ROBLIN Manon

Questions analyse des données

Séance 2 :

La géographie et les statistiques entretiennent une relation relativement complexe car la géographie est avant tout une discipline de recherche. De fait il est courant qu'elle méprise les définitions mathématiques de la statistique car il ne s'agit pas de son domaine traditionnel. C'est ainsi paradoxal puisque la géographie fournie des données de manières massive, que seuls les outils statistiques sont capables d'interpréter. Nous pouvons donc dire qu'en quelques sortes, la géographie adopte une place de méfiance vis-à-vis des statistiques qui relève d'un domaine davantage mathématique que le leur. En géographie, le déterminisme nie le hasard en disant qu'il existe une cause pour tout, tandis que la posture statistique l'accepte mais le voit davantage comme une cause cachée qui sera un jour expliquée. Donc dans un phénomène aléatoire comme l'action des individus sur un territoire nous ne pouvons pas prévoir chaque détail, mais nous pouvons dégager une certitude globale ou bien une tendance de l'action qui semble la plus probable. Le hasard qui est une vision philosophique existe donc bien, bien que l'on puisse quand même trouver des lois générales et faire de la géographie une science en étudiant les échelles.
Notre cours distingue deux grandes séries statistiques qui constituent l'information géographique :

- L'étude des caractéristiques pour une entrée territoriale donnée : Ce sont les attributs dans un S.I.G, comme la population, les caractéristiques économiques (géographie humaine) ou la température, les précipitations (géographie physique), ...

- L'étude de la morphologie des ensembles délimités : Ici ce sont des données nouvelles du S.I.G comme la géographie d'un ensemble géographique.

Au niveau de l'analyse des données, la géographie a ses besoins. En effet, la géographie a besoin de l'outil statistique pour étudier les données massives qu'elle produit comme nous l'avons vu précédemment. L'analyse de données elle-même est le moment mathématique qui permet d'étudier la structure interne des données, ce qui nécessite de fait de mobiliser les probabilités et les statistiques. Donc en quelques sorte, sans statistiques, on ne peut pas donner de sens à la quantité d'informations géographiques qu'on a donc le besoin de la géographie vis-à-vis de l'analyse de données est fondamental.
Nous pouvons toutefois relever des différences entre la statistique descriptive et la statistique explicative. La statistique descriptive sert à étudier et décrire les données (dans la population ou l'échantillon donnée. Son objectif est de dégager des propriétés remarquables afin d'obtenir une image simplifiée de la réalité comme des caractéristiques numériques ou des graphiques et de préparer les comparaisons et les prédictions. L'intuition y joue d'ailleurs un rôle fondamental.
La statistique explicative quant à elle ne cherche plus juste à décrire, mais à relier une variable à expliquer (Y) à des variables explicatives ($X_1, \dots X_2$). Ainsi elle tente d'ajuster un modèle mathématique comme une régression pour expliquer les phénomènes qu'elle étudie.
En géographie il existe plusieurs types de visualisation de données comme l'histogramme, les représentations sectorielles ou encore les diagrammes en bâtons que l'on choisit en fonction du type de variable. Par exemple pour les variables quantitatives continues on

choisit l'histogramme ; pour les variables qualitatives on choisit les représentations sectorielles et enfin pour les variables quantitatives discrètes il est préférable de choisir le diagramme an bâton.
Il existe plusieurs méthodes d'analyse des données possibles :

- Méthode descriptive (analyse des données) qui sert à résumer, visualiser et classer des données multidimensionnelles (ex: ACP, AFC, ACM, CAH).
- Méthodes explicatives qui servent à relier une variable Y a des variables X (ex: Régression, Analyse de la variance).
- Méthodes de prévisions qui servent quant à elles à analyser et prévoir une série chronologique.

Nous pouvons définir les termes suivants :

- Population statistique : il s'agit de l'ensemble sur lequel l'étude porte.
- Individu statistique : il s'agit d'un élément de la population que l'on appelle unité spatiale en géographie.
- Caractères d'un individu : ici, nous parlons des caractéristiques ou des particularités étudiées (qui deviennent une variable statistique).
- Modalité du caractère : il s'agit des valeurs prises par le caractère, elles doivent être incompatibles et exhaustives.

Il existe les caractères Qualitatifs (nominaux ou ordinaux) et Quantitatifs (discrets ou continus). Il n'existe pas de hiérarchie entre eux mais l'identification du type est fondamentale pour choisir les traitements statistiques appropriés.

