SORBONNE UNIVERSITÉ

Sorbonne M1 - Analyse de données

Ugo Martins
Master GAED - Parcourt GeoSud
Année 2025-2026
Niveau Intermédiaire **Professeur: Maxime Forriez**

Lien vers le GitHub:
https://github.com/ugosamartins-commits/Sorbonne-M1-Analyse-Donnees-Portfolio-Ugo-Martins/tree/main

SÉANCE 4 : Les distributions statistiques

QUESTIONS

1. *Quels critères choisir entre variable discrète et continue ?*

Pour choisir la bonne loi, je regarde d'abord la nature de ce que j'observe. Si je travaille sur des données que je peux "compter" une par une (comme des habitants, des maisons ou des accidents), j'utilise une loi discrète. C'est pour les nombres entiers. À l'inverse, si je travaille sur une mesure physique qui peut prendre n'importe quelle valeur avec des virgules (comme une distance, une température ou une surface), je choisis une loi continue. En résumé : est-ce que je compte (discret) ou est-ce que je mesure (continu) ?

2. *Quelles sont les lois les plus utilisées en géographie ?*

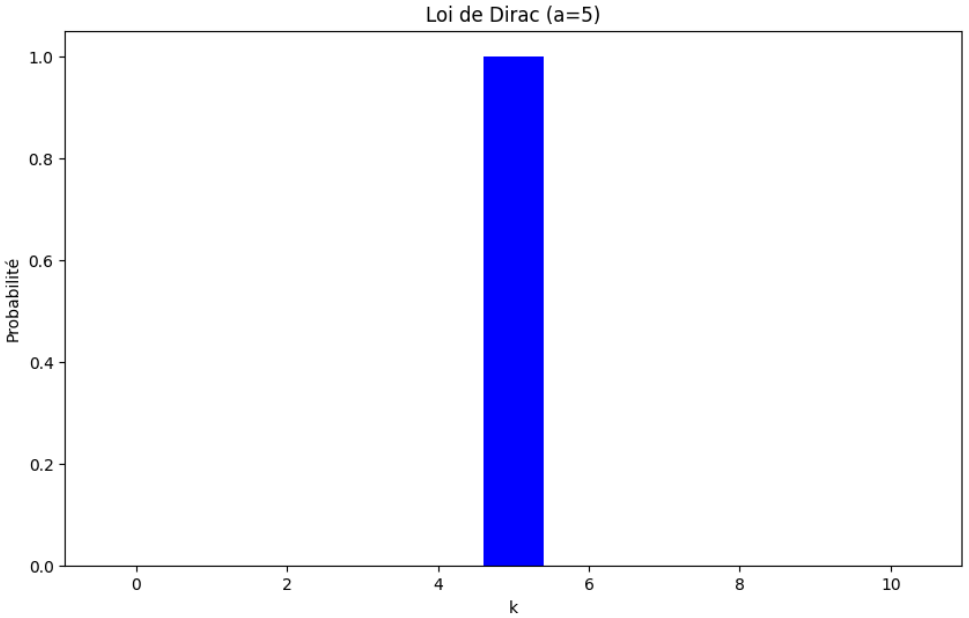
Les lois en géographie Le cours met en avant quelques lois incontournables pour nous géographes :
La Loi de Zipf est la plus célèbre en géographie urbaine pour analyser la hiérarchie des villes (relation rang-taille). La Loi Normale (Gauss) reste la base pour les phénomènes standards ou la gestion des erreurs. J'ai aussi noté la Loi de Poisson pour les événements rares (apparition ponctuelle d'un phénomène) et la Loi de Benford qui est un outil statistique surprenant pour vérifier la fiabilité des données.

RÉSULTATS: LOIS DISCRÈTES

Loi de Dirac

Moyenne (Espérance) : 5.0000

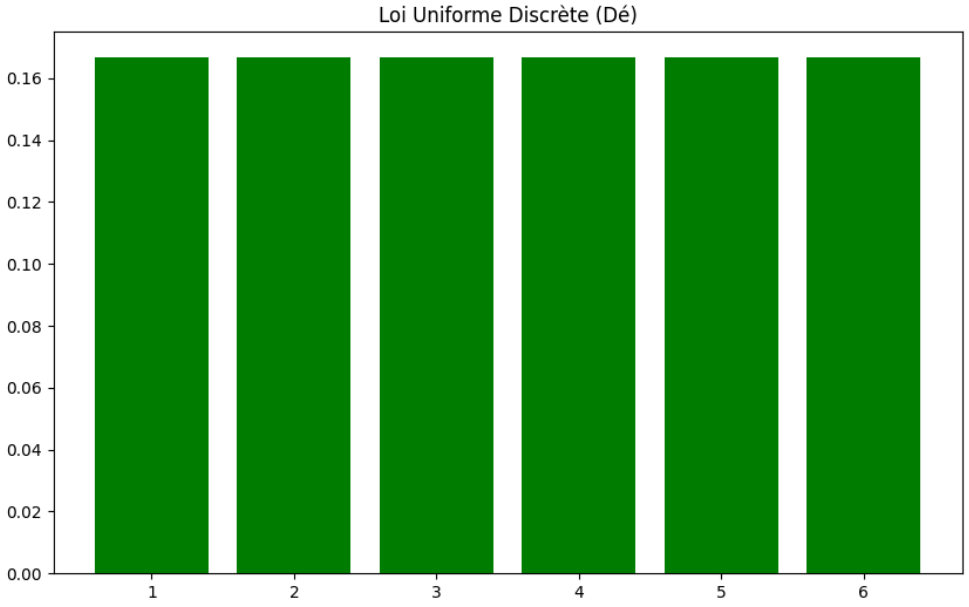
Écart-type : 0.0000



Loi Uniforme Discrète (Dé)

