PERRIN Louise septembre-décembre 2025

Perrin2029@gmail.com

Cours d'analyse de données
Rapport d'activité (séances 2 à 6)

Organisation du rendu : les réponses aux questions sont d'abord présentées, puis la partie de Python, d'abord les données chiffrées, puis les résultats graphiques.

Séance 2
Questions de cours

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

La géographie vient utiliser les statistiques pour analyser les phénomènes dans l'espace géographique. Cela permet de modéliser des processus spatiaux mais aussi vérifier des hypothèses ou quantifier des faits observés sur un territoire donné.

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

On peut dire que le hasard existe en géographie oui, mais il n'est jamais seule causalité. Certains phénomènes ont une part d'aléatoire mais on peut également trouver des causes extérieures précises.

3. Quels sont les types d'information géographique ?

Il existe 5 grands types de catégories d'information géographique :

1. Les données vectorielles (points, lignes mais aussi les polygones)
2. Les données raster (images satellites ou modèles numériques de terrain ou de surface)
3. Les données attributaires : liées à la population, aux températures etc...
4. Les données spatiales qualitatives : les toponymes, typologies...
5. Les données spatio-temporelles : comme l'évolution dans le temps sur les thèmes liés au climat ou à l'immigration.

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

Pour l'analyse de données, la géographie nécessite d'analyser les structures spatiales, modéliser les relations entre variable, intégrer la dimension de temps et d'espaces, visualiser et cartographier ou encore mesurer et décrire les phénomènes. Mais aussi produire des indicateurs (comme la densité, centralité ou la distance) et traiter de grands volumes (big data, GPS ou réseaux sociaux)

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

La statistique descriptive vient résumer sans aucun modèle théorique nécessaire, et répond à un questionnement lié à l'observation (lié aux moyennes, médianes, et histogrammes...) Tandis que la

statistique explicative vient établir des relations et s'appuie sur un modèle théorique pour répondre à la question : "pourquoi cela se produit ?" et "qu'est-ce qui influence quoi ici ?" (lié aux régressions, modèles et causalités)

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

Il existe différents types de visualisation de données et géographie : les cartes en points, cartes raster ou images, les cartes pour les données de taux ou proportions, les anamorphoses, les cours, histogrammes et boites à moustaches, ou encore les diagrammes de flux ou les symboles proportionnels. On les choisit selon la nature des données (qualitatives, quantitatives, continues, discrètes…), la géométrie (point, ligne, surface) et l'objectif (mettre en avant une répartition, un gradient, une relation…)

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

- Statistique descriptive
- Analyse exploratoire spatiale (Moran, LISA, autocorrélation)
- ACP / analyses multivariées (ACP, CAH, AFC)
- Classification / clustering (k-means, DBSCAN)
- Régression (linéaire, logistique, spatiale)
- Interpolation (IDW, krigeage)
- Modèles spatio-temporels
- Analyse de réseaux
- Méthodes bayésiennes

8. **Comment définiriez-vous :**

- (a) population statistique ? : Ensemble complet des unités observées. ex. toutes les communes d'Île-de-France.
- (b) individu statistique ? : Une unité parmi la population. ex. **une** commune.
- (c) caractères statistiques ? : La variable étudiée. ex. densité, revenu, altitude.
- (d) modalités statistiques ? : Les valeurs possibles du caractère. ex. pour "type de commune" : rurale, urbaine.

Quels sont les types de caractères ?

1. Quantitatifs
    a. *continus* (revenu, altitude)
    b. *discrets* (nombre d'habitants)
2. Qualitatifs
    a. *nominal* (couleur de sol, type de climat)
    b. *ordinal* (petit / moyen / grand ; faible / moyen / fort)
3. Binaires
    a. oui/non, vrai/faux

Existe-t-il une hiérarchie entre eux ? : **continu > discret > ordinal > nominal** (car plus il y a de structure, plus on peut faire d'analyses).

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

Amplitude : c'est la différence entre la valeur max et min. : amplitude interquartile = Q1 - Q3

Densité : Utile lorsque les classes n'ont pas la même largeur : on calcule la densité par une division de la quantité totale de l'échantillon (par exemple la population) par la surface totale d'espace.

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

Elles servent à **déterminer le nombre optimal de classes** d'un histogramme.

Elles évitent d'avoir trop de classes ou trop peu.

11. **Comment définir un effectif ?:** c'est est le nombre d'individus correspondant : à une population entière ou à une modalité, ou à une classe de valeurs.

12. **Comment calculer une fréquence et une fréquence cumulée?**

La fréquence relative est la fréquence absolue divisée par l'effectif total de la population ou de l'échantillon. Elle est souvent exprimée en pourcentage, et on la calcule ainsi :

$$\text{Fréquence Relative } (f_i) = \frac{\text{Fréquence Absolue } (n_i)}{\text{Effectif Total } (N)}$$

La fréquence cumulée elle est une somme des fréquences jusqu'a une modalité donnée, on la calcule par cette formule :

$$\text{Fréquence Absolue Cumulée } (N_i) = \sum_{j=1}^{i} n_j$$

13. **Qu'est-ce qu'une distribution statistique ?**

C'est la manière dont les valeurs d'un caractère se répartissent dans la population.Elle peut être : symétrique, asymétrique, multimodale, concentrée ou dispersée. Une distribution est souvent représentée par un histogramme, une courbe de densité ou encore une boîte à moustaches…