Pour mesurer une amplitude ou une densité :

- L'amplitude (A) : il faut prendre la valeur la plus grande de la classe (b) et lui enlever la valeur la plus petite (a). Il s'agit tout simplement de la taille ou la longueur d'une classe, elle indique ainsi l'écart de valeur que cette classe couvre.

- La densité (D) : est calculée en prenant l'effectif de la classe (le nombre d'individus qu'elle contient) et en le divisant par son amplitude (sa longueur). Elle sert à comparer des classes qui n'ont pas la même amplitude. Si une classe est très large et une autre très étroite, comparer uniquement leur effectif (le nombre de personnes dans la classe) n'a pas de sens. La densité permet de "normaliser" cette comparaison. Donc elle est une mesure de concentration de vos observations à l'intérieur d'une classe.

Les formules de Sturges et de Yule sont des outils pratiques utilisés quand vous décidez de regrouper vos données quantitatives (comme des âges ou des salaires) en classes (par exemple : [10-20], [20-30], …). Leur rôle est d'estimer le nombre idéal de classes (k) à créer. En utilisant le nombre total d'observations d'une série, ces formules donnent une valeur approximative de k. C'est une recommandation pour ne pas perdre une quantité d'informations "non négligeable".

Enfin nous pouvons aussi définir les termes suivants :

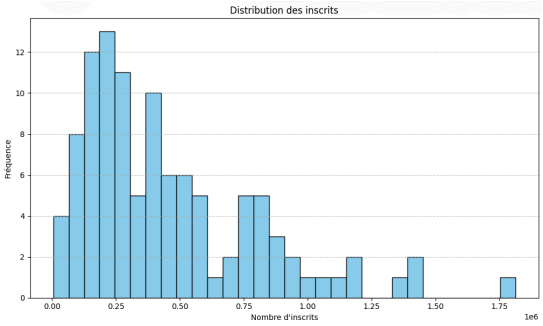
ROBLIN Manon

- Effectif : il s'agit simplement du nombre de fois qu'une valeur, ou une classe de valeurs, apparaît dans une population étudiée. C'est la fréquence absolue.

- Fréquence : C'est une proportion. Elle s'obtient en divisant l'effectif d'une valeur par l'effectif total de toute la série. Cela vous donne une valeur comprise entre 0 et 1 (une probabilité observée).

- Fréquence cumulée : Pour un caractère quantitatif ordonné, elle s'obtient en additionnant la fréquence de la valeur actuelle avec toutes les fréquences des valeurs précédentes. Elle vous dit quelle proportion de la population est inférieure ou égale à une certaine valeur.

- Distribution statistique : Il s'agit du point de départ pour l'interférence. Elle permet de déterminer et de conclure sur le type de loi de probabilité théorique (Loi Normale, ...) qui régit le phénomène étudié. Elle est établie grâce aux fréquences observées. C'est l'image de la répartition des valeurs dans votre série.



Le code de la séance 2 nous donnes l'histogramme suivant ainsi que de nombreux autres diagrammes graphiques et camemberts, relatant pour chacune les données que nous avions dans le dossier Data. Cet histogramme représente sur l'axe des ordonnés la distribution de la fréquence et sur l'axe des abscisses est représenté le nombre d'inscrits au premier tour des élections présidentielles de 2022. Chaque barre que l'on observe représente un intervalle du nombre d'inscrits, tandis que sa hauteur indique le nombre de fois où cette plage d'inscrits a été observée. Lorsque l'on regarde l'histogramme nous constatons que la distribution est positive et fortement asymétrique. Le pic principal des fréquences est situé vers la gauche tandis que sur la droite la distribution est davantage asymétrique. En raison de l'asymétrie à droite, on peut déduire que la Moyenne du nombre d'inscrits par département est probablement supérieure à la Médiane, car elle est tirée vers la droite par les quelques départements ayant un très grand nombre d'inscrits.

J'ai était un peu en difficulté pour la réalisation de ce code, notamment car il s'agit du premier que j'ai réaliser ainsi j'ai découvert via ce dernier la manière dont on codait.

Séance 3 :

Dans le contexte des paramètres statistiques, les caractères quantitatifs sont les plus étudiés. les paramètres statistiques "concernent principalement les variables quantitatives,

et ponctuellement qualitatives." Les paramètres élémentaires (moyenne, variance, écart type) sont en effet conçus pour les variables numériques, ce qui leur confère une importance centrale dans cette discipline, plus que les données à caractères qualitatifs.
Les caractères quantitatifs sont des variables numériques que l'on peut distinguer :

- Discrets : il s'agit de variables dont les calculs des paramètres sont basés sur des sommes et des effectifs.
- Continus : il s'agit ici de variables dont les calculs des paramètres sont basés sur des intégrales et des fonctions de densité.
- Distinction : la distinction est fondamentale car elle détermine la méthode de calcul des paramètres de position (comme la moyenne ou la médiane) et de dispersion.

