dans le texte) :

- **Statistique ordinale:** les catégories sont **ordonnées**.
- **Statistique nominale:** les catégories ne possèdent **pas d'ordre**.

Les statistiques ordinales sont donc plus riches, car elles permettent des comparaisons hiérarchiques.

Type de variables utilisées:

Elle utilise des **variables ordinales**, c'est-à-dire des variables prenant des valeurs qui peuvent être **classées** :

- rangs obtenus par classement,
- positions hiérarchiques,
- indices ordonnés,
- valeur d'un critère permettant un tri.

Matérialisation d'une hiérarchie spatiale:

Ton cours insiste sur ce point : la statistique d'ordre est **au cœur de la géographie humaine**, car les sociétés produisent constamment des **classements d'espaces**, que ce soit :

- rang des villes selon leur population,
- classements des régions selon le revenu,
- classement des quartiers selon l'attractivité,
- hiérarchies de flux,
- hiérarchie urbaine (loi rang–taille).

Ainsi, les mesures ordinales permettent de décrire **la structuration hiérarchique de l'espace**, élément fondamental de la géographie.

2. Quel ordre est à privilégier dans les classifications ?

Le cours est explicite:

L'ordre croissant (ou l'ordre naturel) doit être privilégié dans les classifications.
Cet ordre présente plusieurs avantages :

- il permet d'identifier immédiatement la structure de la distribution,
- il facilite la recherche des valeurs extrêmes :
  - minimum : X(1)X(1),
  - maximum : X(n)X(n),
- il permet d'étudier facilement la **loi de la plus grande valeur**, notamment en géographie physique (crues maximales, séismes les plus élevés…).

Ton cours mentionne que certaines exceptions existent, notamment la **loi rang–taille** (rapport décroissant entre la taille des villes et leur rang), mais elles ne remettent pas en cause le principe général.

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

Corrélation des rangs:

La corrélation des rangs mesure la **relation statistique** entre deux classements portant sur les mêmes objets.
Elle repose sur :

- les différences de rang (test de Spearman),
- ou la comparaison des paires concordantes/discordantes (test de Kendall).

La corrélation répond à la question :

Dans quelle mesure les deux ensembles de rangs évoluent-ils ensemble ?

Une corrélation positive indique que les deux classements donnent des rangs proches.

Concordance de classements

La concordance s'intéresse à savoir si les classements sont **identiques ou non**.
Elle est mesurée notamment par:

- le coefficient τ de Kendall,
- le coefficient W de Kendall (pour *p* classements),
- le coefficient Γ de Goodman-Kruskal.

Elle ne cherche pas seulement une relation monotone, mais une **similarité directe des classements**.
La question sous-jacente est :
Les classements sont-ils équivalents, opposés ou indépendants ?

Résumé de la différence essentielle

Concept	Corrélation des rangs	Concordance des classements
Objectif	Mesurer la liaison monotone entre deux rangs	Vérifier si les classements sont identiques
Méthodes	Spearman, Kendall τ	Kendall τ, Kendall W, Γ, Q
S'intéresse à	Force de l'association	Alignement des classements
Logique	Statistique	Hiérarchique

En géographie humaine, les deux approches sont utiles :

- corrélation pour étudier des relations entre indicateurs,
- concordance pour comparer différents classements urbains ou régionaux.

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

Les deux tests comparent des classements, mais selon des **logiques différentes**, comme le détaille ton cours.

Test de Spearman (1904)
Fondements:

- Basé sur la **covariance des rangs**.
- Définition du coefficient :
  $rs=1-6\sum(ui-vi)2n(n2-1)rs=1-n(n2-1)6\sum(ui-vi)2$

Caractéristiques:

- Test **non paramétrique**.
- Ne doit pas comporter d'ex-aequo (sinon correction).
- Pour $n>30n>30$, rsrs se rapproche d'une **loi normale** :
  $t=n-2\ rs1-rs2t=1-rs2n-2rs$

Interprétation:

- rs=1rs=1 : rangs identiques.
- rs=−1rs=−1 : rangs inverses.
- rs=0rs=0 : indépendance.