Statistique descriptive	**Statistique explicative**
Décrire, résumer	Expliquer, établir des relations
Moyennes, médianes, cartes, histogrammes	Régressions, modèles, causalité
Aucun modèle théorique nécessaire	S'appuie sur un modèle théorique
Répond : *"Qu'est-ce que j'observe ?"*	Répond : *"Pourquoi cela se produit-il ?"*, *"Qu'est-ce qui influence quoi ?"*

Python

Moyenne des échantillons (arrondies) :

- Moyenne 'Pour' : 390.5200 -> 391
- Moyenne 'Contre' : 416.0600 -> 416
- Moyenne 'Sans opinion' : 193.4200 -> 193

Fréquences (Comparaison)

Catégorie	Population mère (théorique)	Échantillons (observés)
Pour	0.39	0.39
Contre	0.42	0.42
Sans opinion	0.19	0.19

Intervalles de Confiance (IC) à 95% (n=1000)

- IC 95% 'Pour' (f=0.3950) : [0.3647 ; 0.4253]
- IC 95% 'Contre' (f=0.3960) : [0.3657 ; 0.4263]
- IC 95% 'Sans opinion' (f=0.2090) : [0.1838 ; 0.2342]

Test de Shapiro-Wilks sur Loi-normale-Test-1

- Statistique W : 0.9639
- P-value : 0.0000
- Seuil (Alpha) : 0.05

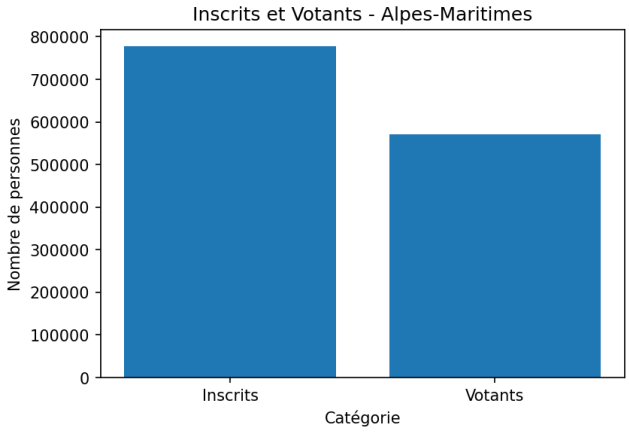
CONCLUSION : La distribution est considérée comme PAS NORMALE (Rejet de H0).

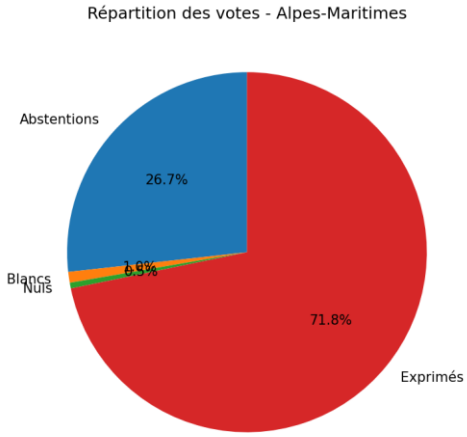
Test de Shapiro-Wilks sur Loi-normale-Test-2

- Statistique W : 0.2609
- P-value : 0.0000
- Seuil (Alpha) : 0.05

CONCLUSION : La distribution est considérée comme PAS NORMALE (Rejet de H0).

CONCLUSION : La distribution est considérée comme PAS NORMALE (Rejet de H0).





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 car il ne requiert aucune notion de nombre : il décrit seulement une catégorie, une qualité, un état, une appartenance (etc...). Alors qu'un caractère quantitatif est un cas particulier de caractère, et ses modalités peuvent être mesurées numériquement (hauteur, revenu, température…). Tout caractère quantitatif est un caractère, mais tout caractère n'est donc pas quantitatif. Le caractère qualitatif est donc plus général, car il englobe plus de situations (catégories, classes, attributs).

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

Les deux sont des caractères quantitatifs, donc exprimés par des nombres, mais ils ne prennent pas les mêmes types de valeurs : les valeurs du caractère quantitatif discret sont dénombrables en général des entiers. Le caractère quantitatif continu peut prendre toutes les valeurs réelles dans un intervalle : il faut donc faire des mesures entre les valeurs, et il existe toujours une infinité d'autres valeurs dans un intervalle.

Il faut les distinguer car ils n'utilisent pas les mêmes méthodes statistiques, les graphiques ne sont pas les mêmes, les traitements informatiques diffèrent et les phénomènes modélisés ne sont pas du même type.

3. **Paramètres de position**

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

Car aucune moyenne unique ne convient à tous les cas : chaque type de moyenne est adapté à un contexte mathématique et statistique particulier. (une seule moyenne ne peut pas représenter tous les types de données)

— Pourquoi calculer une médiane ?

Car elle donne une information claire. Elle est utile pour les distributions asymétriques et existe pour toutes les variables ordinales et quantitatives, et elle est insensibles aux valeurs extrêmes

— Quand est-il possible de calculer un mode ?

C'est la valeur la plus fréquente dans une distribution, et on peut le calculer quand les données sont qualitatives, ordinale ou quantitatives discrètes.