Il existe plusieurs types de moyennes (arithmétique, quadratique, harmonique, géométrique) car chacune est adaptée à un contexte physique ou mathématique spécifique. La moyenne arithmétique est la plus courante, la moyenne Harmonique est adaptée aux taux ou aux vitesses et enfin la moyenne Géométrique est adaptée aux taux de croissance ou aux variables définies par un produit.

La médiane est calculée car elle est la valeur qui partage la série ordonnée en deux parties comprenant exactement le même nombre de données (50 % des effectifs de part et d'autre) mais également car elle est robuste. Elle n'est en effet pas influencée par les valeurs extrêmes (ou aberrantes), contrairement à la moyenne arithmétique. Elle est donc un meilleur résumé pour les distributions fortement dissymétriques.
Le mode qui est la modalité qui correspond à l'effectif maximal ou à la densité maximale peut être calculé dès lors qu'il y a une valeur la plus fréquente (toutefois, le mode n'existe pas toujours et n'est pas toujours unique).

La médiane et l'indice de Gini on tous les deux des intérêts. D'abord, concernant la médiale, elle coupe le groupe en deux parts égales d'argent (50 % de la masse salariale totale), tandis que la médiane coupe le groupe en deux parts égales d'individus (50 % des personnes ont un salaire inférieur). Donc cela veut dire que si la médiale est beaucoup plus élevée que la médiane, il faut un salaire très haut pour que la moitié de l'argent soit atteinte. C'est le signe d'une forte concentration (ou inégalité).
L'indice de Gini quant à lui est un chiffre entre 0 et 1 qui mesure l'inégalité de manière précise. 0 = égalité parfaite et 1 = inégalité maximale. Il est calculé grâce à la Courbe de Lorenz qui montre visuellement à quel point la répartition réelle s'éloigne de l'égalité parfaite

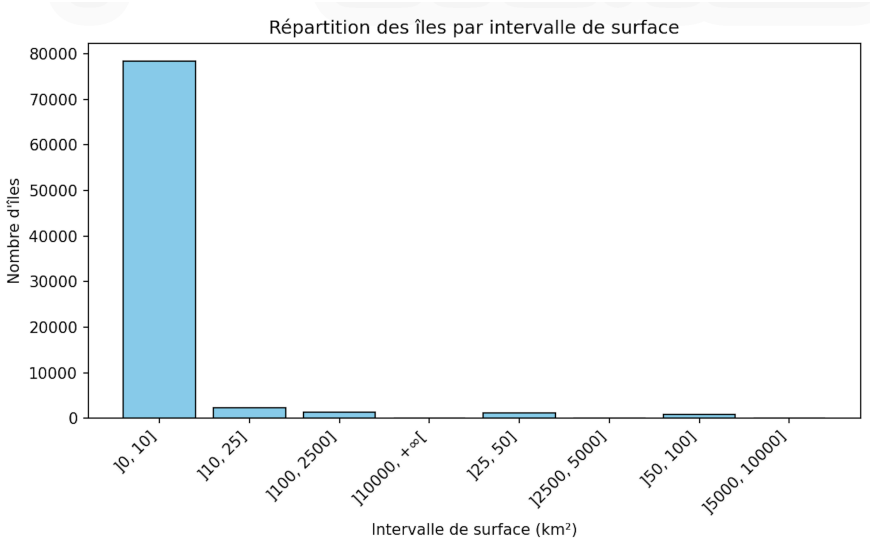
Les paramètres de dispersion permettent de mesurer à quel point les données sont étalées autour de la moyenne. On a la variance et l'écart type pour l'illustrer. La variance sert éviter le zéro, on prend le carré de ces différences. C'est la variance (V (X)). On prend en revanche pour l'écart type la racine carrée de la variance pour revenir à l'unité de mesure de départ (si la moyenne est en kilos, l'écart type est aussi en kilos).
L'étendue est la mesure la plus rapide pour connaître la largeur totale de la série. Pour la calculer on prend la plus grande valeur moins la plus petite valeur.

On a aussi les quantiles et l'écart interquartile. Les quantiles sont des points qui coupent la série ordonnée en parts égales (ex. : les quartiles coupent en 4 parts de 25 %). L'EIQ représente la distance entre le premier quartile et le troisième quartile. Il sert à mesure la dispersion des 50 % des données qui sont au centre. Il est plus fiable que l'étendue, car il ignore les valeurs extrêmes.