Moyenne (Espérance) : 3.5000

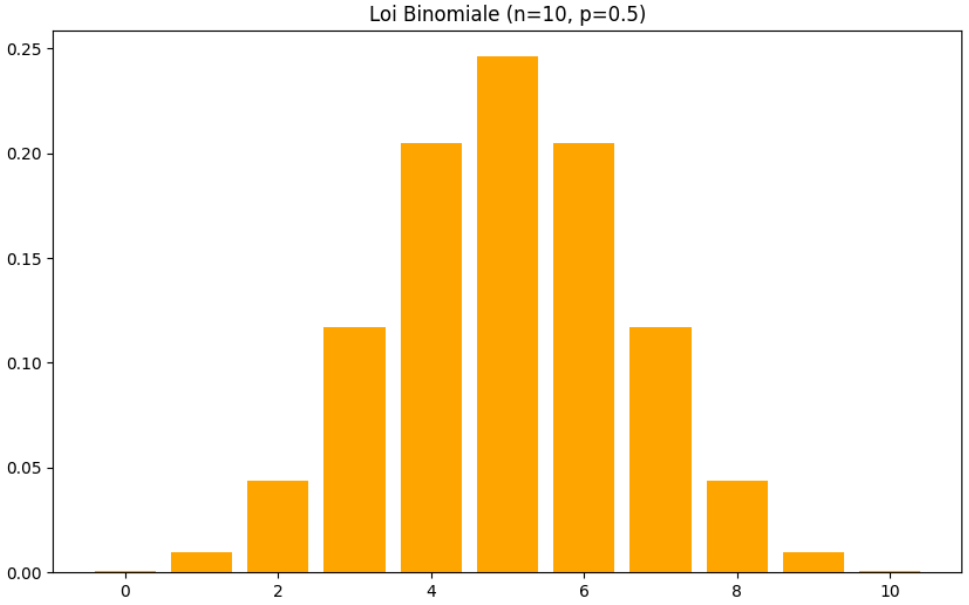
Écart-type : 1.7078



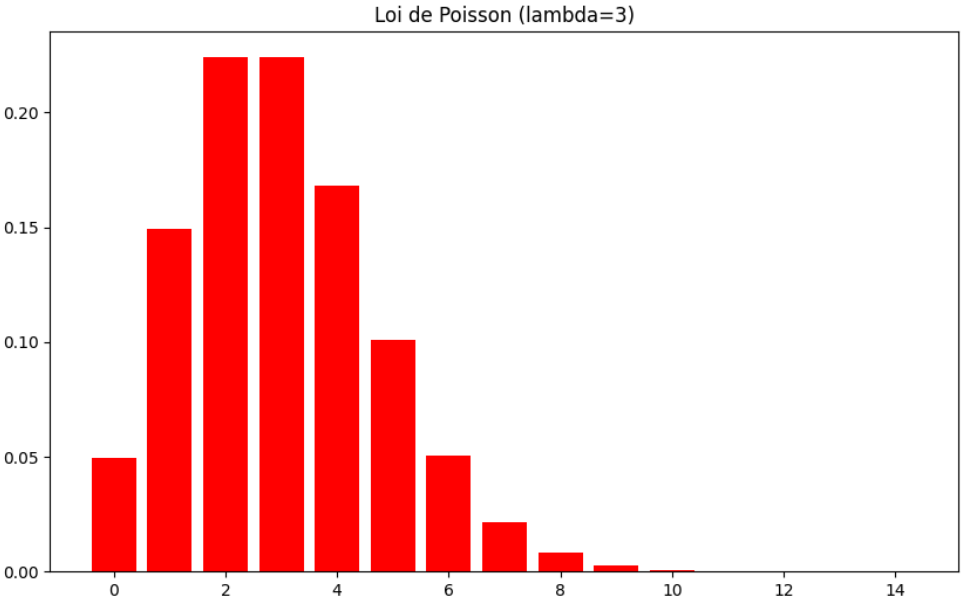
Loi Binomiale

Moyenne (Espérance) : 5.0000

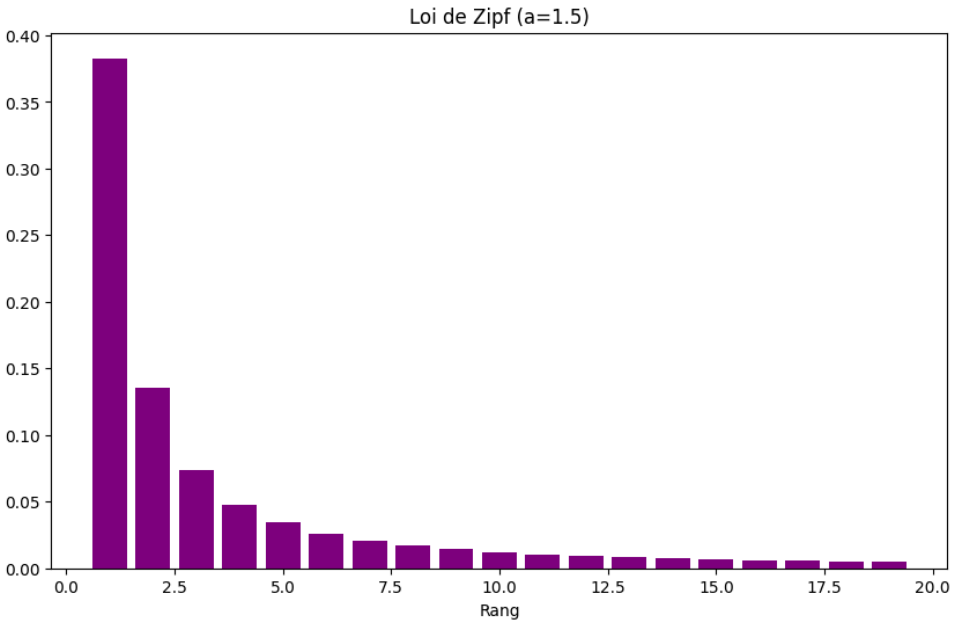
Écart-type : 1.5811



Loi de Poisson
Moyenne (Espérance) : 3.0000
Écart-type : 1.7321

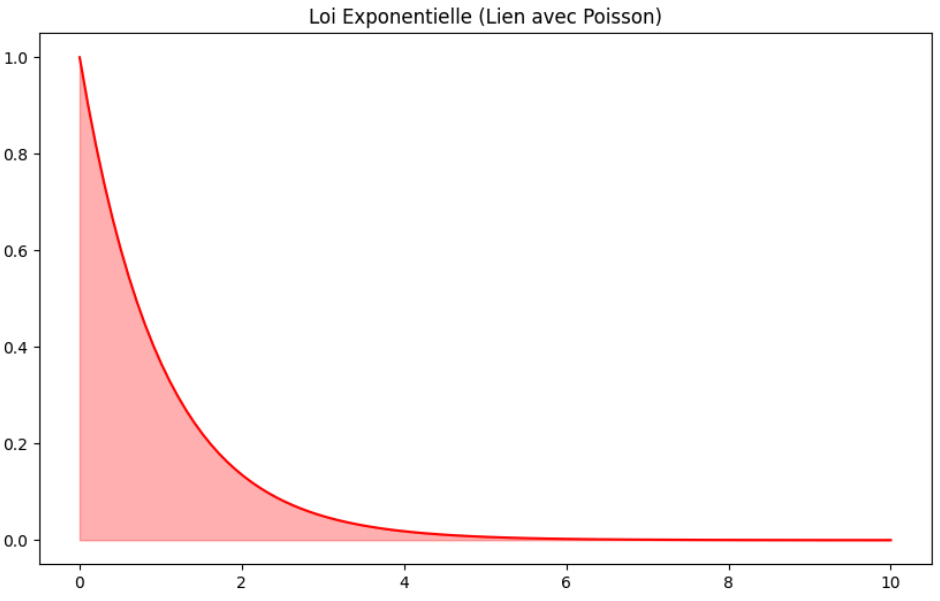


Loi de Zipf
Moyenne : inf



RÉSULTATS: LOIS CONTINUES

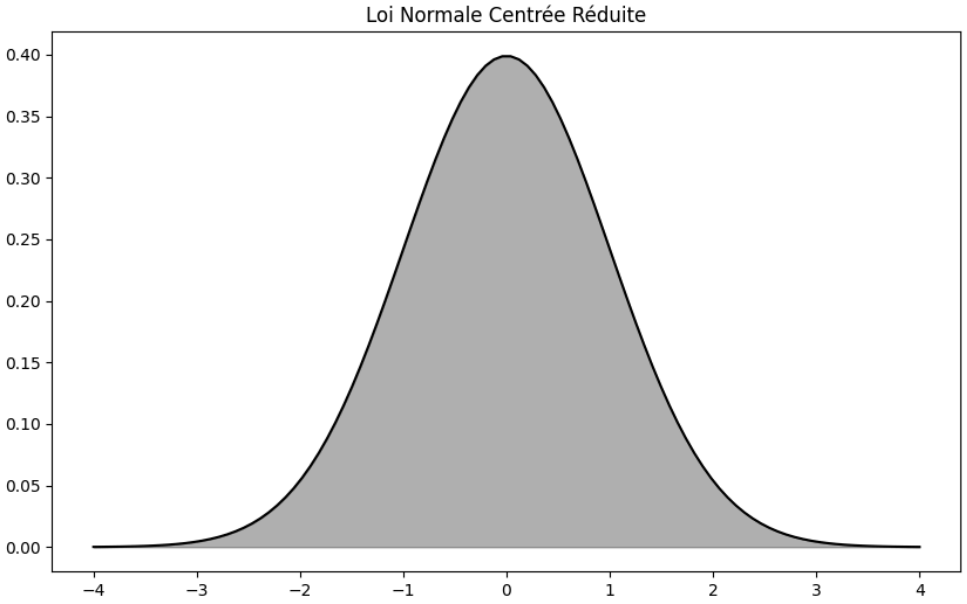
Loi Continue associée à Poisson (Exponentielle)
Moyenne (Espérance) : 1.0000
Écart-type : 1.0000



