

Analyse de données

Rapport d'activité

Séance 1

À l'issue de la séance 1, le but affiché était de se familiariser avec les outils Github, et Dockers après les avoir installés sur sa machine personnelle. Si la première étape de se créer un compte a été relativement aisée, l'installation des applications m'a nécessité quelques semaines supplémentaires, avant même de pouvoir me lancer dans le travail attendu.

J'ai par la suite éprouvé de nombreuses difficultés à installer et opérer Python ainsi que Github, et ce malgré une aide extérieure reçue. Docker reste d'ailleurs pour moi toujours un mystère.

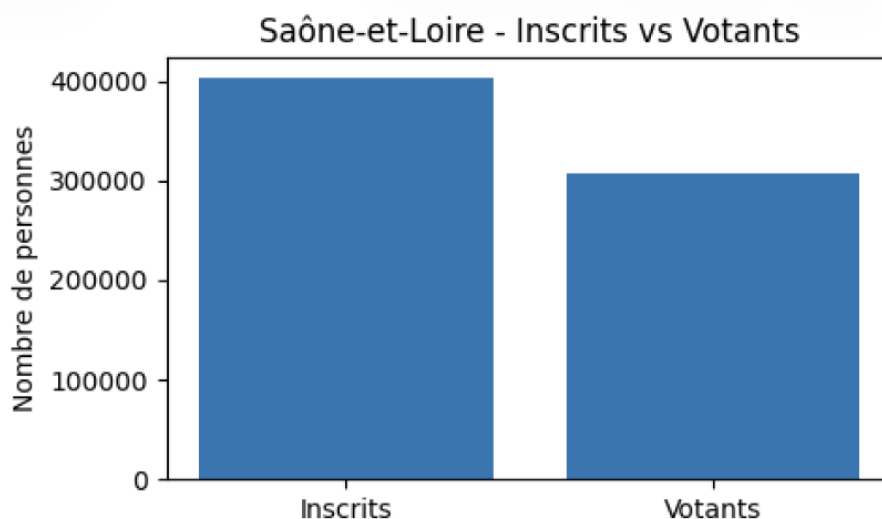
Séance 2

1. Les statistiques en géographie correspondent à l'ensemble de l'information géographique permettant de classer un ensemble de données selon des objets géographiques étudiés.
2. Le hasard existe en géographie même si l'importance qu'il occupe reste mitigée. Dans la pratique de la géographie française, le hasard existe dans la mesure où elle suit le cadre de la géographie vidalienne qui offre la possibilité qu'un événement se produise ou qu'il ne se produise pas. Dans le cadre de la statistique, il est possible de dégager une certitude globale, sans être cependant capable de prévoir le détail des réalisations. En géographie humaine, il est tout simplement impossible de prévoir l'action de chaque individu mais des tendances peuvent être observées.
3. Les types d'information géographiques sont tout ce qui caractérise l'ensemble délimité par des éléments de géographie humaine et physique, ainsi que la morphologie même des ensembles délimités.
4. En matière d'analyse de données, la géographie a besoin de données (documentées par des méta-données) dont elle peut étudier la structure et confronter les résultats obtenus avec la méthode de production des données et les connaissances du phénomène étudié.
5. La statistique descriptive est l'étude des données ayant pour objectif d'en dégager des propriétés remarquables par rapport à une distribution théorique connue afin d'obtenir un tableau simplifié de la réalité. La statistique explicative cherche à déterminer l'influence d'une ou de plusieurs variables sur une autre.
6. Les types de visualisation des données utilisées en géographie sont l'analyse factorielle en composantes principales utilisée pour visualiser des données

quantitatives, l'analyse factorielle des correspondances pour les variables qualitatives, l'analyse factorielle des correspondances multiples qui permet de visualiser plus de deux variables qualitatives et l'analyse factorielle des données mixtes lorsque les données sont hétérogènes. On les choisit donc en fonction du type de données et de la quantité de variables.