ROBLIN Manon

La boîte à moustaches constitue une façon de représenter graphiquement la dispersion et la position. Elle représente les 50 % des données centrales, la ligne à l'intérieur est la médiane et c'est un excellent outil pour comparer visuellement plusieurs groupes de données.

Les paramètres de formes mesurent l'aspect de la distribution (sa symétrie et son aplatissement).
Les moments (centrés) sont des calculs complexes qui permettent de caractériser la distribution. Les moments d'ordres 1 et 2 donnent la position (Moyenne) et la dispersion (Variance), tandis que les moments d'ordres 3 et 4 donnent la forme.
La symétrie sert à savoir si la distribution est équilibrée.
L'aplatissement sert quant à lui à mesurer si la distribution est pointue ou plate, par rapport à une courbe de référence (la loi normale).



Le graphique de la séance 3 illustre la répartition des îles par intervalle de surface (en km2). Comme pour la séance précédente, le résultat du code a donné un diagramme en barres qui montre comme évoqué précédemment la distribution du nombre d'îles en fonction de leur superficie en kilomètres carrés. La catégorie]0, 10] est la caractéristique la plus frappante car la barre pour les îles ayant une surface comprise entre 0 et 10 km2 domine totalement le graphique, atteignant une fréquence de près de 80 000 îles ce qui signifie que l'écrasante majorité des entités enregistrées comme « îles » sont de très petites îles, îlots ou rochers. La distribution est extrêmement asymétrique à droite. Pour la catégorie]10, 25] la fréquence est faible : (environ 2 000 îles), ce qui est déjà négligeable par rapport au premier groupe.
Concernant les autres intervalles (]25, 50],]50, 100],]100, 2500], …), elles présentent des fréquences si faibles qu'elles sont à peine visibles ou nulles sur ce graphique à l'échelle

choisie donc il doit probablement y avoir quelques dizaines ou centaines d'îles seulement dans ces intervalles.
Pour les très grandes îles (îles de plus de 10 000 km2), une barre est visible Sa hauteur est très basse (probablement entre 10 et 20 îles), mais elle est plus visible que les catégories intermédiaires. Ainsi ces îles représentent une minorité. Le graphique démontre que la taille des îles suit une loi de puissance ou une distribution fortement asymétrique : il existe un nombre astronomique de très petites îles/îlots, et très peu de grandes îles.

Pour la réalisation de ce code et comme pour beaucoup d'autres, j'ai rencontré des difficultés à retrouver les fichiers dans le dossier data. Souvent au moment d'exécuter le code, le terminal ne trouvait pas le fichier.

Séance 4 :

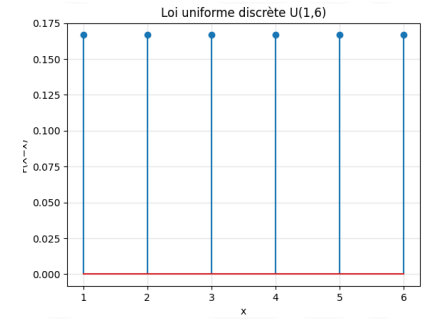
Pour choisir entre une distribution statistique discrète et une distribution continue, le critère principal est la nature de la variable aléatoire (X) à modéliser.
Pour les variables discrètes on utilise une distribution discrète lorsque la variable aléatoire ne peut prendre qu'un nombre fini ou dénombrable de valeurs (souvent se sont des nombres entiers). La variable résulte d'un comptage d'événements ou d'objets. Les lois qui y sont associées la loi de poisson, la loi binominale ou encore la loi hypergéométrique.
Pour les variables continues on utilise une distribution continue lorsque la variable aléatoire peut prendre n'importe quelle valeur réelle dans un intervalle donné (ou non). C'est typiquement le cas des mesures. Ici, la variable résulte d'une mesure comme un temps, une durée, des proportions ou des taux, ... Pour des proportions ou des taux c'est par exemple la loi Bêta qui est utilisée alors que pour un temps ou une durée ce sont plutôt les lois exponentielle, Weibull, ou encore Gamma. On utilise aussi la loi uniforme pour la répartition indifférente ou l'ignorance totale sur un intervalle ou la loi normale pour des grandeurs physiques soumises à de nombreux facteurs.

