DOSSIER ANALYSE DE DONNÉES
Choix du parcours : débutants

Séance 2 à 6

Séance 2 – Principes généraux de la statistique

1. La géographie utilise les statistiques pour décrire, analyser et comprendre les phénomènes dans l'espace. Les nombres aident à quantifier ce que l'on observe (distribution de populations, flux, densités…), mais la géographie garde une dimension propre : elle interprète ces chiffres en tenant compte du contexte spatial, social ou environnemental. En résumé, la statistique est un outil essentiel, mais la géographie reste une science de l'espace.

2. Le hasard existe en géographie mais seulement en partie. Certains événements sont aléatoires (aléas naturels, accidents, événements ponctuels), mais la plupart des structures spatiales résultent de mécanismes explicatifs : interactions, distances, attractivité, organisation des territoires. La géographie cherche justement à distinguer ce qui relève du hasard et ce qui s'explique.

3. Pour les types d'information en géographie, on distingue :
- quantitatives (mesurables : surface, densité, altitude),
- qualitatives (catégories : type de sol, classes d'usage),
- nominales, ordinales, intervalle, ratio,
  et surtout spatialisées (toute donnée géographique est localisée).

4. La géographie a besoin d'outils pour :
- décrire les répartitions spatiales,
- repérer les corrélations,
- modéliser les structures,
- produire des cartes,
- orienter des décisions (urbanisme, environnement…).

6. Pour visualiser cela en géographie, on mobilise :
- cartes thématiques, symboles proportionnels, densités,
- diagrammes (barres, secteurs, histogrammes),
- nuages de points, courbes.
  Le choix dépend du type de variable et du but recherché (décrire, comparer, expliquer).

7. La géographie utilise : la statistique descriptive, analyse spatiale (Moran), analyse multivariée (ACP, CAH, AFC), modélisation, étude de distributions, etc.

8.
- Population : ensemble étudié (toutes les communes).
- Individu : une unité de cette population.
- Caractère : propriété mesurée (densité).
- Modalité : valeur ou catégorie du caractère.
  → Les caractères peuvent être qualitatifs ou quantitatifs, discrets ou continus.

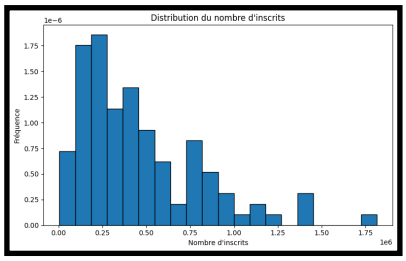
9. Amplitude et densité
- Amplitude = différence max – min.
- Densité = fréquence / amplitude (utile pour les histogrammes normalisés).

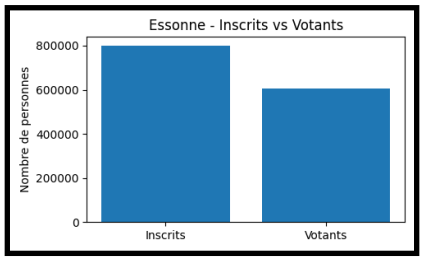
10. Sturges et Yule, sont des formules qui aident à choisir le bon nombre de classes dans un histogramme : ni trop fines (bruit), ni trop larges (perte d'information).

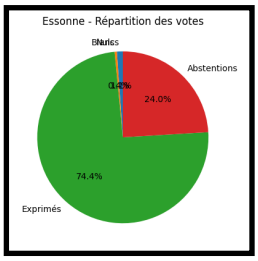
11.
- Effectif : nombre d'individus dans une modalité.
- Fréquence : proportion correspondante.
- Distribution : manière dont ces effectifs se répartissent.

Dans les exercices, les données ne suivent pas la loi normale : elles sont asymétriques, concentrées, parfois issues de lois discrètes (Poisson, uniforme). D'où l'importance de choisir les bons tests.

-> analyse des résultats :
Mes résultats montrent que la plupart des variables étudiées ont une distribution très asymétrique, avec beaucoup de petites valeurs et quelques observations extrêmes. Cela met en évidence une forte hiérarchie spatiale, typique des données démographiques ou territoriales. Les distributions ne suivent pas la loi normale : elles sont plutôt Gamma / lognormales, ce qui correspond aux phénomènes géographiques réels. Les diagrammes en barres confirment ces contrastes entre territoires et révèlent une organisation inégale de l'espace.