4. **Paramètres de concentration**

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

La médiale permet de connaître le point où se concentre la moitié de la distribution (utile en démographie, géographie sociale etc...). C'est un indicateur qui permet de résumer à quel niveau se concentre la masse totale d'une variable, de mesurer où se situe le "centre de gravité" des données, et permet d'être robuste face aux valeurs extrêmes.

L'indice de C. Gini permet de comparer des inégalités, ou la concentration d'une distribution. (utile en économie, aménagement du territoire)

5. **Paramètres de dispersion**

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

Car on ne peut pas mesurer la dispersion globale avec cet indicateur : la somme des écarts est toujours égale à 0. Mais la variance résout ce problème : elle est la première vraie mesure mathématique de dispersion : elle empêche les valeurs positives et négatives de s'annuler, donne une mesure rigoureuse de la dispersion totale et met en évidence les grands écarts, car on les amplifie (carré).

On la remplace par l'écart-type qui est plus lisible et interprétable (c'est la racine carrée de la variance, donc il revient dans les mêmes unités que la variable d'origine.)

— Pourquoi calculer l'étendue ?

L'étendue mesure l'amplitude totale de la distribution : c'est simple à calculer, utile pour repérer la disparité globale, et permet de détecter des valeurs extrêmes.

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

Les quantiles permettent de diviser une distribution en parts égales : ils fonctionnent comme des repères pour comprendre comment les valeurs se répartissent dans la population. Les intérêts sont d'étudier la répartition plutôt que la moyennes, repérer les valeurs extrêmes, comparer des groupes et faire des boîtes à moustache. Les quantiles les plus utilisés sont les quartiles, les déciles, les centiles et les percentiles.

— Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

La boîte de dispersion résume visuellement une distribution grâce aux quartiles, la médiane, l'étendue interquartile, les valeurs extrêmes et les points aberrants. L'intérêt est de visualiser la dispersion, l'asymétrie, les valeurs extrêmes, mais aussi de comparer rapidement plusieurs groupes.

On l'interprète en fonction de la hauteur de la boite (qui montre la dispersion centrale), la médiane qui montre el centre, les moustaches qui montrent l'étendue raisonnable et les points isolés qui détectent les anomalies.

6. Paramètres de forme

— Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ?
Un moment centré est calculé par rapport à la moyenne. L'intérêt est de mesurer la forme mathématique d'une distribution (symétrie, concentration autour de la moyenne).
Un moment absolu est calculé avec la valeur absolue des écarts : les mesures sont plus robustes, utiles pour les distributions avec outliers et plus faciles à interpréter pour les non spécialistes.

— Pourquoi vérifier la symétrie d'une distribution et comment faire ?
Vérifier la symétrie est utile car cela conditionne le choix des indicateurs (moyenne vs médiane) et influence les modèles statistiques car certaines lois supposent la symétrie. Cela informe aussi sur la structure de la population et permet de repérer les déséquilibres dans une distribution.
On peut vérifier la symétrie d'une distribution en utilisant le coefficient d'asymétrie ou en comparant la moyenne et la médiane (très utilisée en géographie).

Python

Statistiques :
Moyennes : [455587.6261682243, 119852.04672897197, 335735.5794392523, 5080.457943925234, 2309.822429906542, 328345.2990654206, 1842.0, 7499.271028037383, 91430.44859813084, 10293.336448598131, 76017.08411214953, 23226.41121495327, 72079.6261682243, 5761.476635514019, 15213.579439252337, 15691.598130841121, 2513.121495327103, 6777.345794392523]

Médianes : [366859.0, 95369.0, 274372.0, 4001.0, 2039.0, 268568.0, 1627.0, 5968.0, 67831.0, 8944.0, 64543.0, 16885.0, 51556.0, 4881.0, 9561.0, 11918.0, 2118.0, 6152.0]

Modes : [5045.0, 2272.0, 2773.0, 4577.0, 17.0, 2701.0, 1203.0, 19.0, 534.0, 17010.0, 459.0, 9657.0, 501.0, 75.0, 72.0, 51.0, 3663.0, 7271.0]

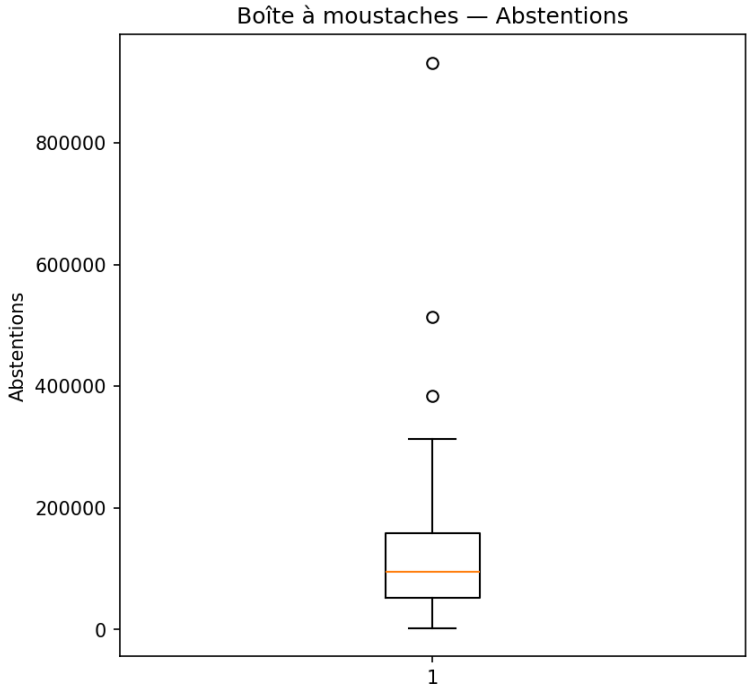
Écarts-types : [351003.7839272608, 117017.79777076835, 258393.8062439694, 3492.5237409016668, 1501.382870814494, 253758.57811511328, 1268.3719545521099, 6501.29125718549, 77226.14228932584, 7464.322239125802, 60278.09642925337, 20760.601914399664, 66210.6824497581, 4581.7903836526575, 14807.61922823019, 13027.130015862478, 1781.4104371191352, 4636.019283220522]