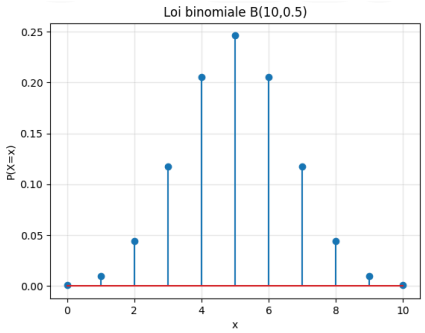
En géographie, les lois statistiques les plus utilisées sont les suivantes :

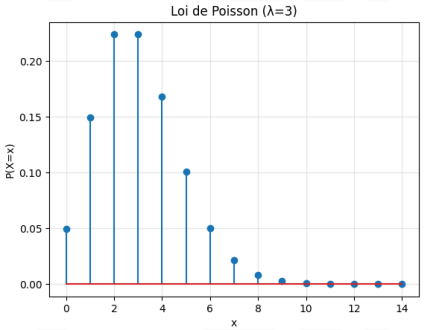
Loi Normale : Elle est utilisée comme un cliché pour les « erreurs » dans les mesures en physique. En géographie, elle peut être utilisée pour modéliser des variables soumises à de multiples facteurs.
La Loi Normale Centrée Réduite est utilisée pour comparer la répartition étudiée à une distribution de probabilité théorique.
La loi de Gibrat est liée à des variables aléatoires positives.
La loi de Benford est utilisée pour le dénombrement de certains objets géographiques et a déjà était testée par exemple pour mesurer la longueur des fleuves du globe.
Les lois des phénomènes rares et extrêmes :
Loi de Pareto qui est une loi fondamentale dans l'étude des phénomènes géographiques et sociaux où les valeurs extrêmes ont une probabilité élevée de se réaliser. Elle a été utilisée dès le début du XXe siècle pour les variables aléatoires concernant la fréquence du dénombrement de certains objets géographiques (lacs, montagnes, ...). Elle modélise des distributions scalantes et est liée à la notion de fractale.
Loi de Fréchet et Loi de Gumbel font quant à elles partie de la théorie des valeurs extrêmes, un domaine souvent appliqué aux phénomènes géographiques comme les crues ou les vents.

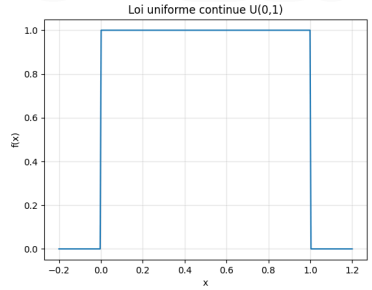
L'ensemble des graphiques illustre visuellement la différence entre les lois continues et les lois discrètes Les formes (asymétrie, symétrie, lourdeur des queues) sont directement

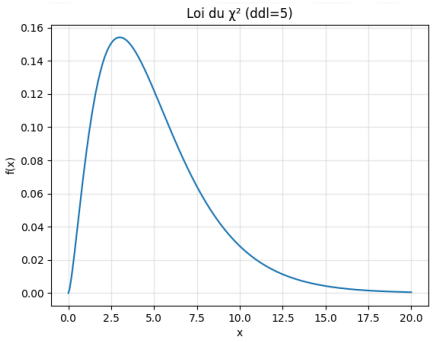
ROBLIN Manon

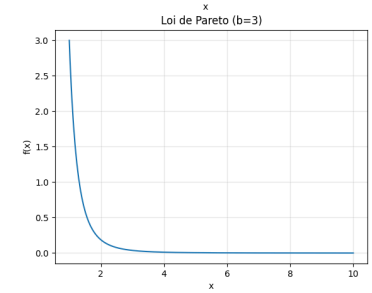


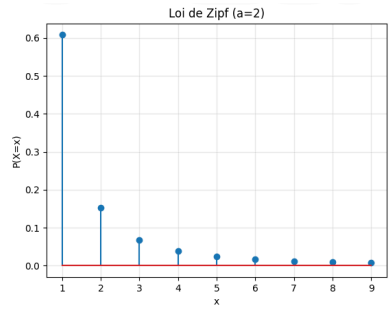




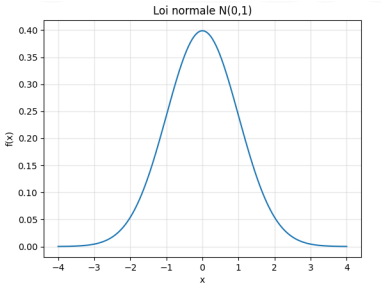


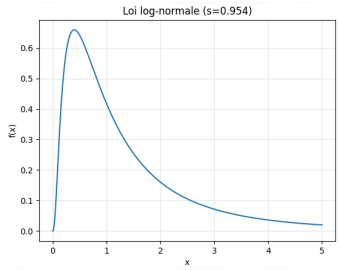


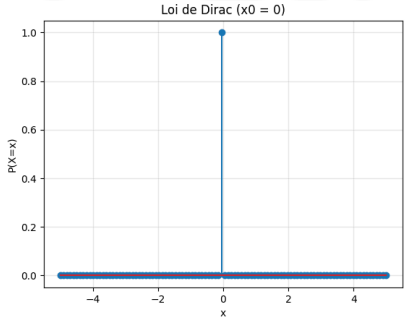




liées aux types de phénomènes qu'elles sont conçues pour modéliser.