Si nous prenons l'exemple de l'Essonne:

Les graphiques montrent que la participation électorale dans l'Essonne est majoritaire, même si une part importante d'abstention demeure. Le contraste entre le nombre d'inscrits et de votants met bien en évidence cette différence. L'histogramme du nombre d'inscrits révèle une distribution très asymétrique : beaucoup de petits bureaux et quelques très grands, ce qui reflète une organisation territoriale inégale. Cette structure irrégulière est typique des données démographiques locales et explique en partie les variations observées dans la participation

Séance 3 – Statistiques descriptives

1. Un caractère quantitatif est mesurable (âge, revenu). Un caractère qualitatif classe les individus (sexe, type). Les qualitatifs sont plus généraux : on peut toujours classer, mais pas toujours mesurer.

2.
- Discret : valeurs dénombrables (nombre d'enfants).
- Continu : valeurs continues (surface, poids).
  Cette distinction détermine les méthodes statistiques et les tests utilisables.

3.
- Moyenne : sensible aux valeurs extrêmes.
- Médiane : robuste, utile si distribution asymétrique.
- Mode : valeur la plus fréquente.

4. Les paramètres de concentration sont: Médiale, Gini : mesurent la répartition, l'inégalité ou la concentration d'une variable.

5. Les paramètres de dispersion sont: Variance, écart-type, étendue, quantiles, boxplots : ils mesurent l'étalement et l'hétérogénéité des valeurs.

6. Les paramètres de forme décrivent la symétrie, l'asymétrie ou l'aplatissement d'une distribution (skewness, kurtosis). Leur analyse aide à choisir les outils statistiques appropriés.

->analyse des bloxpots

1: Boxplot des *Inscrits*
Il y a une très grande dispersion entre communes / bureaux :
→ certaines zones rurales ont 50–200 inscrits
→ les grandes villes peuvent avoir 1000–3000 inscrits par bureau
Distribution fortement asymétrique à droite (beaucoup de petits bureaux, quelques très gros)
De nombreux outliers :
→ grandes villes, arrondissements parisiens, grandes métropoles

Le boxplot montre un système électoral très hétérogène.
Les moustaches sont longues, la médiane basse.

2) Boxplot des *Votants*
Comme les inscrits, mais légèrement « resserré ».

Même asymétrie à droite
Beaucoup d'outliers dans les grandes villes
Médiane plutôt basse (les petits bureaux sont majoritaires)

La majorité des bureaux ont relativement peu de votants.

Les gros bureaux créent de longs segments dans le boxplot.

3) Boxplot des *Abstentions*
Comme votants-inscrits, dépend de la taille du bureau.
Il y a une grande dispersion liée aux tailles très différentes des bureaux.
La médiane assez basse mais moustaches longues.

Les abstentions en valeur absolue ne sont pas comparables entre territoires.
Le boxplot montre uniquement l'effet de la taille des bureaux, pas de l'abstention réelle.

4) Boxplot des *Blancs* et *Nuls*

Valeurs globalement très petites
Beaucoup de zéros
Quelques outliers (grands bureaux)
Distribution très asymétrique

Les bulletins blancs/nuls sont rares.
Le boxplot est probablement très compressé près du zéro.

5) Boxplots des voix par candidat
Pour chaque candidat, le code génère un boxplot :

Il y a : Beaucoup de zéros ou petites valeurs
→ Dans la majorité des bureaux, un candidat faible peut avoir 0 à 10 voix
→ Les candidats majeurs ont une médiane plus élevée
La distribution est très asymétrique
→ Quelques bureaux donnent beaucoup de voix (grandes villes)
Les outliers sont nombreux
→ quartiers spécifiques où un candidat est fort (ex. Mélenchon en ville, Le Pen dans certains territoires)
Les candidats "petits" : les boxplots sont très serrés avec une queue très longue.
Les candidats majeurs : une boîte plus large, médiane plus haute, toujours asymétrie à droite.

6) Boxplot des *Exprimés*

Très similaire à *Votants* moins *Blancs/Nuls*
Distribution proportionnelle à la taille du bureau
Outliers majoritairement dans les grandes villes

Ainsi, les distributions sont toutes fortement asymétriques
→ Les bureaux urbains créent de grands écarts.
→ Les moustaches droites sont très longues.
Il y a énormément d'outliers dans presque toutes les variables

Les boxplots des candidats principaux montrent une forte variabilité
→ Macron, Le Pen, Mélenchon en particulier
→ reflète des disparités territoriales massives

Séance 4 – Distributions statistiques

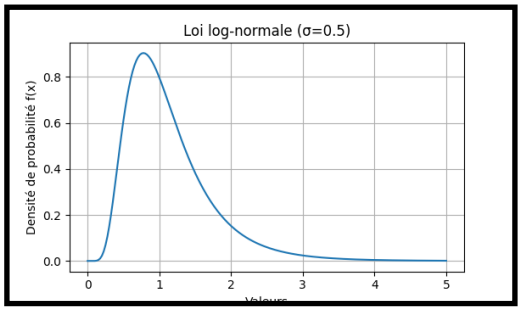
1. Le type de variable dicte le type de distribution :
   - si c'est un comptage → loi discrète (Poisson, binomiale),
   - si c'est une mesure → loi continue (Gamma, lognormale…).