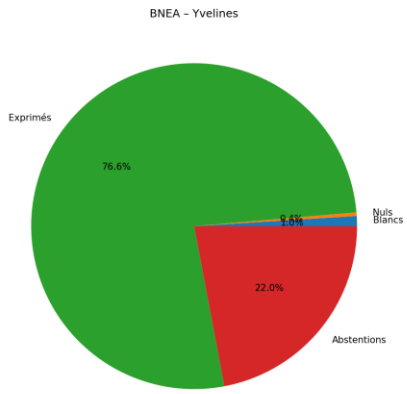
Écarts absolus à la moyenne : [272240.7163944449, 74959.06961306665, 201517.1712813346, 2817.9512621189624, 1131.9931871779197, 197762.1951262119, 977.3644859813085, 4474.96480041925, 59929.14315660756, 5140.371386147263, 42514.7158703817, 15278.362127696742, 49157.0131889248, 3333.335662503275, 11136.568259236614, 9432.008909075028, 1404.5004803913005, 3689.4970739802607]

Étendues : [1808861.0, 929183.0, 1297100.0, 17389.0, 8236.0, 1272080.0, 7651.0, 45883.0, 372286.0, 48168.0, 372668.0, 108537.0, 316871.0, 22826.0, 80196.0, 69513.0, 8686.0, 20535.0]

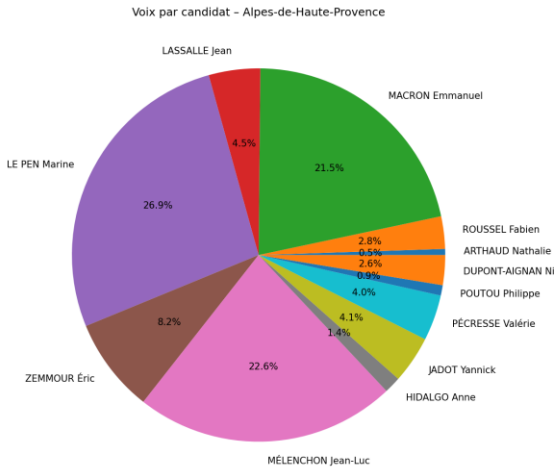
<u>Résultats graphiques :</u>

Boîte à moustahce des abstentions :



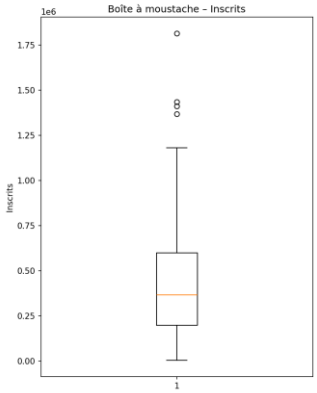


ci-dessus, exemple d'un camembert des votes dans les Yvelines.



Ci-dessus, exemple pris des votes dans les Alpes-de-Haute-Provence.

Boîte à moustache des inscrits :



Catégories des îles :

- 0-10 km² : 78423 îles
- 10-25 km² : 2327 îles
- 25-50 km² : 1164 îles
- 50-100 km² : 788 îles
- 100-2500 km² : 1346 îles
- 2500-5000 km² : 60 îles
- 5000-10000 km² : 40 îles
- 10000+ km² : 71 îles

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 ?**

Il faut compter pour les variables discrètes, et mesure pour les variables continues :

- Les variables discrètes ont des valeurs dénombrables, souvent des entiers.
- Les variables continues ont des valeurs qui peuvent prendre n'importe quelles valeurs réelles dans une intervalle.

Il faut aussi regarder la nature de phénomène étudié : s'il est aléatoire par occurrence, c'est souvent discret, et s'il est aléatoire par intensité, c'est souvent continu.

On choisit aussi en fonction de la loi qui modélise le mieux le phénomène : les phénomènes rares (Poisson, discrète), les répétitions d'essais (binomiale, discrète), la somme de variables indépendants (normale, continue), et la durée entre évènement (exponentielle, continue).

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

En géographie, on modélise souvent des phénomènes humains et physiques. Les lois statistiques utilisées varient selon le type de données, mais certaines reviennent **très souvent, en voici 5 :**

1. **La loi de poisson** : Utilisée lorsque l'on compte des événements rares dans un espace continu.
2. **La loi de Pareto – Log Normale** : Très utilisées en géographie économique.
3. **Loi exponentielle** : Décrit les temps d'attente ou les distances entre événements.
4. **Loi normale (Gaussienne)** : Très utilisée en géographie humaine et physique, car beaucoup de phénomènes se distribuent naturellement autour d'une moyenne et les variables influencées par des facteurs multiples indépendants se normalisent (théorème central limite).
5. **Loi uniforme** : Utile pour des phénomènes répartis de façon homogène dans l'espace.