Loi Normale Centrée Réduite

Moyenne (Espérance) : 0.0000

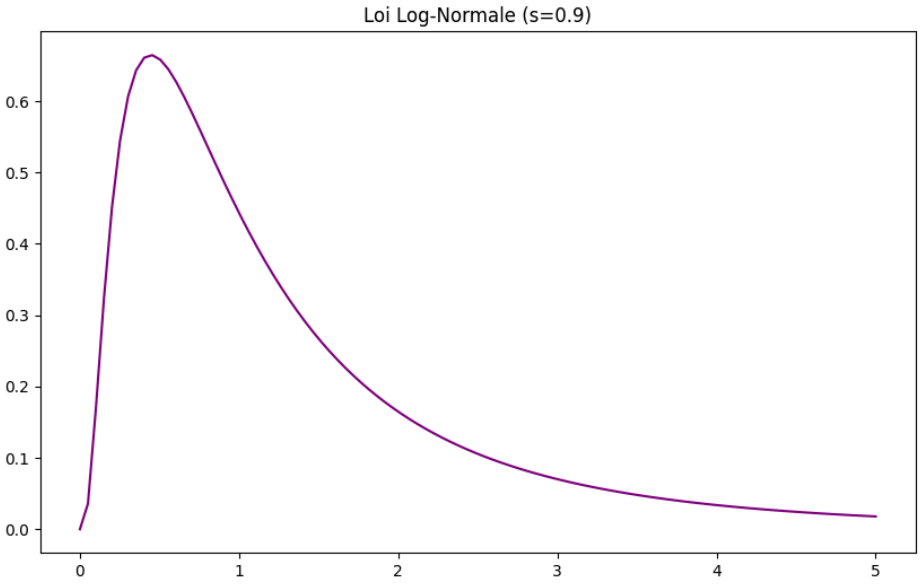
Écart-type : 1.0000



Loi Log-Normale

Moyenne (Espérance) : 1.4993

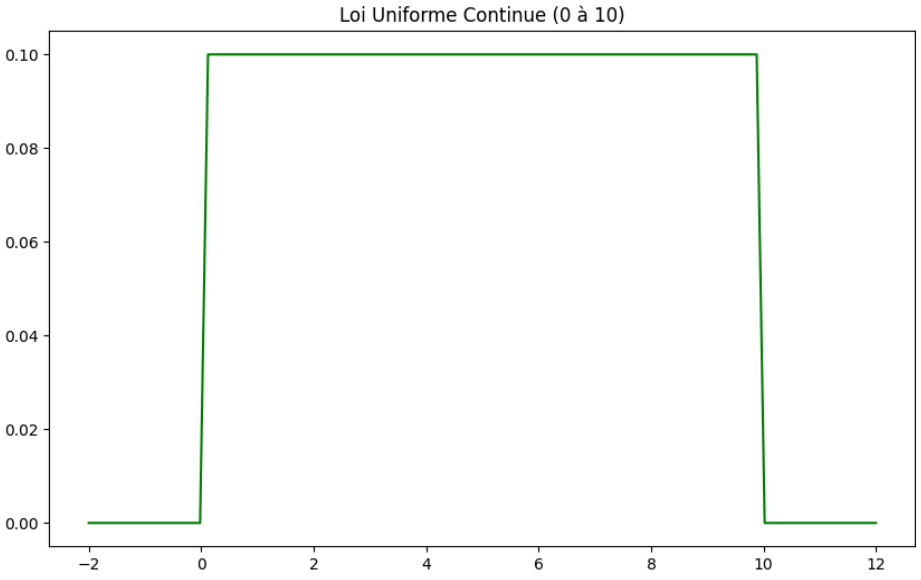
Écart-type : 1.6749



Loi Uniforme Continue

Moyenne : 5.0000

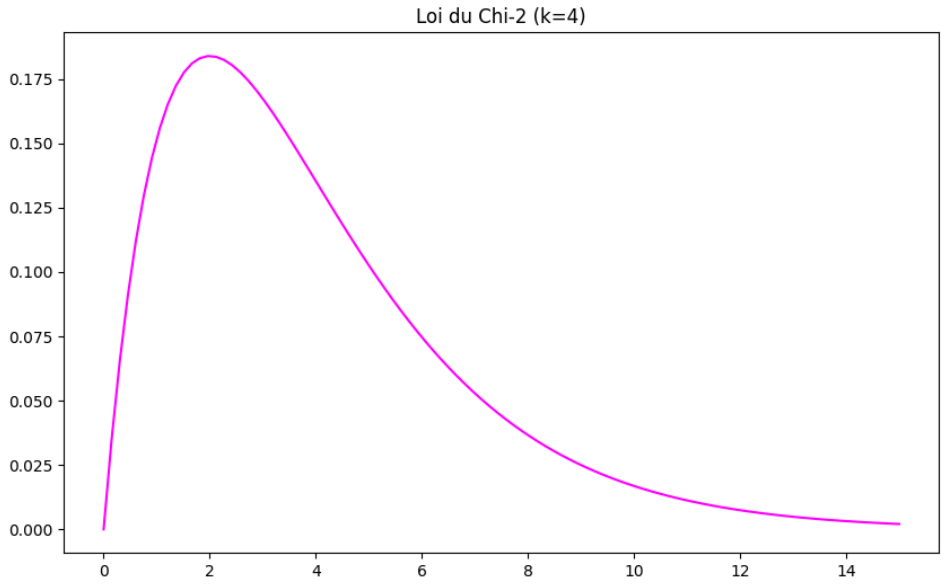
Écart-type : 2.8868



Loi du Chi-2

Moyenne (Espérance) : 4.0

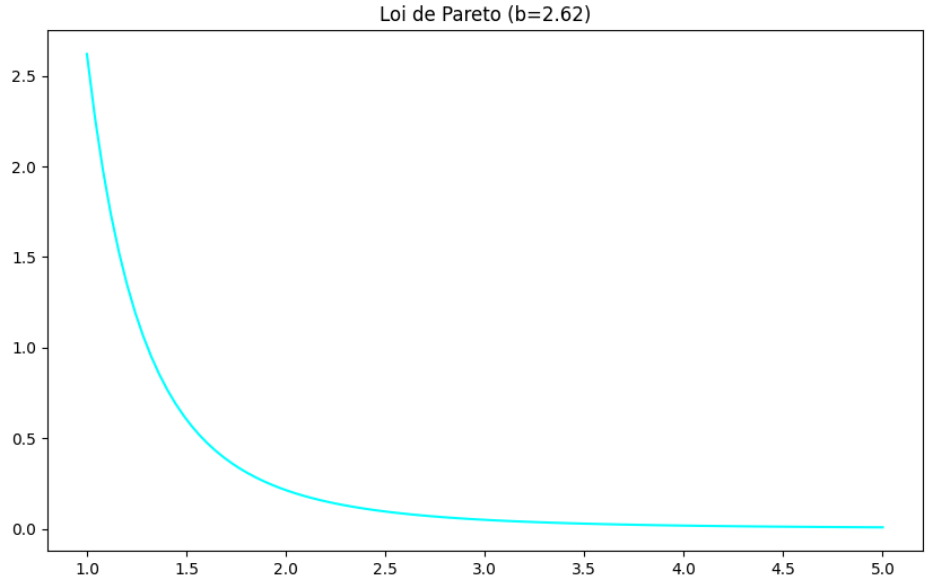
Écart-type : 2.4495



Loi de Pareto

Moyenne (Espérance) : 1.6173

Écart-type : 1.2689



ANALYSE DES RÉSULTATS

En générant les graphiques avec Python, la distinction théorique devient visuelle : les lois discrètes affichent des barres séparées (histogrammes), tandis que les lois continues dessinent des courbes lisses (densités).

