Bien qu'elle produise des données massives que seul l'outil statistique permet d'étudier, la géographie à tendance à mépriser ce terme. En effet, les géographes ont historiquement un rapport compliqué à la statistique. Ces statistiques sont pourtant essentielles pour analyser l'information géographique, ainsi que pour le développement scientifique de la discipline.
La question du hasard est assez ambivalente en géographie, dans la mesure où deux positions philosophiques s'opposent. Pour le déterminisme, le hasard n'existe pas car il existe une cause à tout, tandis qu'une autre posture admet que le hasard existe, mais sera toujours, d'un moment ou un autre, explicable. On note deux types de hasard dans les modélisations mathématiques, à savoir le hasard bénin et le hasard sauvage.
L'information géographique peut se décomposer en deux séries statistiques, à savoir par la caractérisation de l'ensemble délimité (données humaines ou physiques), ainsi que par l'étude de la morphologie de ces ensembles.
Les statistiques présentent un outil essentiel en ce qui concerne l'avenir et la compréhension de l'information géographique. En effet, l'analyse de données repose sur les probabilités et les statistiques, deux éléments essentiels d'étude de la structure interne des données analysées.
La statistique descriptive repose sur la synthétisation et la description de données, afin d'obtenir une image plus simplifiée de la réalité. La statistique explicative (ou mathématique), quant à elle, prévoit des scénarios possibles, et repose donc sur la compréhension des données. Elle se fait à partir des données et de la distribution de probabilité théorique associée.
La visualisation de données, qui se fait sous la forme de graphiques, dépend du type de variable étudiée. Elle peut reposer sur un histogramme pour les valeurs en continues, ou sur une représentation sectorielle pour les variables qualitatives. On note plusieurs choix de graphiques selon la nature de la variable. En effet, elle peut reposer sur le choix de variables qualitatives nominales, souvent représentées par un graphique en secteurs, ou ordinales (histogramme disjoint). Elle peut également reposer sur le choix de variables s'appuyant sur des dénombrements, à savoir par le choix de variables quantitatives discrètes (diagramme en bâtons), ou bien continues (histogrammes, boîtes à moustaches, lissage normal).
On note trois grandes classes en méthodes statistiques d'analyse des données :

- la méthode descriptive : permettant de résumer et d'organiser l'information
- la méthode explicative : cherchant à établir les relations entre les variables
- la méthode de prévision : permettant d'anticiper l'évolution d'une donnée dans le temps

À cela s'ajoutent différents outils ou techniques, comme l'ACP, l'AFC ou encore l'ACM (pour n'en citer que quelques exemples), permettant de représenter ou réduire la complexité des données.
En statistique, la population correspond à l'ensemble des éléments que l'on observe. L'individu est le plus petit élément qui compose cette population. En géographie, ces individus proviennent souvent d'un découpage spatial : ce sont alors des unités spatiales.
Le ou les caractères d'un individu désignent les informations ou caractéristiques que l'on mesure. Les modalités d'un caractère sont les différentes valeurs possibles qu'il peut prendre.
Il existe deux grandes natures de caractères :

- les variables qualitatives, qui décrivent un état ou une catégorie;
- les variables quantitatives, qui expriment une quantité issue d'une mesure ou d'un dénombrement (avec leurs quatre sous-catégories vues auparavant).

Enfin, les modalités constituent une partition du caractère : elles doivent être exhaustives et mutuellement exclusives.
L'amplitude d'une classe correspond à la différence entre ces bornes : b - a, ou a est la valeur minimale et b la valeur maximale.
La densité, noté d, est obtenue en divisant l'effectif de la classe n; par son amplitude. Elle exprime donc la concentration des valeurs dans cette classe.
La formule de Sturges permet de donner une valeur approximative du nombre k de classes pour construire un histogramme.
La formule Yule repose sur la même idée : elle sert également à déterminer un nombre de classes adapté à la distribution étudiée.
L'effectif n correspond au nombre d'individus associés à une valeur ou une classe.
On obtient la fréquence d'une valeur en divisant son nombre d'occurrence par le nombre total d'occurrences. Quant à la fréquence cumulative, elle représente la somme des fréquences des valeurs qui sont inférieures ou égales à k.
Ces fréquences servent à établir une distribution statistique empirique, c'est-à-dire un tableau qui regroupe les classes ou valeurs potentielles ainsi que leurs occurrences et/ou fréquences d'apparition.