Python

Distributions Discrètes :

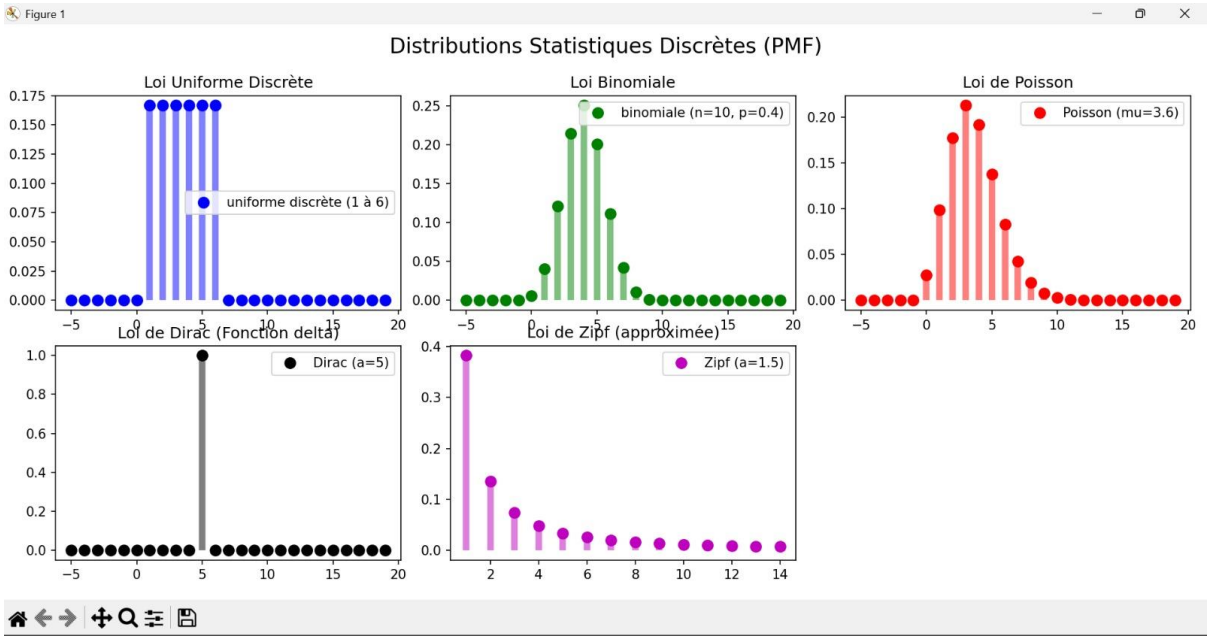
- Uniforme Discrète
  - Moyenne (μ) : 3.5000
  - Écart type (σ): 1.7078
- Binomiale
  - Moyenne (μ) : 4.0000
  - Écart type (σ): 1.5492
- Poisson (Discrète)
  - Moyenne (μ) : 3.6000
  - Écart type (σ): 1.8974
- Dirac (a=5)
  - Moyenne (μ) : 5.0000
  - Écart type (σ): 0.0000
- Zipf (approximée)
  - Moyenne (μ) : inf
  - Écart type (σ): inf

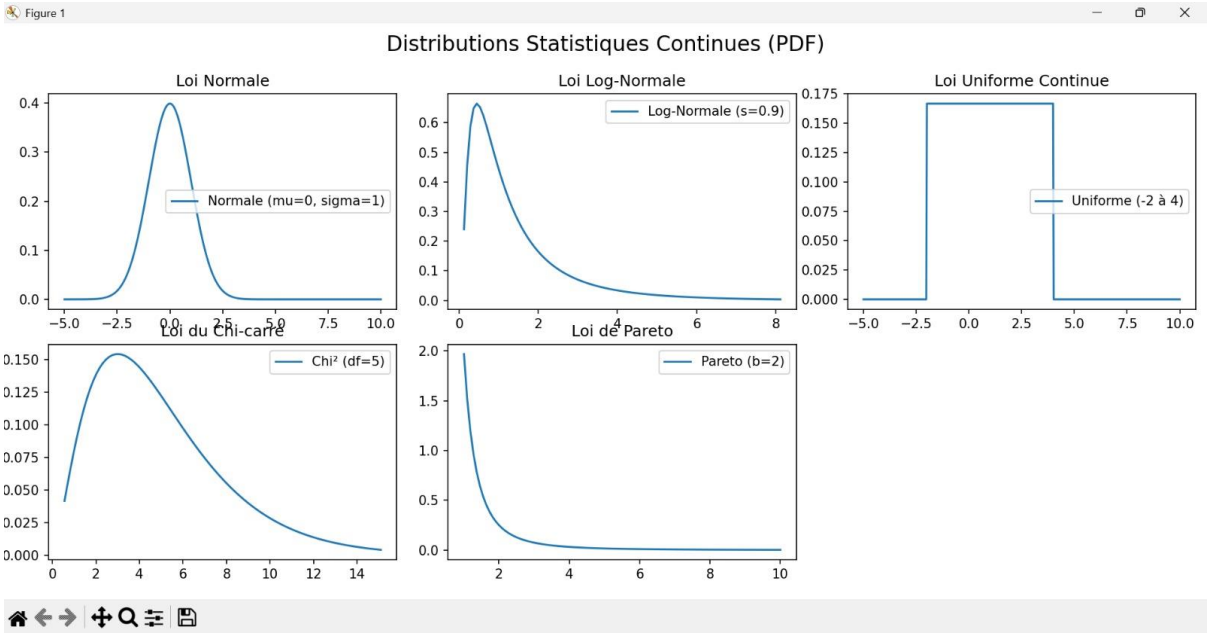
Distributions Continues :