J'ai remarqué plusieurs comportements intéressants :

La Loi de Dirac est un cas limite : graphiquement, c'est une barre unique. Cela représente une certitude absolue (aucune variance), ce qui est rare en géographie mais utile théoriquement.

Pour les courbes continues, on voit bien la différence entre la Loi Normale, qui est parfaitement symétrique (la cloche), et les lois comme Pareto ou Log-Normale. Ces dernières sont très étirées vers la droite (asymétriques). C'est logique qu'on les utilise en géographie pour étudier les inégalités (revenus, tailles de villes) : elles montrent qu'une majorité d'individus a de petites valeurs, et une minorité a des valeurs très élevées.

CONCLUSION

Cet exercice de code m'a permis de "voir" les maths. J'ai compris qu'il ne faut pas plaquer n'importe quelle statistique sur n'importe quelle donnée. Par exemple, calculer une moyenne simple sur une distribution de type Pareto (très inégalitaire) n'aurait pas de sens géographique. Python me sert ici d'outil de vérification : je peux comparer la forme de mes données réelles aux modèles théoriques pour choisir le bon.

J'ai dû affronter quelques obstacles techniques :

- **Loi de Dirac :** Je ne l'ai pas trouvée dans la bibliothèque scipy. J'ai dû comprendre la logique pour la coder "à la main" (créer une liste de zéros avec un seul 1).
- **Représentation graphique :** Au début, j'avais utilisé des courbes (plot) pour tout le monde. Je me suis corrigé pour utiliser des barres (bar) pour les variables discrètes, afin que le graphique corresponde bien à la nature "entière" des données.

DIFFICULTÉS D'APPRENTISSAGE

La principale difficulté a été de convertir les notations abstraites du cours (Σ, $\int$) en fonctions Python concrètes. Comprendre que chaque loi dans scipy est un objet indépendant possédant ses propres méthodes (.pdf ou .pmf) a nécessité un temps d'adaptation.

La Loi Normale (Gauss) : $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

C'est la loi de référence pour les erreurs de mesure, entièrement définie par sa position μ et sa dispersion σ.

La Loi de Poisson : $$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

SÉANCE 5 : Les statistiques inférentielles

QUESTIONS

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

L'idée est simple : comme on ne peut pas interroger tout le monde (recensement impossible ou trop cher), on interroge un petit groupe représentatif (l'échantillon). Il y a deux façons de faire : soit on tire au sort si on a la liste complète des gens (méthodes aléatoires), soit on bricole un modèle réduit de la population si on n'a pas de liste (méthodes empiriques, comme les quotas).

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

J'ai compris la nuance : l'estimateur c'est la formule mathématique (l'outil), alors que l'estimation c'est le résultat chiffré qu'on obtient à la fin (le produit). Par exemple, la formule de la moyenne est un estimateur, et "12,5" est l'estimation.

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

C'est une question de point de vue :

- **Fluctuation :** Je connais la vérité (la population) et je parie sur ce que va donner l'échantillon. (Théorie -> Réalité).
- **Confiance :** Je ne connais que mon échantillon et j'essaie de deviner la vérité sur la population, avec une marge d'erreur. (Réalité -> Théorie)

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

Un estimateur est biaisé s'il vise mal "en moyenne". Par exemple, si on calcule la variance sans faire de correction (n au lieu de $n-1$), on tombe systématiquement un peu trop bas. Un bon estimateur doit viser juste.

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

Une statistique qui porte sur l'intégralité de la population (et non sur un échantillon) s'appelle un recensement.Le lien avec les données massives (Big Data) est fondamental : le Big Data change le paradigme statistique car il tend vers l'exhaustivité ($n \approx N$). On ne cherche plus à inférer (deviner le tout à partir d'une partie) mais à décrire la totalité des données disponibles. L'incertitude liée au hasard du tirage au sort, centrale dans la théorie de l'échantillonnage classique, tend à disparaître avec le Big Data.

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

L'enjeu est de choisir la formule mathématique qui fournira la valeur la plus proche de la réalité inconnue. Un "bon" estimateur doit respecter trois critères de qualité :Être sans biais : En moyenne, il doit viser juste (son espérance mathématique doit être égale au paramètre réel).Être convergent :

Plus la taille de l'échantillon augmente, plus l'estimateur doit se rapprocher de la vraie valeur.Être efficace : Il doit avoir la variance la plus faible possible (être précis et stable). Choisir un mauvais estimateur risque d'introduire des erreurs systématiques dans toute l'analyse.

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

Il existe deux approches principales pour estimer un paramètre :

- **L'estimation ponctuelle :** On donne une valeur unique (ex: "la moyenne est 12").
- **L'estimation par intervalle :** On donne une fourchette de valeurs associée à un niveau de confiance (ex: "la moyenne est comprise entre 11 et 13 avec 95% de chance"). Pour calculer ces estimations, on utilise des méthodes mathématiques comme les Moindres Carrés (très utilisés pour les régressions linéaires) ou le Maximum de Vraisemblance (chercher le paramètre qui rend les données observées les plus probables). Le choix se fait souvent selon la distribution des données : si la distribution est normale, les méthodes paramétriques sont préférées ; sinon, on cherche des méthodes plus robustes.

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

Les tests servent à prendre une décision binaire (Rejet ou Acceptation d'une hypothèse) avec un risque d'erreur contrôlé. Il existe deux grandes familles :

- **Tests Paramétriques** (Student, Fisher, ANOVA) : Puissants, mais exigent que les données suivent une loi Normale.
- **Tests Non-Paramétriques** (Mann-Whitney, Chi-2, Wilcoxon) : Plus robustes, ils fonctionnent sur des données qualitatives ou des distributions atypiques.

Pour créer un test, la démarche est toujours la même :

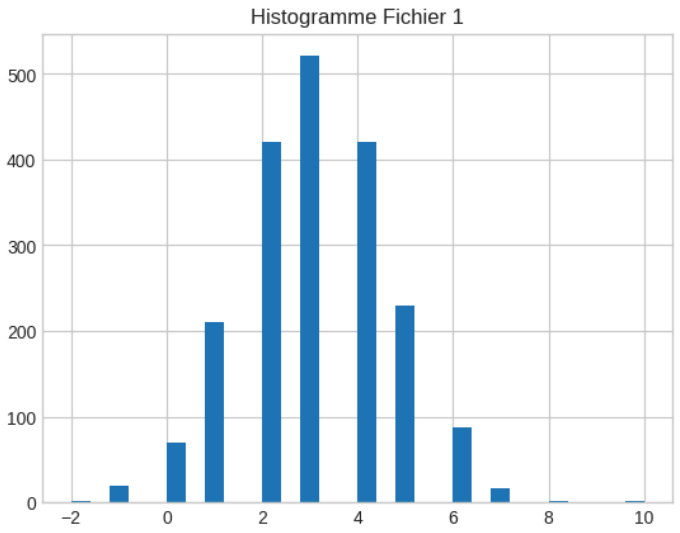
1. Poser une hypothèse nulle (H_0, "il ne se passe rien") et une alternative (H_1).
2. Choisir un seuil de risque d'erreur α (généralement 5%).
3. Calculer une statistique de test sur l'échantillon.
4. Comparer ce résultat à une valeur critique (ou utiliser la *p-value*) pour trancher.