Le caractère le plus général est le caractère quantitatif. En effet, il permet de mesurer et de comparer les données numériquement, ce qui rend possible l'utilisation de l'ensemble des outils statistiques.

- paramètres de position (moyenne, médiane, mode…)
- paramètres de dispersion (variance, écart-type, écart interquartile…)
- paramètres de forme (asymétrie, aplatissement…)

On voit que tous ces paramètres sont construits à partir de valeurs numériques, et donc que l'application des ces paramètres est uniquement possible si le caractère est quantitatif.
*"Les paramètres statistiques concernent **principalement** les variables quantitatives, et **ponctuellement** qualitatives."*

Les caractères quantitatifs **discrets** correspondent à des valeurs numériques qui sont exclusivement des nombres entiers, isolés à l'intérieur d'un intervalle de variation. Les caractères quantitatifs **continus**, au contraire, sont des valeurs pouvant prendre toutes les valeurs possibles à l'intérieur d'un intervalle de variation. Le plus souvent des nombres décimaux ou des %. On peut observer, en prenant l'exemple de la moyenne et de la médiane, que le calcul de ces paramètres change selon la nature de la variable. La distinction est importante entre caractères quantitatif discret et quantitatif continu car elle permet d'adapter correctement les formules et méthodes (calcul des paramètres statistiques) utilisées lors de l'analyse statistique.

Paramètre de position

Il existe plusieurs façons de calculer des moyennes en fonction de la nature (discrète ou continue) des variables. La moyenne arithmétique est la plus utilisée, mais elle ne suffit pas toujours dans la mesure où elle n'est pas toujours adaptée à la nature des données ou à la situation étudiée. Mais il existe d'autres moyens qui vont être plus adaptés : la moyenne quadratique , la moyenne harmonique , la moyenne géométrique, et la moyenne mobile. Et donc plusieurs types de moyenne permettent de choisir la formule la plus adaptée au phénomène observé.

La médiane est présentée comme la valeur (m_e) qui coupe une série de données en deux parties d'effectif égal. On calcule une médiane pour obtenir une mesure de position qui ne soit pas influencée par des valeurs extrêmes. Contrairement à la moyenne arithmétique,

la médiane dépend du classement des données et non de leur amplitude. Elle permet donc de représenter le centre d'une distribution même si la elle est fortement dissymétrique ou contient des valeurs aberrantes.

Il est possible de calculer un mode (m_o) lorsqu'une série statistique présente au moins une valeur qui apparaît plus souvent que les autres. Le mode correspond à la modalité, la valeur dont l'effectif (ou la densité) est maximal, la plus fréquente. Le mode existe seulement si une valeur se détache clairement, et elle n'est pas toujours unique : certaines distributions peuvent être bimodales ou plurimodales.Ainsi, on peut calculer un mode dès qu'une ou plusieurs valeurs dominantes se dégagent dans la distribution.

Paramètre de concentration

La médiane est utile parce qu'elle partage la valeur totale du caractère étudié en deux parts égales. Contrairement à la médiane, qui sépare les individus en deux groupes d'effectif identique, la médiale sépare selon l'importance totale du caractère (par exemple la masse salariale). Elle sert donc à analyser les situations où la répartition du total est plus significative que la simple répartition des individus. L'indice de concentration de C. Gini sert à mesurer le degré d'inégalité ou de concentration d'une distribution. Il est souvent utilisé pour analyser la répartition d'une variable (par exemple les revenus ou la population). L'indice se mesure entre 0 et 1, plus l'indice est élevé (proche de 1), plus la distribution est concentrée, et plus les inégalités sont fortes. Son intérêt est de donner un indicateur synthétique qui résume la manière dont un caractère (revenus, surfaces, ressources…) est réparti..Ainsi, la médiale et l'indice de Gini permettent d'aller au-delà des mesures de position en apportant une information sur la dispersion et la concentration des données.