- Normale:
  - Moyenne (μ) : 0.0000
  - Écart type (σ): 1.0000
- Log-Normale:
  - Moyenne (μ) : 1.4993
  - Écart type (σ): 1.6749
- Uniforme Continue
  - Moyenne (μ) : 1.0000
  - Écart type (σ): 1.7321
- Chi-carré
  - Moyenne (μ) : 5.0000
  - Écart type (σ): 3.1623

- Pareto
    - Moyenne (μ) : 2.0000
    - Écart type (σ): inf

Résultats graphiques :





Séance 5

1. **Comment définir l'échantillonnage ? Pourquoi ne pas utiliser la population en entier ? Quelles sont les méthodes d'échantillonnage ? Comment les choisir ?**

L'échantillonnage est la méthode qui consiste à sélectionner une partie représentative d'une population. Elle sert à étudier cette population sans devoir tous les interroger ou les mesurer.

On n'utilise pas la population en entier car le processus serait trop long, trop coûteux et parfois impossible. Un échantillon permet de produire des résultats plus fiables et plus rapidement.

Il existe deux familles de méthodes d'échantillonnage importantes :

- L'échantillonnage aléatoire (probabiliste) :
    - aléatoire simple : on tire l'échantillon au hasard dans la population
    - Stratifié : la population est divisée en sous-groupes homogènes selon des catégories, puis on tire au hasard dans chaque sous-groupes.
    - Systématique
    - en grappes : on tire au sort de groupes d'individus sur le terrain
- L'échantillonnage non probabiliste :
    - Échantillon de convenance : on interroge les individus facilement accessibles
    - Échantillon par quotas : on sélectionne un échantillon qui ressemble à la population, sans hasard.
    - Échantillonnage raisonné : sélection des individus jugés représentatifs.

On fait le choix en fonction des besoins de l'étude : en géographie on utilise surtout les échantillonnages par quotas, en grappes et stratifiés. On choisit la méthode probabiliste si on souhaite faire le calcul d'un intervalle de confiance et si une rigueur scientifique est nécessaire, tandis qu'on peut choisir la méthode non probabiliste s'il y a des contraintes de temps et de coûts, ou bien que la représentation stricte est impossible.

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

Un estimateur est une formule ou un calcul appliqué à un échantillon pour approcher une valeur inconnue de la population, c'est une fonction mathématique. (par exemple une moyenne ou une variance de l'échantillon).

L'estimation est la valeur numérique obtenue en appliquant l'estimateur sur les données : c'est une méthode et un nombre.

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

L'intervalle de fluctuation est utilisé pour comparer un échantillon à une valeur théorique par les statisticiens. L'intervalle de confiance donne un intervalle dans lequel se trouve la vraie valeur inconnue de la population, c'est un outil fondamental de l'inférence statistique.

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

Un biais est une différence entre la moyenne des estimations et la vraie valeur de la population, c'est un décalage systématique qui n'est pas dû au hasard.

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

C'est une statistique exhaustive. Avec les données massives (big data) on collecte presque toutes les données en s'approchant d'une collecte quasi exhaustive.

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

Il faut faire le choix d'un estimateur en effectuant un compromis entre exactitude, simplicité et robustesse : il faut qu'il soit convergent, stable et sans biais, mais aussi facile à calculer.

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

Il en existe plusieurs

- L'estimation ponctuelle : on donne une valeur unique pour le paramètre
- L'estimation par intervalle de confiance : donne une intervalle qui contient probablement la vraie valeur.
- L'estimation par maximum de vraisemblance : on cherche la valeur du paramètre qui rend l'échantillon le plus probable
- L'estimation bayésienne : combine les données observées avec un connaissance préalable.

On les sélectionne en fonction des situations :

- Si l'on a un petit échantillon on peut utiliser l'estimation bayésienne
- Si l'on veut quantifier l'incertitude on utilisera l'estimation intervalle de confiance
- Si l'on n'a pas d'a priori on utilisera l'estimation par maximum de vraisemblance
- Si l'on fait une étude simple on utilisera une estimation ponctuelle

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

Les tests statistiques sont des réponses à une question sur le hasard des résultats obtenus, ils permettent de montrer si le modèle choisit est le bon ou non et si une différence est significative.

Quelques tests statistiques existants : le test de moyenne, le test de proportion, d'ajustement, d'indépendance ou de corrélation...

Pour créer un test il faut formuler des hypothèses, choisir un niveau de test adapté et un niveau de risque avant de calculer la statistique de test, la population évaluée, et de décider si les résultats sont satisfaisants pour les garder ou non.

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

La statistique inférentielle peut s'avérer utile mais est imparfaite : un échantillon mal choisi peut mener à des conclusions fausses et les hypothèses peuvent être irréalistes. L'interprétation des résultats peut donc conduire à des généralisations dangereuses car faussées.