Séance 5 :

L'échantillonnage consiste à prélever un sous-ensemble (un échantillon) d'une population mère pour étudier ses caractéristiques. On n'utilise pas la population en entier car c'est souvent impossible (population trop grande, infinie, ou nécessitant une destruction pour la mesure) ou trop coûteux/long en temps et en ressources. Le choix de la méthode dépend de la nature du phénomène étudié (pour choisir entre loi discrète ou continue) et de la forme de la distribution. Par exemple, si le phénomène étudié ne peut prendre que des valeurs dénombrables (nombres entiers, comptages, succès/échecs), il faut choisir une loi discrète.

L'estimateur est défini comme la variable de décision. Un estimateur est une statistique (une fonction des observations de l'échantillon) utilisée pour approximer la valeur d'un paramètre inconnu de la population. En revanche, l'estimation est la valeur numérique particulière que prend l'estimateur pour un échantillon donné.

L'intervalle de confiance sert à estimer un paramètre inconnu de la population à partir de l'échantillon, avec un certain niveau de confiance et utilise la valeur observée comme point de référence. L'intervalle de fluctuation, sert quant à elle à tester si un échantillon est compatible avec une population connue. Il est centré sur la valeur théorique de la population et donne la plage de valeurs où l'on s'attend à trouver la statistique de l'échantillon.

Dans la théorie de l'estimation, un estimateur est sans biais si son espérance mathématique est égale à la vraie valeur du paramètre qu'il estime. Par conséquent, un biais dans la théorie de l'estimation correspond à l'écart entre l'espérance mathématique de l'estimateur et la vraie valeur du paramètre qu'il est censé estimer.

On a coutume d'appeler les caractéristiques d'une population totale des paramètres. l'inférence statistique a été développée pour permettre aux statisticiens de généraliser les

ROBLIN Manon

conclusions d'un petit échantillon à l'ensemble de la population. Toutefois l'émergence des BigData a modifié ce paradigme. En effet, les technologies actuelles permettent souvent de collecter et d'analyser des ensembles de données qui approchent, voire incluent, la quasi-totalité de la population, c'est ce qu'on appelle un changement d'échelle. Dans le contexte du Big Data, les tests statistiques classiques peuvent devenir trop sensibles. Une différence minuscule et sans importance pratique peut être déclarée "statistiquement significative" simplement parce que la taille de l'échantillon (n) est très grande. Cela nécessite de se concentrer davantage sur la signification pratique plutôt que sur la seule signification statistique.

Les enjeux autour du choix de l'estimateur sont liés à la qualité de l'estimateur : s'assurer qu'il est sans biais, convergent (se rapproche du paramètre réel quand la taille d'échantillon augmente), et efficace (possède la plus petite variance possible). Il y a donc de fait un enjeu de la validité (absence de biais), un enjeu de la précision (efficaité) pour réduire la largeur de l'intervalle de confiance et avoir ainsi plus de précisions, un enjeu de la pertinence à long therme our garantir que l'estimation devient fiable et proche du paramètre réel lorsque l'on collecte de plus en plus de données ou encore un enjeu de la robustesse afin de garantir que l'estimation reste stable même si les hypothèses sur la distribution de la population sont légèrement violées.

L'estimation d'un paramètre, comme la moyenne ou l'écart type d'une population, repose sur deux approches principales. D'une part, l'estimation Graphique qui est une méthode visuelle rapide où les paramètres sont déterminés à partir de la pente et de l'intersection, après avoir linéarisé la fonction de répartition de la loi théorique (par exemple, la droite de Henry pour la Loi Normale ou une anamorphose logarithmique pour la Loi Exponentielle).
D'autre part, la Méthode du Maximum de Vraisemblance est une approche analytique rigoureuse. Elle consiste à trouver les valeurs des paramètres qui rendent maximalement probable l'observation de l'échantillon tel qu'il a été prélevé (par exemple, pour estimer les paramètres B et N de la Loi de Weibull).