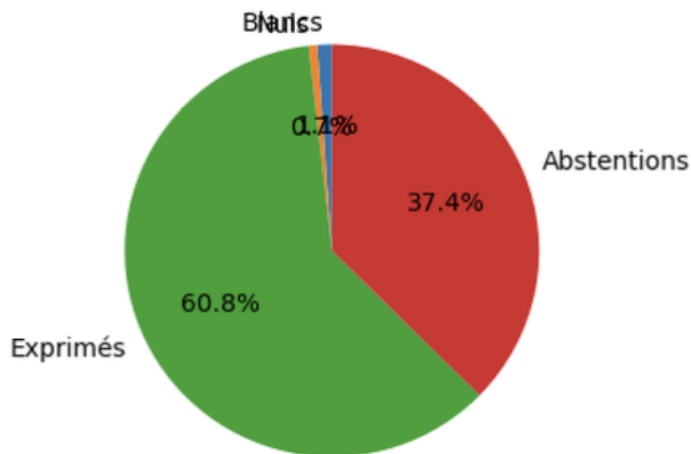
7. Les méthodes d'analyse de données possibles sont les méthodes descriptives, les méthodes explicatives, et les méthodes de prévision.
8. a) Une population statistique est l'ensemble d'individu étudiés ; b) l'individu statistique est un élément de la population statistique ; c) les caractères statistiques sont les caractères, les éléments particuliers de l'individu dans la population statistique étudiée ; d) les modalités du caractère correspondent aux valeurs prises par un caractère. Ce caractère peut-être une variable qualitative ou quantitative.
Non.
9. On mesure une amplitude en soustrayant la valeur maximale à la valeur minimale de la classe. On appelle densité le rapport entre l'effectif divisé par l'amplitude.
10. Les formules de Sturges et de Yules permettent de donner une valeur approximative du nombre de classes.
11. L'effectif (ou fréquence absolue) d'une variable est le nombre de fois où cette variable apparaît dans la population. On calcule une fréquence en divisant l'effectif par l'effectif total. On calcule une fréquence cumulée en réalisant la somme des effectifs associés aux valeurs du caractère qui sont inférieurs ou égales à 1.

Résultats :



Diagrammes en barres comparant le nombre d'inscrits et le nombre de votants pour chaque département.