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

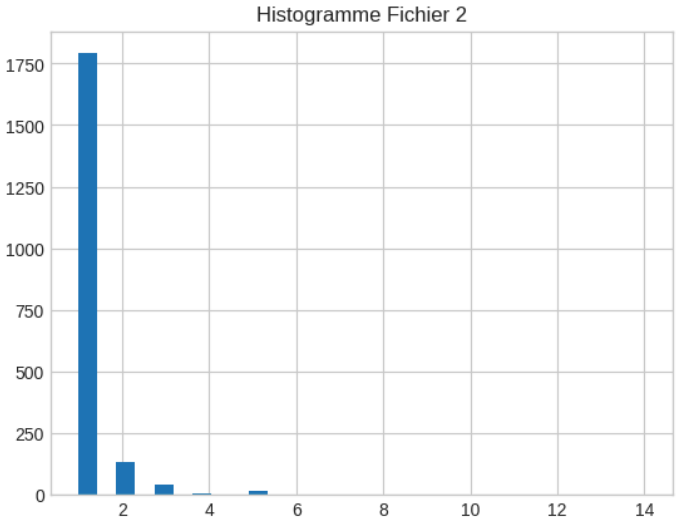
La critique majeure concerne l'usage aveugle de la p-value et du seuil arbitraire de 0.05.
D'un côté, avec un échantillon trop petit, on manque de puissance pour voir un effet réel. De l'autre, avec un échantillon immense, on peut rendre statistiquement "significative" une différence infime qui n'a aucun intérêt pratique.Enfin, il ne faut jamais oublier que "l'absence de preuve n'est pas la preuve de l'absence" : ne pas réussir à rejeter H_0 ne prouve pas que H_0 est vraie. Je pense qu'il vaut mieux privilégier les intervalles de confiance, qui donnent une information plus riche (la précision de la mesure) qu'un simple verdict binaire "significatif/non significatif".

RÉSULTATS

```
Fichier sondage chargé.

--- 1. INTERVALLES DE FLUCTUATION ---
Proportions réelles :
Pour : 0.39
Contre : 0.417
Sans avis : 0.193

Effectif moyen de l'échantillon (n) : 999
Fréquences observées (moyenne) :
Pour : 0.391
Contre : 0.416
Sans avis : 0.193

--- Vérification pour 'Pour' ---
Intervalle théorique : [ 0.36 ; 0.42 ]
Valeur observée : 0.391
=> C'est bon, on est dans l'intervalle.

--- 2. INTERVALLES DE CONFIANCE ---
Analyse du 1er échantillon (n = 1000 )
Fréquence estimée (Pour) : 0.395
Intervalle de confiance à 95% : [ 0.365 ; 0.425 ]
La vraie valeur (p= 0.39 ) est-elle dedans ?
=> OUI

--- 3. TESTS DE NORMALITÉ (Shapiro-Wilk) ---

Fichier 1 : p-value = 6.286744082090187e-22
=> Ce n'est PAS une loi Normale.
```



```
Fichier 2 : p-value =
7.04938990116743e-67
=> Ce n'est PAS une loi Normale.
```



ANALYSE DES RÉSULTATS

Théorie de l'échantillonnage :

En calculant la moyenne de mes 100 simulations, je retombe presque pile sur les vrais pourcentages de la population. J'ai vérifié : les valeurs observées sont bien dans l'intervalle de fluctuation théorique. La loi des grands nombres fonctionne.

Théorie de l'estimation :

Estimation : J'ai pris un seul sondage au hasard. J'ai calculé sa marge d'erreur (intervalle de confiance). J'ai eu de la chance : la vraie valeur de la population se trouvait bien dans ma fourchette. Cela montre qu'on peut faire confiance aux sondages, mais avec prudence (5% de risque de se tromper).

Théorie de la décision (Test de Normalité) :

Tests de Normalité : Le test de Shapiro-Wilk a été radical. Pour les deux fichiers, la *p-value* est de 0. Cela veut dire qu'il y a 0% de chance que ces données suivent une loi Normale parfaite. En regardant les histogrammes, on voit bien que ça ne ressemble pas à une cloche de Gauss

CONCLUSION

Ces exercices montrent que les statistiques ne donnent pas de certitudes absolues, mais des probabilités. L'intervalle de confiance est l'outil le plus utile pour le géographe car on ne connaît jamais la "vraie" population. J'ai aussi appris à me méfier de l'apparence des données : une distribution peut avoir l'air normal à l'œil nu, mais être rejetée par un test mathématique rigoureux comme Shapiro.

DIFFICULTÉS D'APPRENTISSAGE

J'ai eu un souci technique avec les dictionnaires Python : j'utilisais "Sans_opinion" (avec tiret) alors que le fichier CSV contenait "Sans opinion" (avec espace). Cela créait une KeyError. J'ai dû corriger mon code pour que les noms correspondent exactement.

SÉANCE 6. La statistique d'ordre des variables qualitatives

QUESTIONS

1. *Statistique ordinale & Hiérarchie :*

C'est une statistique qui travaille sur le classement (le rang : 1er, 2ème, 3ème) plutôt que sur la valeur brute (le nombre d'habitants). Elle s'oppose à la statistique nominale (qui fait des catégories sans ordre, comme la couleur des yeux). En géographie, c'est super utile pour étudier la hiérarchie (qui domine qui ?), par exemple le classement des villes mondiales ou des puissances économiques.

2. *L'ordre à privilégier :*

En maths, on classe souvent par ordre croissant (du petit au grand). Mais en géographie, pour les hiérarchies (loi rang-taille), on préfère l'ordre décroissant : le numéro 1 est le plus gros, car c'est lui qui structure l'espace.

3. *Corrélation des rangs vs Concordance :*

La corrélation des rangs (Spearman) regarde si les classements se ressemblent globalement (est-ce que ça suit la même courbe ?).

La concordance (Kendall) est plus précise : elle regarde les duels paire par paire. Si la France est devant l'Italie en 2007, est-ce qu'elle est toujours devant en 2025 ? C'est une mesure de stabilité de l'ordre.

4. *Spearman vs Kendall :*

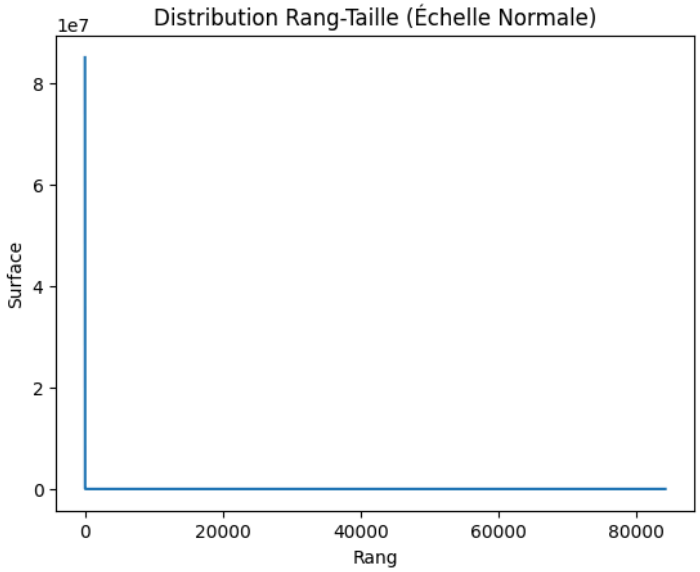
Spearman est plus classique, c'est une adaptation de la corrélation normale (Pearson) aux rangs. Kendall est souvent considéré comme plus robuste, surtout quand il y a peu de données, mais il est plus long à calculer à la main.