Le choix de la loi théorique à estimer s'opère selon une démarche logique : Il faut d'abord choisir entre une loi discrète (pour les comptages, comme un nombre de défauts) et une loi continue (pour les mesures, comme la durée de vie) ? nature du phénomène. Ensuite, on a l'observation de l'histogramme de l'échantillon aide à affiner le choix : une forme symétrique suggère souvent la Loi Normale, tandis qu'une forme dissymétrique oriente vers des lois comme Weibull ou Gamma ? c'est la forme de la distribution.

Il existe de nombreux textes statistiques :

Tests Paramétriques Classiques : test sur la moyenne, sur l'écart type et test de fischer Snedecor (pour l'égalité des variances). ? servent à déterminer si un paramètre (moyenne, écart-type) est égal à une valeur spécifiée.

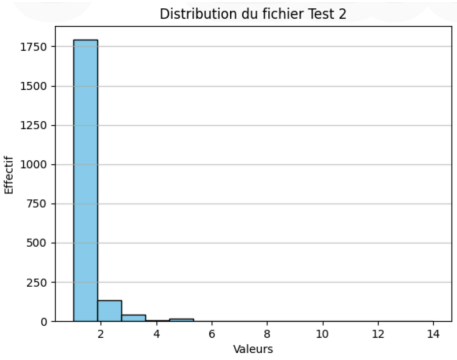
Test d'ajustements : test du X_2 de Pearson, de Kolmogorov-Smirnov, de Cramér-Von Mises, de Shapiro-Wilk ? visent à juger l'adéquation entre une situation réelle et un modèle théorique (ils vérifient si un échantillon suit une loi spécifiée).

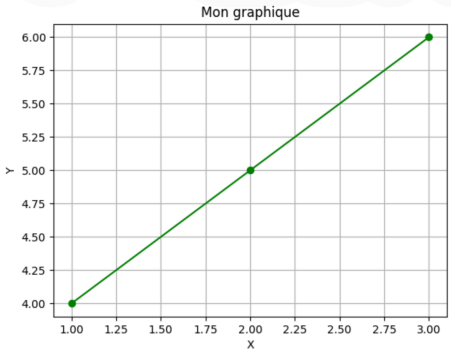
Tests de comparaison : test de comparaison des moyennes, testes non paramétriques (Test de Smirnov, Test de Wilcoxon/Mann-Whitney) ? il sert à décider si deux ou plusieurs échantillons sont issus de la même population.

ROBLIN Manon

Pour créer un test il faut dans un premier temps définir les hypothèses. Il faut ensuite définir une variable de décision, établir la loi de la statistique sous l'Hypothèse H_0, fixer ensuite un risque de première espèce, déterminer la région critique (en fonction bien sûr de la loi), calculer la valeur de la statistique à partir de l'échantillon et enfin énoncer une règle de décision.

Ce que je peux retenir et émettre comme critiques, c'est que la statistique inférentielle, en soi, n'est pas un mauvais outil, mais qu'il faut l'utiliser avec dissernement. au lieu de se contenter d'un simple "C'est significatif/Ce n'est pas significatif", il faudrait d'avantage utiliser des outils qui nous montrent à quel point l'effet est important dans la réalité. C'est ça qui fait qu'on doit utiliser la statistique inférentielle avec discernement. Par exemple, l'intervalle de confiance ou encore la taille d'Effet semblent être plus précis sur les effets et leurs conséquences.





Pour la séance 5 nous avons ces deux graphiques Le graphique distribution du fichier test 2 montre la distribution des fréquences d'une variable quantitative, ce qui est essentiel pour comprendre la structure des données. Ici, les valeurs sont regroupées en classes donc en colonnes. L'axe où est renseigné l'effectif indique le nombre d'occurrences pour chaque classe de valeurs.
Pour le graphique n°2 est illustré la relation entre la variable X (en abscisse) et la variable Y (en ordonnée). On a une fonction qui relie X à Y et la relation entre ces fonctions est linéaire car les points sont parfaitement alignés.

Séance 6 :

Quand on parle de statistique ordinale nous faisons référence à une suite d'observations que l'on nomme X_i, classées le plus souvent par ordre croissant. Elle s'oppose implicitement aux statistiques utilisées pour les variables nominales où il n'existe pas d'ordre naturel. Elle utilise principalement des variables quantitatives ou ordinales. Elle est le cœur de la géographie humaine et matérialise une hiérarchie en opérant des classements d'objets géographiques (villes, régions, ...) pour évaluer s'ils ont monté,

stagné ou descendu dans la hiérarchie. L'ordre à privilégier dans les classifications est l'ordre croissant.