Haute-Corse - Répartition des votes



Diagrammes circulaires représentant la répartition des votes

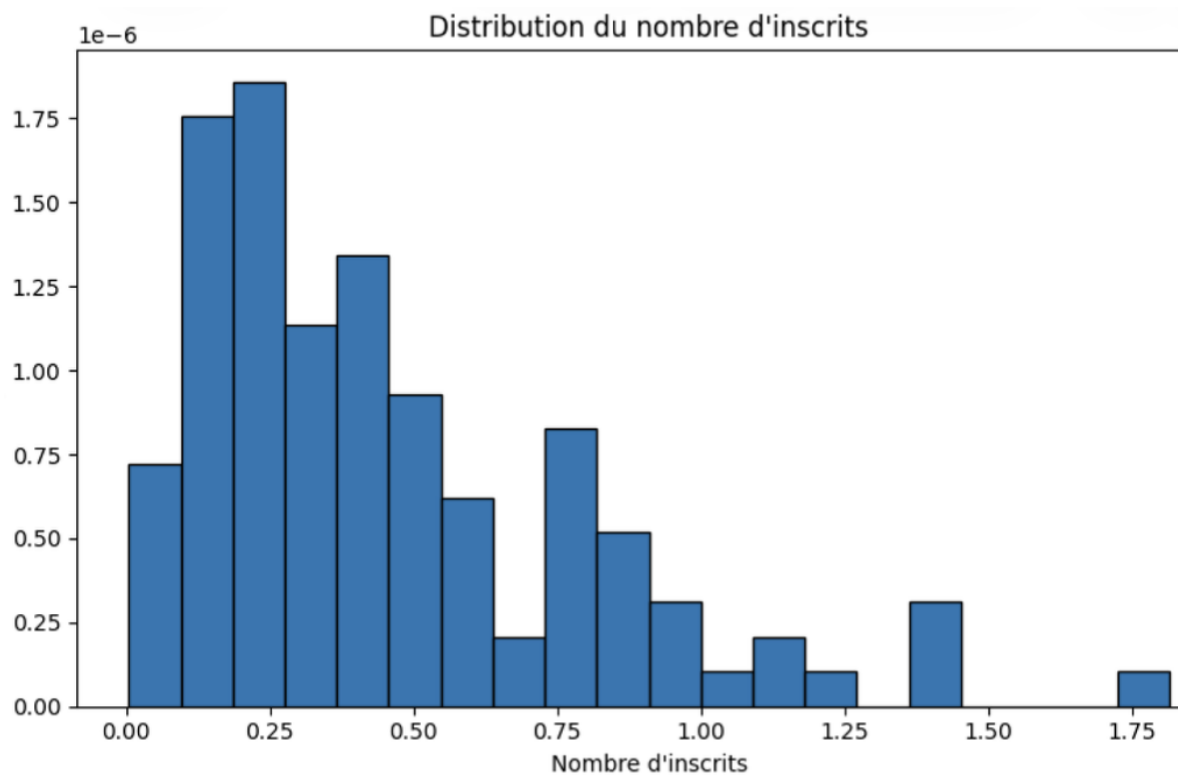


Diagramme en barres pour chaque département:

Séance 3 :

1. Le caractère qualitatif est le concept le plus général en statistique. Un caractère représente une propriété observée sur un individu, comme la couleur des yeux ou la taille. Les caractères qualitatifs : décrivent des catégories ou attributs non mesurables numériquement (ex. : "Bleu", "Femme", "Ville"). Les caractères

quantitatifs : ils sont un cas particulier, décrivant des propriétés mesurables par un nombre (ex. : 1,75 m, 20 ans).

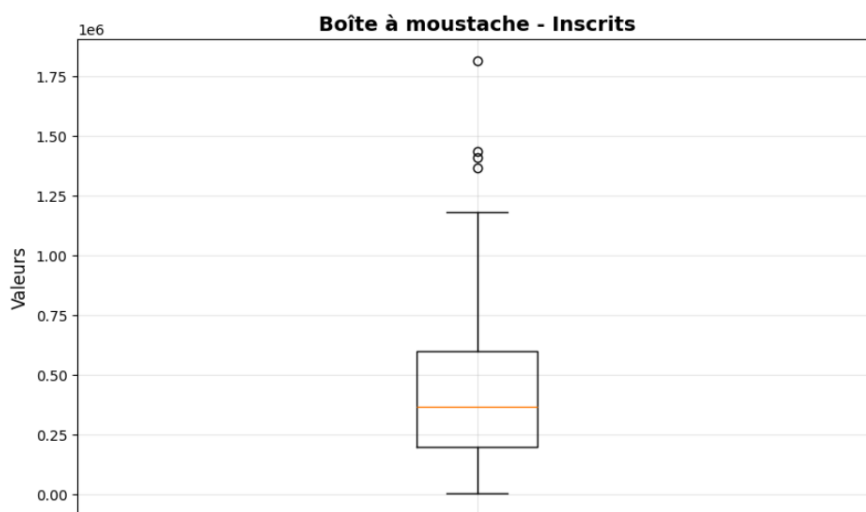
2. Les caractères quantitatifs peuvent être discrets ou continus selon l'ensemble des valeurs qu'ils peuvent prendre. Les caractères quantitatifs discrets ne prennent que des valeurs précises et dénombrables, souvent entières. Les caractères quantitatifs continus peuvent prendre toutes les valeurs possibles dans un intervalle donné. Ils nécessitent un regroupement en classes pour être représentés, par exemple via un histogramme.
3. Les paramètres de position permettent d'identifier le centre d'une distribution de données. Il existe plusieurs types de moyennes (arithmétique, quadratique, harmonique, géométrique). La moyenne arithmétique est la plus courante mais sensible aux valeurs extrêmes. D'autres moyennes, comme la géométrique, sont utilisées selon le contexte (ex. taux de croissance). La médiane est la valeur qui partage les données en deux parties égales après classement. Elle est robuste aux valeurs extrêmes et utile pour les distributions asymétriques. Le mode est la valeur la plus fréquente. Elle peut être unique ou multiple (distribution bimodale ou multimodale). Le mode est le seul paramètre applicable aux données qualitatives nominales.
4. Les paramètres de concentration mesurent comment la masse totale d'un caractère est répartie parmi les individus. La médiale est proche de la médiane, mais elle divise la masse totale de la variable en deux parties égales. Par exemple, pour les revenus, elle sépare les individus selon la moitié des revenus totaux. La comparaison avec la médiane permet d'évaluer la concentration. Indice et courbe de Gini : ils décrivent la concentration d'une population (souvent pour les revenus). L'indice de Gini varie entre 0 (égalité parfaite) et 1 (concentration maximale).
5. Les paramètres de dispersion mesurent l'étalement des données autour d'un centre. Variance et écart-type : la variance calcule la moyenne des carrés des écarts à la moyenne, donnant plus de poids aux valeurs éloignées. L'écart-type (racine carrée de la variance) permet de revenir à l'unité originale. Petit écart-type : données regroupées autour de la moyenne. Grand écart-type : données dispersées. Étendue : différence entre la valeur maximale et minimale. Simple mais dépend uniquement des extrêmes. Les quantiles et écart interquartile divisent la série ordonnée en parties égales. L'écart interquartile ($Q3 - Q1$) contient 50% des observations centrales. La boîte à moustaches : représente graphiquement la valeur minimale, $Q1$, la médiane, $Q3$ et la valeur maximale. Elle permet de visualiser la dispersion, l'asymétrie et de comparer plusieurs séries.
6. Les paramètres de forme. Les moments caractérisent la forme globale d'une distribution (symétrie, aplatissement). Moments centrés : moyenne des puissances des écarts à la moyenne, utilisés pour mesurer l'asymétrie (moment d'ordre 3). Moments absolus : moyenne des valeurs absolues des écarts à la moyenne, moins sensibles aux valeurs extrêmes. Vérification de la symétrie : importante pour choisir les méthodes statistiques. Comparer moyenne, médiane et mode : différences importantes \rightarrow forte asymétrie. Observation graphique : histogrammes ou boîtes à moustaches. Coefficients d'asymétrie basés sur les moments centrés pour une mesure quantitative.