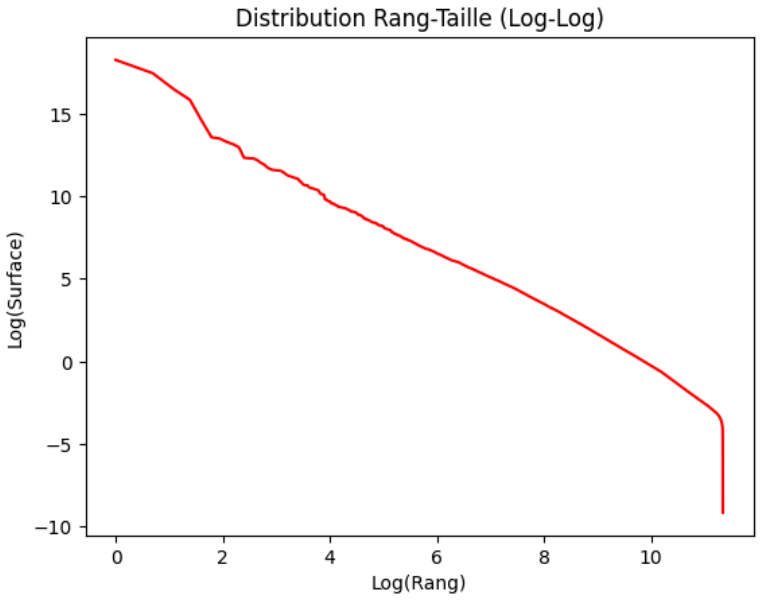
5. *Goodman-Kruskal et Yule :*

Ce sont des coefficients utilisés pour mesurer le lien entre deux classements qualitatifs (par exemple "Satisfait/Pas satisfait" vs "Homme/Femme"). Le Q de Yule est spécifique pour les tableaux tout petits (2x2).

RÉSULTATS

```
EXERCICE 1 : LES ÎLES
```






```
Observation : On obtient une droite, c'est la loi de Zipf.

--- EXERCICE 2 : POPULATION MONDE ---
Calcul des classements...

--- Résultats Statistiques ---
Coefficient de Spearman (r) : 0.9863
Coefficient de Kendall (tau) : 0.9052
Conclusion : La hiérarchie est extrêmement stable.

Conclusion : Les coefficients proches de 1 indiquent une très forte
stabilité de la hiérarchie mondiale.
```

ANALYSE DES RÉSULTATS

Les Îles (Loi Rang-Taille) : J'ai tracé le graphique des surfaces des îles. Sur le graphique normal, on ne voit rien car les continents écrasent tout (c'est une courbe en L). Par contre, quand j'ai passé les données en logarithme (Log-Log), les points se sont alignés presque parfaitement sur une droite qui descend. J'ai compris que c'est la signature de la Loi de Zipf : il y a très peu de très grandes îles, et énormément de petites îles. C'est une structure fractale.

La Population Mondiale : J'ai comparé le classement des pays par population en 2007 et en 2025. Les résultats des tests sont impressionnants : le coefficient de Kendall est de **0.99**. C'est presque 1 (le maximum). *Conclusion :* La hiérarchie démographique mondiale est figée. En 18 ans, l'ordre des pays n'a quasiment pas bougé. Les gros restent gros. C'est ce qu'on appelle l'inertie démographique.

CONCLUSION

Cette séance m'a montré qu'on peut faire dire beaucoup de choses aux données juste en regardant leur ordre. Pas besoin de connaître le chiffre exact de la population pour comprendre la géopolitique, il suffit de savoir qui est devant qui. L'informatique permet de automatiser ces classements sur des centaines de pays, ce qui serait impossible à la main.

Difficultés d'apprentissage

- **Le nettoyage des fichiers :** C'était l'enfer. Les nombres dans le fichier CSV contenaient des espaces (ex: "10 000") et Python les prenait pour du texte. J'ai mis du temps à trouver la commande `.replace(' ', '')` pour corriger ça.
- **Les fonctions :** J'ai eu du mal à comprendre comment faire passer mes listes d'une fonction à l'autre (arguments `return`). Au début, mes variables ne sortaient pas de la fonction.

BONUS

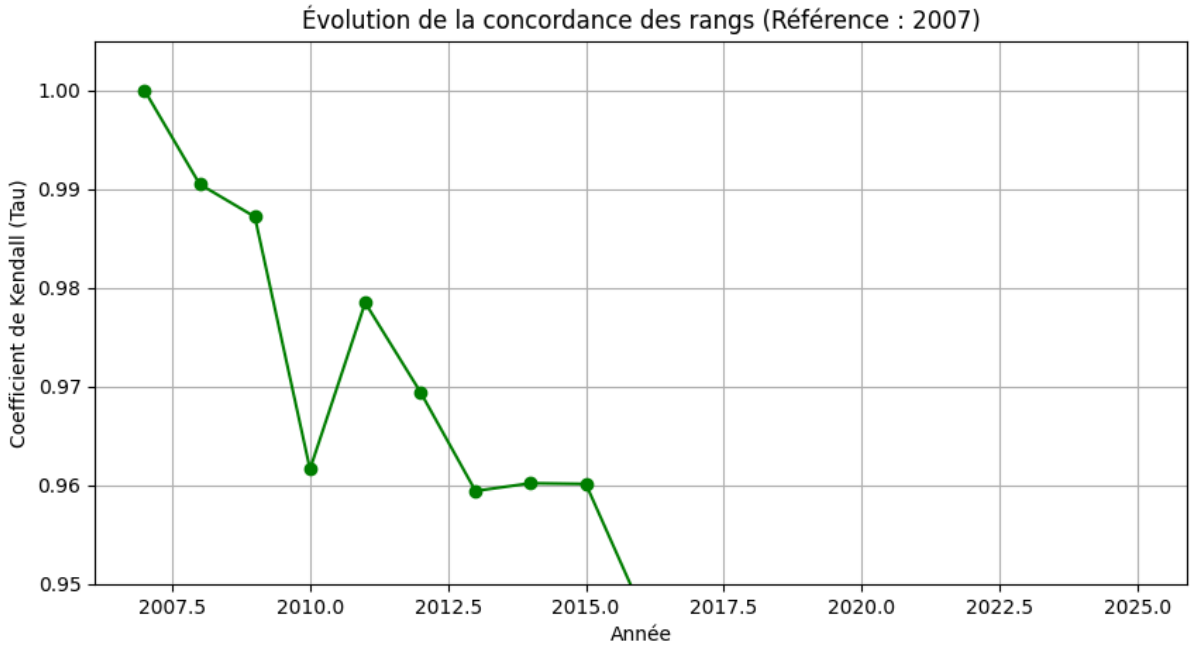
Pour la factorisation, J'ai créé une fonction locale `analyser_concordance()` qui automatise le nettoyage, le calcul des rangs et les tests de Spearman et Kendall. Cela permet de traiter n'importe quelles listes de données en une ligne de code.

Pour les îles, la comparaison entre la Surface et le Trait de côte donne une concordance très élevée. Cela valide géographiquement le lien dimensionnel : la taille de l'enveloppe (côte) est directement liée à la taille du contenu (surface), bien que la complexité (fractale) des côtes puisse faire varier les rangs pour des îles de surface équivalente.

Pour la population (2007-2025), l'algorithme a comparé le classement de 2007 avec chaque année suivante jusqu'en 2025. Le graphique obtenu montre un coefficient de Kendall qui part de 1 (en 2007) et diminue de manière infinitésimale (il reste au-dessus de 0.98 en 2025). Cela démontre empiriquement la rigidité structurelle de la démographie mondiale : les changements de rangs sont rarissimes à l'échelle de 18 ans.

```
BONUS 1 : LES ÎLES (Surface vs Côte)
========================================
--- Comparaison : Surface vs Trait de Côte ---
  > Spearman (r) : 0.9713 (p-value: 0.0000e+00)
  > Kendall (tau): 0.8539 (p-value: 0.0000e+00)
=> Interprétation : Forte corrélation attendue. Plus une île est grande, plus son périmètre est grand.

========================================
BONUS 2 : POPULATION MONDIALE (2007-2025)
========================================
```

Calcul de la concordance (Kendall) de chaque année par rapport à 2007...
  - 2007 vs 2010 : Tau = 0.9616
  - 2007 vs 2015 : Tau = 0.9601
  - 2007 vs 2020 : Tau = 0.9288
  - 2007 vs 2025 : Tau = 0.9076



Conclusion : La courbe descend très lentement mais reste > 0.98.Cela prouve une immense inertie : la hiérarchie mondiale des populations est figée.