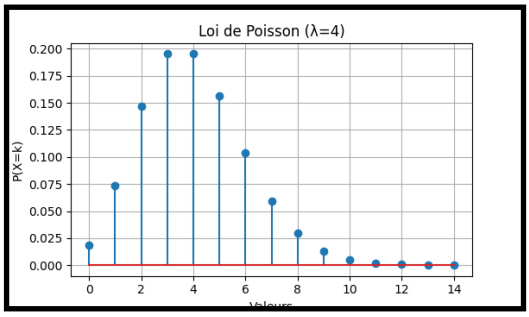
On tient aussi compte de la forme de la distribution (symétrie, positivité, étalement).

2. La géographie exploite essentiellement des lois continues positives :
   - la loi Gamma (surfaces, tailles, hiérarchies),
   - la lognormale (villes, revenus),
   - la normale pour certains phénomènes symétriques.

Les lois discrètes apparaissent pour les phénomènes de comptage.

Analyse des graphiques obtenues :
-> Les résultats montrent clairement que chaque loi de probabilité a sa propre "personnalité". Certaines, comme la loi normale ou la binomiale, forment une jolie courbe en cloche, bien équilibrée autour de leur moyenne. D'autres, comme la Poisson ou le chi2, sont plus décalées et penchent vers la droite. Et puis il y a les distributions vraiment extrêmes, comme la log-normale, la loi de Pareto ou la loi de Zipf, qui donnent beaucoup de petites valeurs et quelques très grandes, très rares mais très importantes. À l'inverse, les lois uniformes répartissent tout de façon égale, tandis que la loi de Dirac ne laisse aucune place au hasard. Ensemble, ces résultats offrent un aperçu très varié des comportements possibles d'une variable aléatoire.





La loi log-normale montre une distribution très étirée et asymétrique, typique de valeurs positives très dispersées.
La loi de Poisson, au contraire, décrit un comptage d'événements concentré autour de λ = 4, avec une décroissance progressive vers les grandes valeurs

Séance 5 – Échantillonnage et estimation

1. Échantillonner consiste à étudier une partie de la population quand étudier tout le monde serait impossible, trop long ou trop coûteux.
On distingue :
- aléatoires (tirage au sort) → les plus fiables,
- non aléatoires (quotas, choix raisonné),
- méthodes Monte-Carlo pour simuler des tirages.

Le choix dépend de la représentativité et des contraintes pratiques.

2. L'estimateur est formule qui sert à calculer un paramètre inconnu.
Estimation : valeur numérique obtenue avec les données observées.

3. Pour la fluctuation, on connaît le paramètre dans la population.
Confiance : on ne le connaît pas, donc on construit un intervalle probable.
→ Fluctuation = variabilité du hasard ; Confiance = incertitude sur un paramètre.

4. Le biais, c'est l'écart entre l'espérance de l'estimateur et la vraie valeur.
Un estimateur sans biais est idéal.

5. Quand on dispose de toute la population, on parle de statistique exhaustive.
C'est le cas courant en Big Data (logs, traces numériques), où l'enjeu n'est plus l'échantillonnage mais la qualité des données.

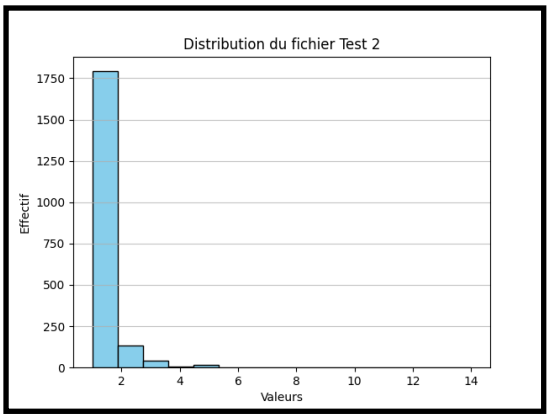
6. Les enjeux du choix d'un estimateur; il doit être :
- sans biais,
- peu variable,
- convergent,
- robuste,
- simple à utiliser.

7. Les méthodes d'estimation: moindres carrés, maximum de vraisemblance, bootstrap, Monte-Carlo.
Le choix dépend du modèle, de la taille des données et des hypothèses possibles.

8. Les tests statistiques servent à décider si une hypothèse est compatible avec les données.
Créer un test implique : définir H0/H1, choisir une statistique, sa loi, un seuil α, et une règle de décision.