Test de Kendall (1938)

Fondements:

- Repose sur les **paires de rangs** :
  - concordantes : NaNa,
  - discordantes : NdNd.
- Coefficient :
  τ=Na−NdNa+Ndτ=Na+NdNa−Nd

Caractéristiques:

- Très interprétable : compare directement l'ordre des couples.
- La variance théorique est connue :
  V(τ)=2(2n+5)9n(n−1)V(τ)=9n(n−1)2(2n+5)
- Pour n≥8n≥8, τ suit approximativement une **loi normale**.

Interprétation:

- τ = 1 : concordance parfaite.
- τ = -1 : inversion parfaite.
- τ = 0 : absence de concordance.

Résumé

Aspect	Spearman	Kendall
Logique	Différence des rangs	Concordances/discordances
Formule	∑(ui−vi)2∑(ui−vi)2	Na−NdNa−Nd
Interprétation	Relation monotone	Similitude d'ordre
Robustesse	Moins robuste	Plus robuste aux anomalies

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

Coefficient Γ de Goodman-Kruskal

Γ=Na−NdNa+NdΓ=Na+NdNa−Nd

Ce coefficient mesure:

Le "surplus" de paires concordantes par rapport aux paires discordantes.

Il permet :

- d'évaluer **l'association** entre deux classements,
- de tester si deux ordres sont alignés,
- de compléter les approches de Spearman et Kendall.

Utilité :

- analyse de dépendance,
- comparaison de classements en géographie,
- mesure non paramétrique robuste.

Un test de Student peut être utilisé pour en tester la significativité.

Coefficient Q de Yule

Appliqué aux **tables 2×2**, cas particulier du précédent.

$Q=ad-bcad+bcQ=ad+bcad-bc$

Il mesure :

➡️ **l'association entre deux variables dichotomiques.**

Il varie entre −1 (association négative parfaite) et +1 (association positive parfaite).
Il se relie aussi à l'odds ratio :

$Q=OU-1OU+1Q=OU+1OU-1$

Utilité :

- tableaux croisés simples,
- association entre deux réponses binaires,
- analyse de dépendance dans des systèmes simples (ex. oui/non, présent/absent).

Synthèse question de cours séance 2:

La géographie, longtemps marquée par un héritage descriptif et qualitative issu de la tradition vidalienne, a progressivement consolidé son statut de science en intégrant les outils statistiques. L'usage des statistiques permet en effet de dépasser la simple observation des faits spatiaux pour identifier des régularités, dégager des structures et modéliser les

phénomènes socio-spatiaux. Grâce à elles, la géographie peut traiter le hasard, puisqu'elles offrent un cadre pour distinguer les fluctuations aléatoires des tendances significatives, tout en reconnaissant que certains phénomènes restent soumis à des formes de hasard « bénin » – des variations normales, modélisables – ou à un hasard « sauvage », rare et imprévisible. Dans cette perspective, les statistiques n'annulent pas la complexité des comportements individuels, mais elles permettent de dégager des tendances générales à l'échelle des territoires.

L'information géographique repose sur deux grandes catégories de données. Les données attributaires regroupent toutes les caractéristiques sociales, économiques, démographiques ou environnementales mesurées sur les entités spatiales. Les données géométriques ou morphologiques concernent les formes, les limites, les superficies ou les structures physiques des objets géographiques. Leur analyse exige des données fiables, correctement normées et accompagnées de métadonnées, afin de permettre une interprétation rigoureuse. La démarche statistique en géographie requiert ainsi une chaîne complète : collecte ou production de données, description et synthèse, visualisation cartographique ou graphique, puis application de méthodes statistiques adaptées à la nature des questions de recherche.