Paramètre de dispersion

On ne peut pas utiliser directement l'écart à la moyenne car, lorsqu'on additionne tous les écarts, le résultat est toujours égal à zéro : les valeurs au-dessus et en dessous de la moyenne s'annulent. Pour éviter ce problème, on calcule la variance en élevant les écarts à la moyenne au carré, ce qui permet de mesurer correctement la dispersion des données. Cependant, la variance est exprimée dans une unité au carré, ce qui la rend difficile à interpréter. C'est pourquoi on utilise généralement l'écart type, qui correspond à la racine carrée de la variance. L'écart type est plus parlant, car il s'exprime dans la même unité que la variable étudiée. Ducoup, la variance sert surtout au calcul, tandis que l'écart type est plus simple à comprendre et à utiliser pour décrire la dispersion des données.

On calcule l'étendue car il s'agit d'une mesure simple et facile à calculer pour avoir une première idée de la dispersion des données. Elle correspond à la différence entre la valeur maximale et la valeur minimale d'une série statistique. Cela permet de voir rapidement sur quelle plage les valeurs sont réparties.

Créer un quantile permet de découper une série statistique ordonnée en plusieurs parts contenant le même nombre d'observations. Cela sert à mieux comprendre la répartition des données et à repérer leur position dans la distribution. Les quantiles sont utiles pour décrire la structure d'une série sans être trop influencés par les valeurs extrêmes. Les quantiles les plus utilisées sont les quartiles: le premier quartile (Q1), la médiane (Q2) et le troisième quartile (Q3) sont très courants. Ils permettent de décrire la position centrale des données et leur dispersion.

On construit une boîte de dispersion (boite à moustache) pour résumer une série statistique de manière simple et visuelle. Elle permet de décrire une distribution statistique

en montrant son étendue statistique, les valeurs min et max,la médiane (quartile 2), le premier (Q1) et le dernier (Q3) quartile. C'est un outil statistique descriptif pratique pour comparer la distribution statistique de différentes variables. Pour l'interpréter : la boîte (le rectangle) représente l'intervalle entre le premier (Q_1) et le troisième (Q_3) quartile. Le trait à l'intérieur de la boîte correspond à la médiane. Les moustaches indiquent les valeurs minimale et maximale. Plus la boîte est grande, plus la dispersion est forte. La position de la médiane dans la boîte permet aussi de repérer une éventuelle dissymétrie de la distribution.

Paramètre de forme

- Les moments permettent de caractériser une distribution. La différence entre les moments centrés et les moments absolus est la référence sur lequel se base leur calcul. On va avoir les moments absolus (*k)* calculés directement à partir des valeurs de la variable. Ils dépendent donc de l'origine choisie et servent surtout à décrire la position générale des données. Et les moments centrés *(r)* qui sont calculés à partir des écarts des valeurs par rapport à la moyenne. Les moments centrés ne dépendent donc plus de l'origine des variables et permettent de mieux décrire la dispersion et la forme de la distribution, comme la variance pour la dispersion ou l'asymétrie et l'aplatissement pour la forme. On les utilise parce qu'ils apportent des informations différentes mais complémentaires : les moments absolus donnent une idée du niveau des données, tandis que les moments centrés permettent d'analyser comment les valeurs sont réparties autour de la moyenne.

On vérifie la symétrie d'une distribution pour comprendre comment les valeurs se répartissent. Dans une distribution symétrique, le mode, la médiane et la moyenne arithmétique sont égaux, ce qui rend l'interprétation des données plus facile et permet de comparer la distribution à des modèles comme la loi normale. Pour mesurer cette symétrie, on utilise le coefficient d'asymétrie β_1, basé sur le moment centré d'ordre 3 et l'écart-type :

- Si $\beta_1 = 0$, la distribution est symétrique.
- Si $\beta_1 > 0$, elle est étalée vers la droite (asymétrie positive).
- Si $\beta_1 < 0$, elle est étalée vers la gauche (asymétrie négative).