Séance 4 :

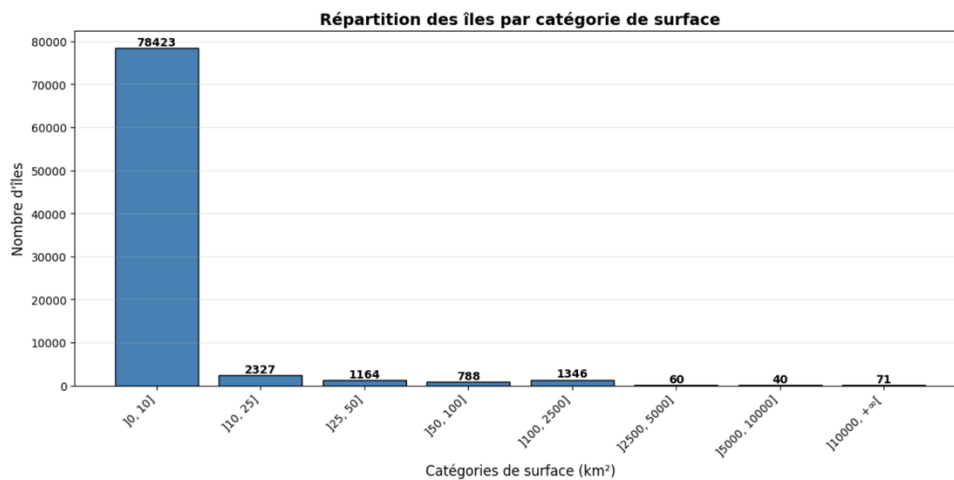
Le choix entre une distribution statistique à variables discrètes ou à variables continues dépend principalement de la nature du phénomène analysé. Lorsqu'une variable ne peut prendre qu'un nombre limité ou dénombrable de valeurs, comme lors du comptage de personnes, d'événements ou de résultats, il est alors approprié d'utiliser une distribution discrète. En revanche, lorsque la variable peut adopter toutes les valeurs possibles sur un intervalle continu, comme c'est le cas pour la température, la durée, la distance ou encore le revenu, une distribution continue est plus pertinente.