Python

Moyennes des Échantillons (Arrondies) === Moyenne 'Pour' : 390.5200 -> 391 Moyenne 'Contre' : 416.0600 -> 416 Moyenne 'Sans opinion' : 193.4200 -> 193

Fréquences (Comparaison)

Catégorie	Pop. Mère (Théorique)	Échantillons (Observé)
Pour	0.39	0.39
Contre	0.42	0.42
Sans opinion	0.19	0.19

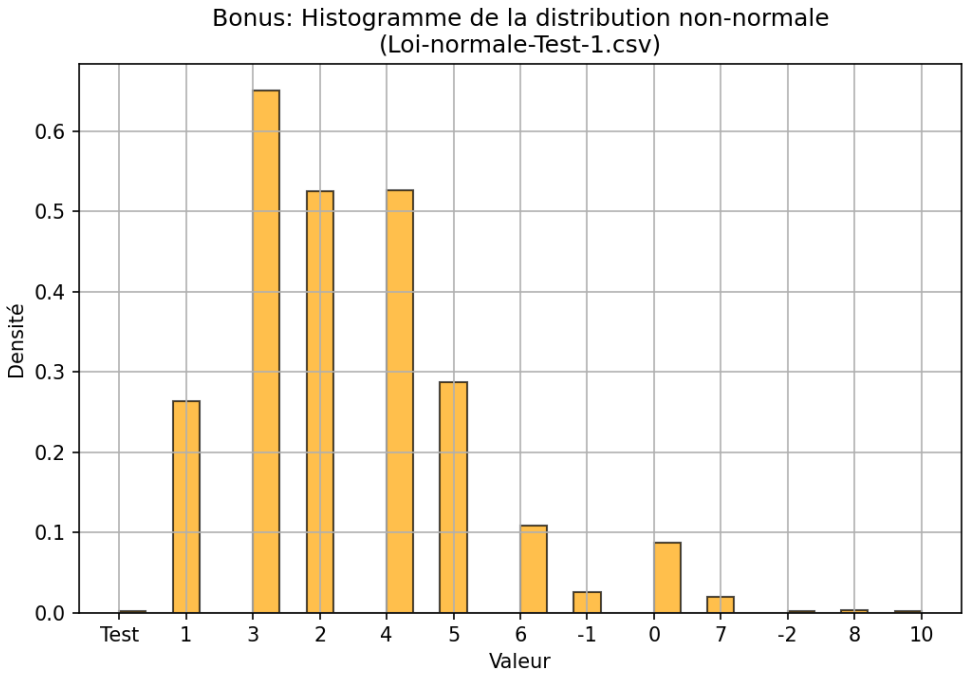
Intervalles de Confiance (IC) à 95% (n=1000) : IC 95% 'Pour' (f=0.3950) : [0.3647 ; 0.4253] IC 95% 'Contre' (f=0.3960) : [0.3657 ; 0.4263] IC 95% 'Sans opinion' (f=0.2090) : [0.1838 ; 0.2342]

Test de Shapiro-Wilks sur Loi-normale-Test-1 --- Statistique W : 0.9639 P-value : 0.0000 Seuil (Alpha) : 0.05 CONCLUSION : La distribution est considérée comme PAS NORMALE (Rejet de H0).

Test de Shapiro-Wilks sur Loi-normale-Test-2 --- Statistique W : 0.2609 P-value : 0.0000 Seuil (Alpha) : 0.05 CONCLUSION : La distribution est considérée comme PAS NORMALE (Rejet de H0).

BONUS : Identification

Observez l'histogramme. Si la distribution est asymétrique à droite et ressemble aux graphiques de la Séance 4 (comme la loi de Pareto, Log-Normale ou Chi-carré), c'est la distribution probable.



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 ?**

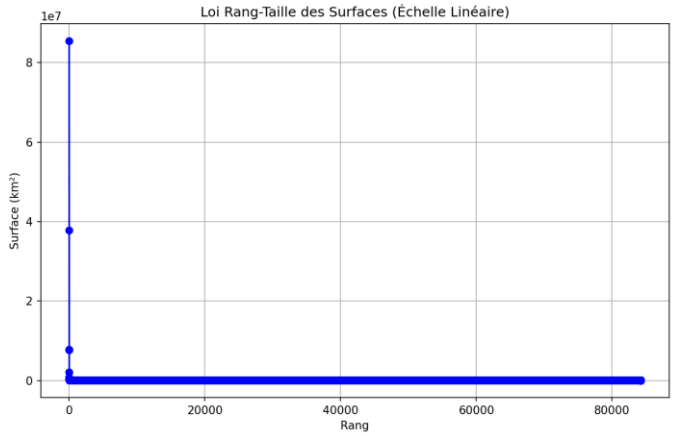
2. **Quel ordre est à privilégier dans les classifications ?**

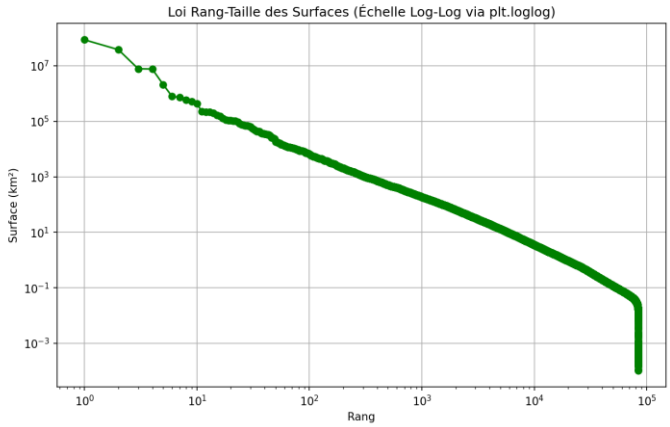
3. **Quelle est la différence entre une corrélation des rangs et une concordance de classements ?**