On distingue classiquement la statistique descriptive, qui vise à résumer l'information disponible à travers des indicateurs (moyenne, écart-type), des tableaux, des visualisations ou encore des méthodes d'exploration comme les ACP ou les analyses factorielles, et la statistique explicative, centrée sur les relations entre variables à l'aide de modèles tels que les régressions ou l'analyse discriminante. Les deux approches sont complémentaires : la première permet d'explorer et de comprendre les données, tandis que la seconde cherche à établir des liens causaux ou à prévoir des valeurs.

La visualisation occupe une place essentielle dans ce processus. Selon la nature des variables et les objectifs de l'analyse, on privilégiera des histogrammes pour les distributions quantitatives, des diagrammes en secteurs pour les proportions de modalités qualitatives, des boîtes à moustaches pour comparer des dispersions ou encore des polygones de fréquence et courbes cumulées pour suivre des évolutions continues. Le choix d'une représentation dépend donc du type de variable (qualitative ou quantitative, discrète ou continue) et du message que le chercheur souhaite faire émerger.

L'analyse statistique implique également une connaissance des concepts fondamentaux comme la population et les individus statistiques, les caractères et leurs modalités, ainsi que la distinction entre variables qualitatives et quantitatives. Cette dernière conditionne les traitements possibles, les variables quantitatives permettant notamment le recours à des tests paramétriques et à un éventail plus large de modèles explicatifs. Pour comprendre une distribution, on mobilise des notions comme l'effectif, la fréquence relative, la fréquence cumulée, ainsi que l'amplitude des classes et leur densité, notamment lors de la construction d'histogrammes. Les formules de Sturges et de Yule aident à déterminer un nombre optimal de classes lors de la discrétisation des données continues, garantissant un découpage ni trop fin ni trop grossier.

Enfin, dans une perspective plus large, les méthodes disponibles en géographie se répartissent entre analyses descriptives (ACP, AFC, ACM, nuées dynamiques, analyses

factorielles variées), méthodes explicatives (régressions, modèles discriminants, modélisation statistique) et outils de prévision (séries chronologiques, modèles reliant état passé et état actuel des phénomènes). Ces outils permettent au géographe d'explorer la structure interne des données, de tester des hypothèses, de modéliser les comportements spatiaux et d'anticiper certaines dynamiques territoriales. L'ensemble constitue une boîte à outils indispensable pour appréhender la complexité du réel et proposer une lecture scientifique des phénomènes spatialisés.

Synthèse question de cours séance 3

Dans l'analyse statistique appliquée à la géographie, la première distinction fondamentale concerne la nature des caractères étudiés. Le caractère qualitatif est considéré comme le plus général, car il regroupe l'ensemble des propriétés descriptives attribuables à des individus, qu'elles soient catégorielles, ordinales ou même codées numériquement sans signification métrique. Le caractère quantitatif n'en constitue en réalité qu'un cas particulier : il correspond à une situation où les modalités du caractère sont mesurables par des nombres et organisables selon une échelle numérique rigoureuse. Ainsi, tout caractère quantitatif peut être vu comme un caractère qualitatif dont les modalités sont des valeurs numériques, mais l'inverse n'est jamais vrai. Cette hiérarchie explique pourquoi la catégorie « qualitatif » est la plus englobante d'un point de vue statistique.

Parmi les caractères quantitatifs, on distingue les variables discrètes, qui prennent un nombre limité de valeurs distinctes (comme le nombre d'enfants ou le nombre de logements), et les variables continues, susceptibles d'adopter une infinité de valeurs dans un intervalle (comme une altitude ou un revenu). Cette distinction n'est pas seulement conceptuelle : elle conditionne les traitements statistiques possibles. Les variables continues nécessitent souvent d'être regroupées en classes afin d'être analysées ou représentées, ce qui fait intervenir des notions comme l'amplitude et la densité. Les variables discrètes, quant à elles, se décrivent par des effectifs directs et des fréquences simples. La manière de résumer et de modéliser une distribution dépend donc fortement du type de caractère étudié.