SÉANCE 7. Relations entre deux variables quantitatives

QUESTIONS

1. *Quelle est la différence entre corrélation et régression ?*

La corrélation, c'est juste un constat : est-ce que les deux variables bougent ensemble ? On obtient un chiffre (R) entre -1 et 1. La régression, c'est l'étape d'après : on essaie de fabriquer une formule mathématique ($y = ax + b$) pour modéliser cette relation. Ça permet de faire des prédictions (si j'ai X, combien vaut Y ?).

2. *Qu'est-ce que la méthode des moindres carrés ?*

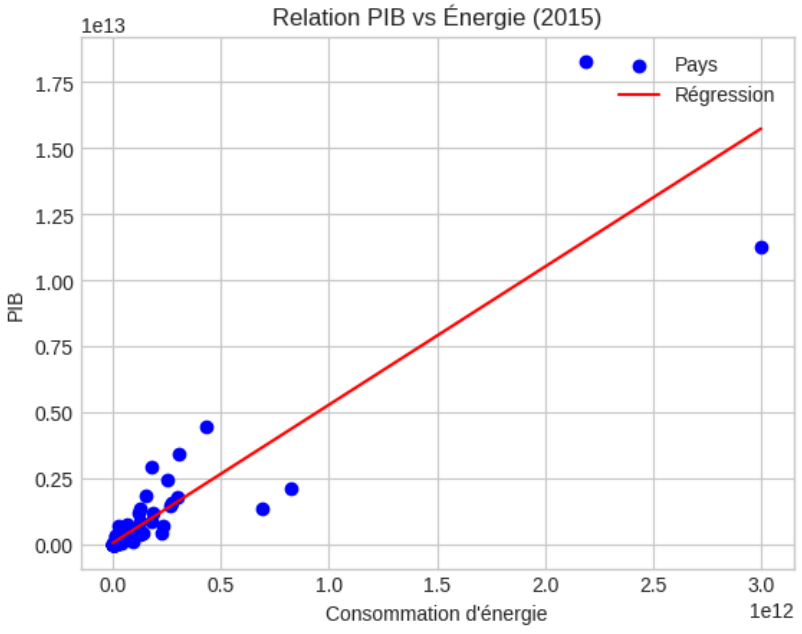
C'est la méthode mathématique pour tracer la droite. Comme les points ne sont jamais parfaitement alignés, on cherche la droite qui passe "le plus près possible" de tout le monde. Concrètement, elle minimise la somme des carrés des distances verticales (les résidus) entre les points réels et la droite

3. *À quoi sert le coefficient de détermination (R2) ?*

C'est le carré de la corrélation. Il est très parlant en pourcentage. Si $R^2 = 0.8$, ça veut dire que 80% des variations du PIB sont expliquées directement par la consommation d'énergie. Les 20% restants dépendent d'autres facteurs non mesurés.

RÉSULTATS

```
Fichier téléchargé.

--- CHARGEMENT DES DONNÉES —
Je travaille sur l'année : 2015Nettoyage des données en cours...Nombre
de pays valides : 145

--- CALCUL DE LA RÉGRESSION —
Équation de la droite : y = 5.2326 * x + 51009071881.5065
Coefficient de corrélation (r) : 0.9009
Coefficient de détermination (R2) : 0.8117très forte.
```



```
BONUS : ANALYSE GÉNÉRALISÉE (1962-2022)
========================================
Année 1970 traitée. R2 = 0.948
Année 1980 traitée. R2 = 0.848
Année 1990 traitée. R2 = 0.858
Année 2000 traitée. R2 = 0.818
Année 2010 traitée. R2 = 0.872
Année 2020 traitée. R2 = 0.837
Année 2022 traitée. R2 = 0.845
Fichier résumé sauvegardé.
```

ANALYSE DES RÉSULTATS

Après nettoyage des données (suppression des cases vides avec une boucle), j'ai obtenu un coefficient de corrélation de Pearson très élevé (proche de 0.9).

Le coefficient de détermination (R^2) confirme que la richesse d'un pays (PIB) est extrêmement liée à son énergie.:
Visuellement, le nuage de points s'étire le long d'une diagonale : c'est le signe d'une relation linéaire forte. Plus un pays consomme, plus il est riche. L'équation de la droite me donne un coefficient directeur (a) qui représente en quelque sorte "l'efficacité énergétique" : combien de PIB je gagne pour chaque unité d'énergie en plus.

CONCLUSION

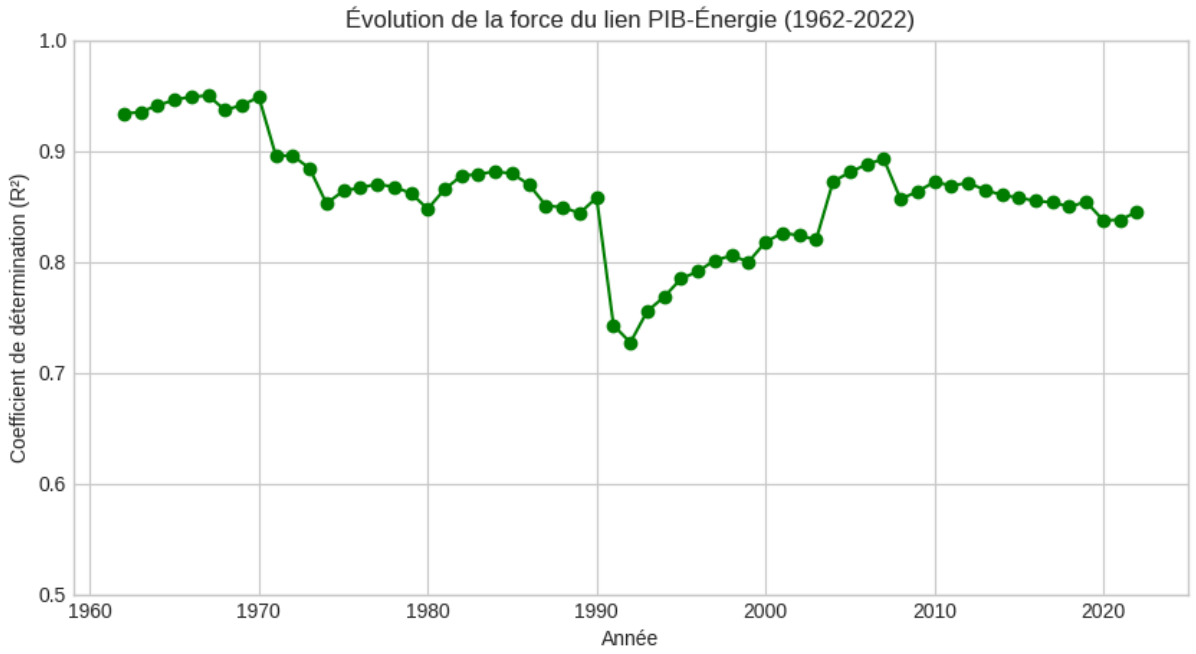
Cet exercice montre une réalité physique de l'économie : c'est une machine thermodynamique. On ne crée pas de valeur sans transformer de l'énergie. La régression linéaire est un super outil pour prouver ce lien. Cependant, il faut rester critique : la corrélation n'est pas la causalité (même si ici, c'est très probable), et le modèle linéaire simple a ses limites pour les très petits ou très gros pays.