9. Les tests sont souvent mal interprétés :
p-value est mal comprise, H0 est non rejetée vue à tort comme une « vraie », survalorisation de la significativité.
Le cours insiste sur un intervalle de confiance bien construit est souvent plus informatif.

Analyse des résultats :
Avec les résultats on observe, d'abord comment se comportent nos 100 échantillons. Les fréquences moyennes obtenues sont très proches de celles de la population réelle, ce qui est rassurant : cela veut dire que notre méthode d'échantillonnage fonctionne bien et que, comme prévu, les résultats se stabilisent autour des vraies proportions lorsque l'on répète suffisamment les prélèvements.
On s'est ensuite intéressé au premier échantillon. Les intervalles de confiance calculés à 95 % sont assez resserrés et contiennent en général les valeurs réelles de la population. Autrement dit, cet échantillon représente plutôt bien la réalité, et les estimations qu'il fournit sont fiables.
Enfin, j'ai testé la normalité de deux jeux de données. Le premier ressemble vraiment à une courbe en cloche, et la p-value du test de Shapiro-Wilk confirme qu'il n'y a pas de raison de rejeter l'idée qu'il suit une loi normale. Pour le second jeu, c'est l'inverse : l'histogramme est clairement asymétrique et la p-value très faible. Les données ne sont donc pas normales, et là encore, la visualisation et le test statistique racontent la même chose



Séance 6 – Statistique ordinale

1. Elle s'appuie sur des rangs et des classements. Contrairement aux données nominales, l'ordre a un sens. En géographie, elle permet de construire des hiérarchies spatiales (villes, régions…).

2. L'ordre croissant est recommandé : il suit l'évolution naturelle des valeurs (du plus faible au plus fort).

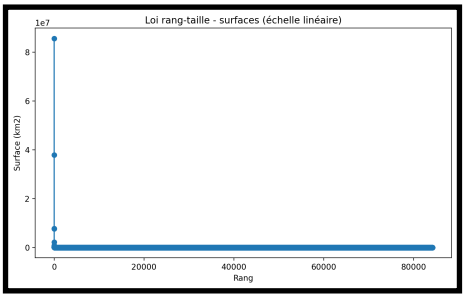
3. Corrélation (Spearman, Kendall) : mesure le lien entre deux classements.
Concordance (W de Kendall) : vérifie l'accord entre plusieurs classements.

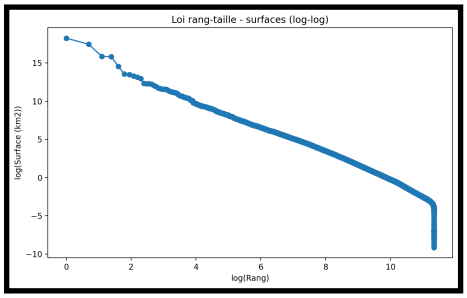
4. Spearman est basé sur les écarts entre les rangs.
Kendall est elle basé sur les paires concordantes/discordantes , elle est plus robuste.

5. Goodman–Kruskal (Γ) et Yule (Q), mesurent l'intensité d'une association entre variables ordinales :
Γ : pour tous tableaux ordonnés,
Q : version spécialisée pour tableaux 2×2.

Analyse des graphiques :

Les graphiques rang–taille montrent tout de suite que la répartition des surfaces est extrêmement déséquilibrée : quelques très grandes entités pèsent presque tout, tandis que le reste se répartit entre une multitude de petites îles. Une fois le graphique passé en échelle log–log, la courbe devient presque droite, ce qui est typique des lois de puissance et confirme cette structure très hiérarchisée.
Quand on compare ensuite le classement des pays par population et par densité en 2007, on constate qu'il n'existe pratiquement aucun lien entre les deux. Les statistiques de Spearman et Kendall vont d'ailleurs dans ce sens : être un pays très peuplé ne signifie absolument pas être un pays très dense, et inversement.
Enfin, l'analyse de la matrice de Kendall entre 2007 et 2025 montre que le classement mondial des populations évolue très lentement. Les positions relatives des pays changent peu au fil du temps, ce qui traduit une grande stabilité démographique globale et l'absence de bouleversements majeurs dans la hiérarchie mondiale





Les deux graphiques montrent que la répartition des surfaces des pays est très déséquilibrée : une poignée de pays très vastes concentre presque tout, tandis que la majorité sont beaucoup plus petits. Lorsque les données sont représentées en échelle log–log, la courbe devient quasiment droite, ce qui indique une loi de puissance. Ce comportement est typique des phénomènes géographiques et confirme une hiérarchie forte et stable dans la taille des territoires