Pour caractériser une distribution, les statisticiens utilisent divers paramètres de position, qui permettent de saisir le « centre » d'une série. La multiplicité des moyennes s'explique par la diversité des phénomènes observés : la moyenne arithmétique est adaptée aux situations additives et largement utilisée, mais elle demeure sensible aux valeurs extrêmes. La moyenne géométrique correspond plutôt aux processus multiplicatifs ou aux taux de croissance, la moyenne harmonique est pertinente pour les phénomènes exprimés en ratios, tandis que la moyenne quadratique apparaît dans certains contextes géométriques ou physiques. La médiane, elle, conserve un rôle particulier : insensible aux valeurs extrêmes, elle indique la valeur centrale réelle de la distribution et devient essentielle lorsque la série est asymétrique ou présente des valeurs atypiques. Le mode, enfin, correspond à la

modalité la plus fréquente ; il n'existe que lorsque l'une ou plusieurs des valeurs se répètent suffisamment pour constituer une dominante. Son utilité est grande dans les distributions multimodales, mais il peut être absent si toutes les modalités sont équivalentes.

L'étude des données exige également de comprendre leur concentration et leur dispersion. Parmi les indicateurs de concentration, la médiale découpe une population en deux groupes contenant chacun une moitié de la masse totale (par exemple 50 % du revenu global). Elle permet de saisir non pas seulement le nombre d'individus, mais la façon dont la « masse » d'un phénomène est répartie. L'indice de Gini, mesuré à partir de la courbe de Lorenz, quantifie ce degré d'inégalité : plus il est élevé, plus une faible partie de la population concentre une grande part de la masse observée. Ces outils sont particulièrement utiles pour étudier des phénomènes comme les revenus, les superficies agricoles ou la distribution des équipements.

La dispersion quant à elle s'appréhende par différents paramètres. Les écarts simples à la moyenne ne peuvent être utilisés directement, car leur somme est toujours nulle ; c'est pourquoi la variance est introduite : en élevant au carré les écarts, elle mesure effectivement la dispersion totale autour de la moyenne. Toutefois, la variance est exprimée dans une unité au carré, ce qui rend l'écart-type plus intuitif car il revient à l'unité initiale et facilite les comparaisons. L'étendue, qui correspond à la différence entre les valeurs extrêmes, fournit une information immédiate sur l'amplitude totale, bien qu'elle soit très sensible aux valeurs isolées. Les quantiles – quartiles, déciles, centiles – divisent la distribution en parties égales et résument efficacement les séries asymétriques. Leur importance est grande en géographie, notamment pour comparer des catégories socio-économiques ou analyser des distributions de prix ou de surfaces. La boîte à moustaches synthétise plusieurs de ces éléments visuellement : dispersion centrale, position de la médiane, étendue et valeurs atypiques. Elle est particulièrement utile pour comparer plusieurs distributions entre elles ou identifier des asymétries importantes.

Enfin, les paramètres de forme permettent de caractériser la distribution au-delà de sa position et de sa dispersion. Les moments centrés, calculés par rapport à la moyenne, renseignent sur la variance (moment d'ordre deux), l'asymétrie (ordre trois) ou l'aplatissement (ordre quatre). Les moments absolus, construits à partir des écarts absolus, sont moins sensibles aux valeurs extrêmes et souvent préférés lorsque l'on se réfère à la médiane plutôt qu'à la moyenne. Vérifier la symétrie d'une distribution est une étape essentielle : dans une distribution symétrique, moyenne, médiane et mode coïncident, et les indicateurs classiques sont pleinement pertinents. En revanche, dans une distribution asymétrique, la médiane devient souvent un meilleur indicateur central que la moyenne. L'asymétrie peut être identifiée par des coefficients normés, par la comparaison des paramètres de position ou par une simple inspection visuelle d'un histogramme ou d'un boxplot.