Nous pouvons faire une différence entre une corrélation des rangs et une concordance de classements. La corrélation des rangs mesure l'association ou l'accord entre deux classements d'un même ensemble d'objets tandis que la concordance des classements mesure l'accord global entre classements d'un même ensemble d'objets. C'est une généralisation de la corrélation des rangs de Kendall.

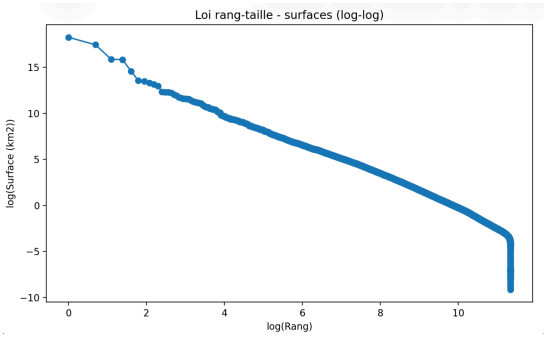
Les tests de Spearman et de Kendall sont des tests non paramétriques pour comparer deux classements. Ils reposent sur des principes de calcul différents :

Test Spearmen : Basé sur la distance, il mesure la force de la relation monotone en quantifiant la distance entre les rangs des deux classements. Il utilise la somme des carrés des différences de rangs. Il est plus sensible aux grandes différences de rangs, car les écarts sont mis au carré. Une seule grande différence a un impact important.

Test de Kendall : Basé sur l'ordre, il mesure la probabilité relative d'accord en comptant le nombre de paires d'objets classées dans le même ordre(concordantes) ou dans un ordre opposé (discordantes). Il utilise la différence nette entre les paires concordantes (+1) et discordantes (-1). Il moins sensible aux grandes différences de rangs, car chaque inversion de paire compte seulement comme , quelle que soit l'ampleur de l'écart.

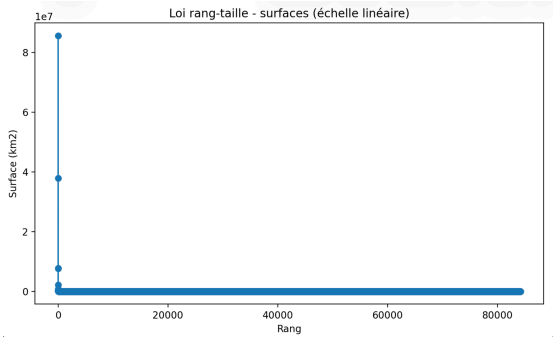
Les coefficients de Goodman-Kruskal et de Yule sont des mesures d'associations pour les variables catégorielles.

Le coefficient de Goodman-Kruskal sert à mesurer l'association entre variables ordinales tandis que le coefficient d'association de Yule Sert à mesurer l'association pour les matrices 2 x 2.



Ici nous avons un graphique qui représente la relation entre le logarithme du rang et le logarithme de la surface (en km2). Il s'agit de la visualisation standard de la loi de Zipf (ou loi Rang-Taille). Pour les petits rangs (à l'extrême gauche, log(Rang) proche de 0), on observe une légère courbure vers le haut. Ces points représentent les quelques entités les plus grandes (ex: les plus grandes villes ou les plus grands pays). Pour les grands rangs (à l'extrême droite, log(Rang) proche de 12), la courbe plonge brutalement, représentant les très nombreuses entités de très petite taille (petites villes, petites surfaces). Le graphique confirme que la distribution

des surfaces suit la Loi de Zipf : les grandes surfaces sont très rares (Rang 1, 2, etc.), et il y a une multitude d'entités de très petite surface (Rang élevé).



Ce graphique représente exactement les mêmes données que le précédent, mais en utilisant des échelles linéaires pour les deux axes. Le premier point, correspondant au Rang 1 (la plus grande surface), écrase complètement le graphique. Sa surface est de l'ordre de 8×107 km2 (80 millions de km2). La surface du Rang 2 est déjà beaucoup plus petite (autour de 4×107 km2). Dès le Rang 3 ou 4, la surface a chuté à des valeurs si petites qu'elles sont indiscernables de l'axe des X (Surface = 0). Pour tous les rangs supérieurs à un petit nombre, les points se confondent avec l'axe horizontal. L'échelle de l'axe Y, dominée par la surface du Rang 1, ne permet pas de distinguer les variations entre les dizaines de milliers d'entités de petite taille. I illustre pourquoi l'échelle linéaire est inadéquate pour visualiser une distribution de type Rang-Taille/Loi de Zipf.
Il montre visuellement et de manière spectaculaire que l'entité classée première est immensément plus grande que toutes les autres.

ROBLIN Manon