En résumé, calculer β_1 permet de **quantifier la dissymétrie** et de mieux comprendre la forme de la distribution avant d'appliquer d'autres analyses statistiques.

Pour choisir entre une distribution statistique avec des variables continues ou une distribution avec des variables discrètes, le critère principal est la nature des variables, du phénomène étudié.Si la variable correspond à un dénombrement ou à des valeurs bien séparées, comme un nombre d'individus, de défauts, de succès ou d'événements, alors une distribution discrète est plus adaptée. Dans ce cas, les valeurs possibles sont limitées et clairement identifiables, souvent des entiers. La distribution discrète permet donc de représenter fidèlement ce type de phénomène. En revanche, si la variable peut prendre n'importe quelle valeur dans un intervalle, comme une durée, une taille, un poids ou une température, il est plus logique d'utiliser une distribution continue. Ici, la précision de mesure joue un rôle important : même si l'on arrondit les données, le phénomène sous-jacent reste continu. Je tiendrais aussi compte du contexte du phénomène étudié est essentiel. Certains modèles théoriques ou pratiques sont naturellement associés à des lois discrètes ou continues. Le bon choix est donc celui qui décrit le plus simplement et le plus réalistement la situation observée.

Certaines lois sont, en géographie, plus utilisées que d'autres, car elles permettent de décrire des phénomènes très fréquents dans l'espace ou dans les populations. Concernant les variables discrètes, on retrouve notamment les lois de Bernoulli, binomiale et de Poisson. La loi Bernoulli permet de démontrer sa capacité lorsque l'on étudie un phénomène à deux modalités, comme la présence et l'absence. La loi binomiale trouve sa pertinence lorsque l'on répète plusieurs fois ce type d'observation. La loi de Poisson permet quant à elle de modéliser des événements rares mais comptables.

En ce qui concerne les variables continues, les lois les plus courantes sont la loi uniforme, l'exponentielle, la gamma, la loi log-normal et surtout la loi normale. La loi uniforme permet de décrire des variables continues ayant une probabilité constante. La loi exponentielle est utilisée afin de modéliser la durée de vie, ce qui est utile dans l'analyse de certains processus naturels. Les lois gamma et loi-normale sont adaptées lorsque les valeurs sont très asymétriques, comme certaines intensités climatiques. Pour finir, la loi normale donne la probabilité qu'une variable aléatoire suivant une distribution normale soit inférieure ou égale à une valeur donnée. Elle constitue la loi la plus utilisée en géographie, ce qui la rend indispensable.

Un échantillon correspond à un sous-ensemble de la population totale, choisi pour la représenter lorsqu'il est impossible, trop long ou trop coûteux d'observer chaque individu. Cela permet par exemple d'estimer un phénomène sans devoir mesurer toute la population.
On distingue deux grandes familles de méthodes d'échantillonnage. On peut tout d'abord noter les méthodes probabilistes qui regroupent l'aléatoire simple, le tirage systématique, l'échantillonnage en grappe ou encore le stratifié. Les méthodes non probabilistes reposent au contraire sur des choix guidés par des critères pratiques, comme les quotas, la convenance ou la méthode dite "boule de neige".
Le choix des méthodes dépend de plusieurs éléments, à savoir l'objectif de l'étude, la structure de la population et les contraintes matérielles ou organisationnelles.
Un estimateur se définit comme étant une statistique utilisée pour approcher un paramètre inconnu de la population. Il est calculé à partir d'un échantillon, avec l'idée que cette valeur soit la meilleure approximation possible de ce que l'on obtiendrait si l'on observait toute la population.
L'estimation correspond, elle, à la valeur numérique prise par l'estimateur une fois calculé sur un échantillon. Autrement dit, c'est la valeur réellement observée dans les données.
L'intervalle de fluctuation est lié à la variabilité d'une statistique lorsqu'on répète différents échantillonnages. Il constitue un outil important en statistique, car il permet l'aide à la décision avec un risque d'erreur contrôlé. L'intervalle de confiance, lui, vise à encadrer la valeur du paramètre qui est utilisée pour estimer un inconnu de la population avec un niveau de confiance fixé à l'avance. Contrairement à l'intervalle de fluctuation, l'intervalle de confiance fournit un degré de certitude sur la précision de l'estimation.