Ce choix est également influencé par la forme de la distribution, c'est-à-dire par la façon dont les données sont réparties. Les indicateurs statistiques de la série, tels que l'espérance, la médiane, la variance, l'écart type ou le coefficient d'asymétrie, constituent des éléments essentiels pour déterminer la loi la mieux adaptée. Par ailleurs, le nombre de paramètres propres à chaque loi statistique peut aussi intervenir, certaines distributions offrant un ajustement plus satisfaisant que d'autres en fonction de la complexité des données étudiées.

Parmi les lois statistiques couramment utilisées en géographie, la loi de Zipf occupe une place importante. Elle met en relation le rang d'un élément et sa taille, en montrant que cette dernière est inversement proportionnelle à son rang. Cette loi est notamment employée pour étudier les hiérarchies urbaines et les répartitions de population, car elle permet de faire apparaître des régularités dans l'inégale distribution des villes au sein d'un espace donné.



Boîte à moustache



Graphique en bâton de la répartition des îles du monde selon leur superficie.

Séance 5 :

L'échantillonnage consiste à extraire de manière aléatoire une partie d'une population de référence afin d'en déduire des informations concernant l'ensemble de cette population. Cette démarche relève de la statistique inférentielle. L'analyse complète d'une population est le plus souvent irréalisable ou trop coûteuse sur le plan matériel et organisationnel, en particulier lorsque les effectifs sont importants. Par ailleurs, lorsque la taille exacte de la population mère est inconnue, il devient difficile de sélectionner un modèle statistique adapté. On distingue principalement deux types d'échantillonnage : l'échantillonnage non biaisé, fondé sur le hasard et l'égalité des probabilités de sélection, et l'échantillonnage biaisé. Le choix de la méthode dépend des contraintes pratiques et des objectifs poursuivis par l'étude.

Un estimateur est une fonction définie à partir d'un échantillon permettant d'approcher au mieux la valeur inconnue d'un paramètre de la population mère. L'estimation correspond quant à elle à la valeur numérique obtenue lorsque cet estimateur est appliqué à un échantillon donné, fournissant ainsi un résultat chiffré basé sur les observations.

Il est essentiel de distinguer l'intervalle de fluctuation de l'intervalle de confiance. L'intervalle de fluctuation désigne l'ensemble des valeurs qu'une fréquence observée a de fortes chances de prendre, en supposant connu un paramètre de la population ; il permet d'évaluer l'ampleur des variations possibles d'un échantillon. À l'inverse, l'intervalle de confiance vise à estimer un paramètre inconnu de la population mère à partir d'un échantillon, en déterminant une plage de valeurs dans laquelle ce paramètre est susceptible de se situer.

Le biais correspond à la différence entre l'espérance mathématique d'un estimateur et la valeur réelle du paramètre étudié. Lorsque cette différence est nulle, l'estimateur est qualifié de sans biais ; dans le cas contraire, il est considéré comme biaisé.

Une statistique reposant sur l'analyse de l'ensemble de la population est dite exhaustive. Elle exploite toutes les informations disponibles afin de conserver l'intégralité des données utiles. Les données massives s'inscrivent dans cette logique d'exhaustivité : elles portent sur des volumes très importants d'informations couvrant l'ensemble de la population connue et nécessitent des techniques spécifiques de traitement en raison de leur ampleur.

Le choix d'un estimateur constitue un enjeu majeur en statistique. Un estimateur pertinent doit fournir une estimation ponctuelle aussi proche que possible du paramètre inconnu, quel que soit l'échantillon considéré. Les principaux critères de qualité sont l'absence de biais, une variance faible garantissant une bonne précision, ainsi qu'une résistance aux valeurs aberrantes.