Synthèse question de cours Séance 4:

Le choix entre une distribution statistique avec des variables discrètes ou continues repose avant tout sur la nature du phénomène étudié et sur les propriétés des données disponibles.

Les variables discrètes correspondent généralement à des comptages : elles ne peuvent prendre que des valeurs isolées, souvent entières, comme le nombre d'individus, de logements ou d'occurrences d'un événement particulier. À l'inverse, les variables continues représentent des mesures susceptibles de varier de manière infiniment fine dans un intervalle donné, comme la température, la distance, le revenu ou la concentration d'un polluant. Le premier critère de choix est donc la nature des valeurs : sont-elles dénombrables ou mesurables ? Ce critère conditionne directement le type de traitement statistique et la forme de la distribution.

Outre la nature des données, la structure empirique de la distribution influence également ce choix. Une distribution asymétrique, fortement concentrée ou très étalée n'appellera pas les mêmes lois que des distributions régulières et symétriques. Les caractéristiques statistiques observées, telles que la moyenne, la médiane, la variance, l'asymétrie ou l'aplatissement, orientent le choix vers des lois adaptées et fiables pour modéliser correctement les phénomènes étudiés. De même, le nombre de paramètres d'une loi statistique joue un rôle : certaines lois à plusieurs paramètres sont plus flexibles et permettent de représenter des phénomènes complexes, tandis que des lois plus simples sont suffisantes pour des distributions régulières et symétriques. Le choix entre une distribution discrète ou continue ne relève donc pas d'une simple formalité, mais conditionne directement la qualité et l'adéquation du modèle statistique à la réalité géographique.

En géographie, les lois statistiques les plus mobilisées dépendent des types de variables et de la nature des phénomènes. Pour les variables discrètes, la loi de Poisson est largement utilisée afin de modéliser des événements rares ou des comptages dans l'espace, tels que le nombre d'accidents, d'incidents naturels ou d'équipements urbains par unité territoriale. La loi binomiale sert principalement pour les phénomènes de présence/absence ou les résultats d'échantillonnages, tandis que certaines lois spécifiques, comme la loi de Zipf ou la loi de Zipf-Mandelbrot, sont essentielles en géographie urbaine pour décrire les distributions rang–taille des villes et la hiérarchie des systèmes urbains. Ces lois permettent de comprendre la répartition des populations, des fonctions urbaines ou des infrastructures dans un système de villes hiérarchisé.

Pour les variables continues, la loi normale occupe une place centrale, car elle décrit de nombreux phénomènes résultant de l'addition de processus indépendants, conformément au théorème central limite. Elle est particulièrement adaptée aux variables symétriques et centrées. Cependant, en géographie, de nombreux phénomènes présentent des distributions asymétriques avec des inégalités importantes. La loi log-normale est ainsi très utilisée pour modéliser des variables comme la taille des villes, les revenus, les surfaces ou les phénomènes de croissance multiplicative. De même, la loi de Pareto est employée pour représenter les distributions à queue lourde, typiques des richesses, des superficies ou des tailles de villes. Enfin, les lois des valeurs extrêmes (Gumbel, Fréchet, Weibull) sont indispensables pour analyser les événements rares et extrêmes, notamment dans la géographie physique et environnementale, tels que les crues, sécheresses, tempêtes ou pics de pollution.

Dans l'ensemble, la géographie s'appuie sur un large éventail de lois statistiques, allant des lois de comptage aux lois continues asymétriques et aux lois extrêmes, en passant par les distributions hiérarchiques. Cette diversité reflète la complexité et la variété des

phénomènes étudiés, qu'ils soient humains, économiques ou environnementaux, et souligne l'importance d'adapter les outils statistiques à la nature précise des données et des questions de recherche.