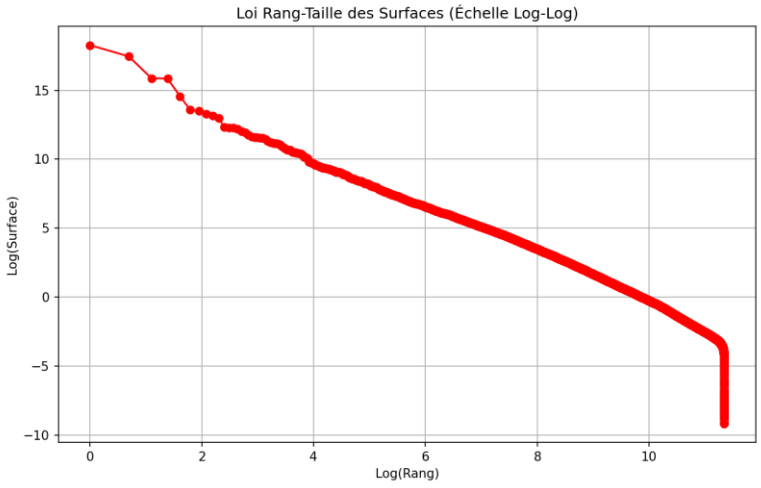
4. **Quelle est la différence entre les tests de Spearman et de Kendal ?**

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

Python







Interprétation des résultats :

La corrélation est très faible voir inexistante entre les deux variables. Le calcul révèle une relation faible et non significative entre le classement des états par population et leur classement par densité. C'est

plutôt attendu car la population est largement indépendante de la densité, qui elle est déterminée par la surface géographique des états.

QUESTION 7 REPONSE

Non, il n'est pas possible de faire un test statistique paramétrique (comme le test de Student ou Shapiro-Wilks) directement sur les rangs.

Les rangs ne sont pas des données continues issues d'une distribution de probabilité mais des données ordinales artificiellement créées.

cependant, une fois transformée en échelle loglog, la relation linéaire (loi de Zipf)peut être testée par régression linéaire (test de corrélation, p-value du coefficient de régression).

Partie sur les Populations des états du monde

QUESTION 10 : isoler les colonnes des états

Partie Populations Mondiales

Extraction réussie de 195 états et des données de population/densité.

listes prêtes pour les fonctions d'ordre et de classement

QUESTION 11 : Classements par population et densité

Classement Pop 2007 :

[[1, 'Chine'], [2, 'Inde'], [3, 'États-Unis'], [4, 'Indonésie'], [5, 'Brésil'], [6, 'Pakistan'], [7, 'Bangladesh'], [8, 'Russie'], [9, 'Nigeria'], [10, 'Japon']]

Classement Pop 2025 :

[[1, 'Inde'], [2, 'Chine'], [3, 'États-Unis'], [4, 'Indonésie'], [5, 'Pakistan'], [6, 'Nigeria'], [7, 'Brésil'], [8, 'Bangladesh'], [9, 'Russie'], [10, 'Éthiopie']]

Classement Densité 2007 :

[[1, 'Singapour'], [2, 'Malte'], [3, 'Bangladesh'], [4, 'Maldives'], [5, 'Bahreïn'], [6, 'La Barbade'], [7, 'Maurice'], [8, 'Taïwan'], [9, 'Corée du Sud'], [10, 'Pays-Bas']]

Classement Densité 2025 :

[[1, 'Monaco'], [2, 'Singapour'], [3, 'Bahreïn'], [4, 'Malte'], [5, 'Maldives'], [6, 'Bangladesh'], [7, 'La Barbade'], [8, 'Taïwan'], [9, 'Maurice'], [10, 'Liban']]

QUESTION 12 : utilisation de classementPays et Tri, Comparaison Population vs. Densité (2007)

| Rang Pop 2007 | Rang Densité 2007 | État |

| :---: | :---: | :--- |

| 1 | 39 | Chine |

| 2 | 14 | Inde |

| 3 | 123 | États-Unis |

| 4 | 48 | Indonésie |

| 5 | 134 | Brésil |

| 6 | 28 | Pakistan |

| 7 | 3 | Bangladesh |

| 8 | 159 | Russie |

| 9 | 38 | Nigeria |

| 10 | 16 | Japon |

QUESTION 14 : Coefficients de Corrélation des Rangs

Corrélation entre le classement par Population 2007 et Densité 2007 :

1. Spearman (rho) : 0.0928 (p-value: 0.2245)

2. Kendall (tau) : 0.0668 (p-value: 0.1919)

Analyse pour le Rapport : La corrélation n'est PAS statistiquement significative ($p > 0.05$). Le coefficient de Spearman (0.0928) est proche de 0, indiquant une relation monotone (tendance) très FAIBLE entre le classement par Population et par Densité.

Conclusion : La taille d'un État (population) n'est pas un bon prédicteur de sa densité de population, et inversement.