Différentes méthodes existent pour estimer un paramètre. La méthode des moindres carrés est utilisée lorsque les grandeurs à estimer correspondent à des espérances. La méthode du maximum de vraisemblance et la méthode des moments sont également fréquemment employées. Le choix entre ces approches dépend des propriétés attendues de l'estimateur, telles que sa convergence, son efficacité et sa robustesse.

Les tests statistiques permettent de valider ou d'infirmer des hypothèses à partir de données observées. On distingue les tests paramétriques, qui portent sur des paramètres comme la moyenne, l'écart type ou la forme de la distribution (par exemple le test de Student), et les tests non paramétriques, fondés sur des statistiques comme les effectifs ou la médiane, tels que le test de Mann-Whitney. Il existe plusieurs catégories de tests, notamment les tests de conformité, d'homogénéité, d'indépendance, d'adéquation à une loi de probabilité ou encore les tests sur séries appariées. Ces outils permettent d'étudier les relations entre plusieurs échantillons afin de mettre en évidence d'éventuels liens entre des phénomènes ou des événements.

La construction d'un test d'hypothèse commence par la formulation d'une hypothèse de travail, ainsi que d'une hypothèse nulle et d'une hypothèse alternative. Après avoir fixé le seuil de risque et choisi la loi de probabilité appropriée, on détermine les paramètres du test, notamment l'échantillon et sa variance. Les données collectées permettent ensuite de calculer la statistique de test. En comparant cette valeur à la région critique définie au préalable, il est alors possible d'accepter ou de rejeter l'hypothèse nulle et de conclure le test.

Les critiques formulées à l'encontre de la statistique inférentielle mettent en avant sa complexité et questionnent parfois la fiabilité de ses résultats. Toutefois, même si ces critiques peuvent fragiliser la confiance qui lui est accordée, elles ne remettent pas en cause son utilité pour l'analyse des phénomènes, à condition que ses méthodes soient appliquées avec rigueur.

Séance 6 :

La statistique ordinale concerne des données qui peuvent être ordonnées selon un classement. Elle se différencie de la statistique nominale, laquelle se limite à regrouper les données en catégories sans établir de hiérarchie. La statistique ordinale s'appuie sur des variables quantitatives et permet de classer des entités à partir d'un critère commun. En géographie, elle sert notamment à représenter des hiérarchies spatiales en ordonnant les territoires ou les unités spatiales selon leur position relative.

Dans les opérations de classement, il est généralement recommandé d'utiliser un ordre croissant, considéré comme l'ordre le plus intuitif, car il facilite la compréhension et l'interprétation des résultats.

La corrélation des rangs a pour objectif de déterminer si deux séries de données présentent un ordre similaire, en mesurant la relation entre deux classements. La notion de concordance des classements s'intéresse davantage à la cohérence des critères utilisés pour organiser les données au sein d'un ensemble.

Les tests de Spearman et de Kendall permettent tous deux d'évaluer une corrélation entre rangs, bien qu'ils reposent sur des approches de calcul distinctes. Le coefficient de Spearman est fondé sur l'écart entre les rangs attribués à chaque paire d'éléments : il mesure la similarité des distances entre les positions occupées dans deux classements différents.

Le test de Kendall, quant à lui, considère l'ensemble des paires possibles et compare le nombre de paires dont l'ordre est identique dans les deux classements (paires concordantes) à celles dont l'ordre est inversé (paires discordantes).

Les coefficients de Goodman-Kruskal et de Yule permettent également d'évaluer le degré de concordance entre des classements. Le coefficient de Goodman-Kruskal mesure l'association entre variables ordinales en comparant le nombre de paires concordantes et discordantes ; sa valeur est comprise entre -1 et $+1$ et son interprétation est proche de celle des coefficients de Spearman et de Kendall.

Le coefficient de Yule est une forme particulière du coefficient de Goodman-Kruskal, utilisée lorsque les données sont présentées sous la forme d'un tableau de contingence 2×2 . Dans ce contexte, une valeur négative indique une relation inverse, une valeur positive traduit une association parfaite, tandis qu'une valeur nulle signifie l'absence de lien entre les variables.