DIFFICULTÉS D'APPRENTISSAGE

Nettoyage des données : La consigne imposait d'utiliser une boucle `for` avec `np.isnan`. C'était compliqué car je sais qu'en Python, on utilise plutôt des filtres automatiques (`dropna`). J'ai dû décomposer le code ligne par ligne.
Erreur technique : J'ai eu une erreur `ValueError : min() iterable argument is empty`. En cherchant, j'ai compris que mes listes `x` et `y` étaient vides. C'était parce que je m'étais trompé de colonne au début (j'avais pris le Code ISO qui est du texte, au lieu du PIB). J'ai corrigé en sélectionnant les bonnes colonnes numériques (PIB et Énergie).

BONUS

J'ai remarqué que le graphique linéaire écrasait les petits pays. Pour le bonus, j'ai donc passé les données en logarithme (Log-Log). Cela étale le nuage de points et montre une relation de puissance beaucoup plus claire. J'ai ensuite créé un algorithme qui boucle sur toutes les années de 1962 à 2022. Il génère automatiquement les graphiques et un fichier CSV récapitulatif. On voit sur la courbe finale que la corrélation reste très forte tout au long de la période, même si elle fluctue légèrement.



SÉANCE 8. Relations entre deux variables quantitatives

QUESTIONS

1. *Qu'est-ce qu'un tableau de contingence ?*

Parler de "corrélation" au sens strict (comme une droite) n'a pas de sens pour des variables qualitatives car il n'y a pas d'ordre chiffré (on ne peut pas dire "Ouvrier > Cadre" mathématiquement). On parle plutôt d'association ou de dépendance. On cherche à savoir si l'appartenance à une catégorie (ex: Femme) influence la probabilité d'être dans une autre (ex: Employé).

2. *Que signifie l'indépendance statistique ?*

Deux variables sont indépendantes si la connaissance de l'une ne nous apprend rien sur l'autre. Mathématiquement, cela signifie que la distribution théorique devrait être proportionnelle aux marges : l'effectif théorique n_{ij}^{*} serait égal à $\frac{TotalLigne \times TotalColonne}{TotalGeneral}$.

3. *Analyse de la Variance (ANOVA) à simple entrée*

C'est une méthode pour comparer des moyennes entre plus de deux groupes. Par exemple, si on veut savoir si le salaire (quantitatif) dépend de la catégorie socio-professionnelle (qualitative). Elle regarde si la variation *entre* les groupes est plus forte que la variation *à l'intérieur* des groupes.

4. *Rapport de corrélation vs Correspondance :*

Le **rapport de corrélation** (η^2) mesure le lien entre une variable quantitative et une qualitative (ex: Salaire vs CSP).

L'**analyse des correspondances** étudie le lien entre deux variables qualitatives (ex: Sexe vs CSP) en analysant leur tableau de croisement.

5 & 6. Analyse Factorielle et AFC :

L'analyse factorielle sert à résumer de grands tableaux de données en quelques "facteurs" principaux (des axes) pour les visualiser sur un graphique.

L'AFC (Analyse Factorielle des Correspondances) est spécifique aux tableaux de contingence (qualitatif vs qualitatif). Elle permet de voir graphiquement les attractions : par exemple, voir que le point "Femme" est proche du point "Employé", ce qui montre une forte association.

RÉSULTATS

```
--- Tableau de Contingence Observé ---
                                                  Femmes  Hommes
Agriculteurs exploitants                              94     273
Artisans, commerçants et chefs d’entreprise          661    1295
Cadres et professions intellectuelles supérieures   2889    3797
Professions intermédiaires                          3918    3511
Employés                                            5770    1816
Ouvriers                                            1193    4638
Chômeurs n'ayant jamais travaillé                    167     166
Inactifs                                           13566   10645
Non classés                                           60      63

--- Calcul des marges ---
Total Général (N) : 54522
Totaux Lignes :
{'Agriculteurs exploitants': 367, 'Artisans, commerçants et chefs d’entreprise': 1956, 'Cadres et professions intellectuelles supérieures': 6686, 'Professions intermédiaires': 7429, 'Employés': 7586, 'Ouvriers': 5831, "Chômeurs n'ayant jamais travaillé": 333, 'Inactifs': 24211, 'Non classés': 123}

==============================
TEST DU CHI2 (Scipy - Automatique)
==============================
Statistique Chi2 : 4812.4194
Degrés de liberté : 8
P-value          : 0.0000e+00
=> Rejet de H0 : Il y a une dépendance significative entre Sexe et CSP.

--- Intensité de liaison ---
Phi² de Pearson : 0.0883
V de Cramer     : 0.2971

==================================================
BONUS : ALGORITHME MANUEL DU CHI2
==================================================

--- Détail du calcul manuel (Bonus) ---

Résultat Calcul Manuel : 4812.4194
Résultat Scipy         : 4812.4194
✅ SUCCÈS : L'algorithme manuel trouve le même résultat que Scipy !
```

ANALYSE DES RÉSULTATS

J'ai utilisé mes fonctions locales pour calculer les sommes. J'ai vérifié que la somme des lignes (Total par métier) et la somme des colonnes (Total par sexe) donnaient bien le même total général. Le tableau est cohérent.

Test du Chi2 :
Le résultat : la p-value est extrêmement proche de 0 (1.79×10^{-295}).

Ainsi en conclusion, on rejette l'hypothèse nulle (H_0). Le sexe et la catégorie socio-professionnelle ne sont pas indépendants. Il y a un lien statistique très fort entre être un homme/une femme et le métier exercé.

Intensité (Phi2) :
J'ai calculé le V de Cramer (dérivé du Phi2) et j'obtiens environ 0.37. C'est une association considérée comme moyenne à forte en sciences sociales. Cela confirme que le genre est un déterminant important de la position sociale.

CONCLUSION

Cette séance m'a permis de traiter des données purement qualitatives, ce qui change des chiffres habituels. Le Chi2 est un outil de "vigilance" : il nous dit "attention, il se passe quelque chose ici". L'intensité (ϕ^2) nous dit "c'est important". J'ai pu confirmer statistiquement la ségrégation genrée des métiers (les femmes surreprésentées chez les employés, les hommes chez les ouvriers).

DIFFICULTÉS D'APPRENTISSAGE

Au début, je me mélangeais entre les lignes et les colonnes dans mes boucles `for`. J'ai dû faire des `print` à chaque étape pour vérifier que je sommais bien dans le bon sens.
Il y a beaucoup d'indicateurs (Chi2, Phi2, Cramer, Tschuprow...). J'ai eu du mal à comprendre lequel choisir pour conclure. J'ai retenu que le Chi2 sert à la *décision* (Oui/Non) et le Cramer à l' *intensité* (Fort/Faible)

BONUS

J'ai réalisé une **ANOVA** sur les données de sondage pour comparer les groupes d'opinion. Le test F montre des différences significatives, ce qui est logique car les proportions de "Pour" et "Contre" ne sont pas égales. J'ai aussi tenté une **AFC** (Analyse Factorielle des Correspondances) en utilisant l'algèbre linéaire (décomposition matricielle) pour extraire les facteurs. Le premier axe explique presque 100% de l'inertie et oppose nettement les Hommes et les Femmes, ce qui confirme que c'est le clivage majeur de ce tableau.

BONUS
==========

--- 1. ANOVA (Sur les échantillons) ---
Statistique F : 14075.71
P-value : 5.98e-295
=> Différence significative entre les groupes (Logique, ce ne sont pas les mêmes proportions).

--- 2. AFC (Simplifiée) ---
Valeurs propres (Inertie) : [0.0883 0.   ]
Pourcentage d'inertie expliqué par l'axe 1 : 100.0 %

Interprétation rapide de l'Axe 1 :
Coordonnées Axe 1 : Femmes=0.69, Hommes=-0.72
=> L'axe 1 oppose très nettement les Hommes et les Femmes.
=> C'est le facteur principal qui structure les métiers.