En statistique, un estimateur est une fonction de données qui permet d'estimer un paramètre de la population. On dit qu'il est biaisé lorsque son espérance ne correspond pas à la valeur réelle du paramètre qu'il cherche à approcher.
Le recensement est le nom de statistique que l'on utilise lorsque l'on travaille sur la population entière. Elle repose sur une opération du dénombrement de la population

étudiée. Les données massives permettent d'approcher la population totale sans échantillonnage classique.
Le choix d'un estimateur présente plusieurs enjeux. En effet, il repose sur le compromis entre le biais et la variance. Idéalement, un bon estimateur doit alors être consistant, efficace et, si possible, non biaisé.
On distingue plusieurs méthodes d'estimation d'un paramètre, à savoir :

- la méthode des moments, qui consiste à estimer les paramètres en égalisant certains moments théoriques ;
- le maximum de vraisemblance, qui cherche les valeurs des paramètres rendant l'observation de l'échantillon plus probable ;
- la méthode des moindres carrés, utilisée pour déterminer la droite ou la courbe qui s'ajuste le mieux à un ensemble de points.

Le choix entre ces méthodes dépend des propriétés que l'on souhaite obtenir, comme le biais, la variance, la simplicité d'application ou encore les hypothèses nécessaires.
On retient plusieurs types de tests en statistique, à savoir les tests paramétriques, comme le test t, le test Z, ou l'ANOVA, ainsi que les tests non paramétriques, tels que le test de Mann-Whitney, pour n'en citer qu'un.
Ces tests permettent de déterminer si une hypothèse statistique peut être rejetée ou non.
Pour construire un test, il faut définir les hypothèses H0 et H1, choisir une statistique de test, fixer une région critique et un risque a.
L'interférence est critiquée pour la dépendance aux conditions d'échantillonnage, aux hypothèses fortes et à l'interprétation des p-values.

Une statistique ordinale (ou d'ordre) est une statistique qui peut être ordonnée.Elle consiste à ordonner/hiérarchiser des objets géographiques selon un critère donné. La statistique ordinal s'oppose à une autre statistique catégorielle, la statistique nominale dont les valeurs ne peuvent être ordonnées. Les statistiques ordinales utilisent des variables ordinales c'est-à-dire des variables qualitatives particulières pour lesquelles on ne mesure/quantifie pas de valeur numérique mais qu'on leur attribue un rang. Cela peut matérialiser une hiérarchie spatiale, on va établir des rapports de dominations, de centralité ou d'importance dans l'espace; entre objets géographiques (par exemple).

L'ordre à privilégier dans les classifications est l'ordre croissant aussi dit ordre naturel (de la plus petite valeur à la plus grande), le classement par ordre croissant permet de mettre en avant les valeurs aberrantes (trop petites ou trop grandes) d'une distribution statistique, d'une série d'observation.

Lorsqu'on établit un classement il y a une certaine subjectivité de celui qui établit le classement. On va donc mettre en place différent moyen d'analyse statistique pour divers classement possible. Et bien que ces deux méthodes cherchent à comparer des classements, il faut faire la différence entre la corrélation de rang et la concordance des classements. La corrélation de rang consiste à établir le degré de relation entre deux classement (d'un même objet d'étude). Pour chacun de ces objets, on regarde **le rang** qu'il occupe dans chaque classement. La corrélation des rangs consiste alors à mesurer si les objets placés dans le premier classement sont aussi bien placés dans le second, et inversement. Cela correspond au test de Spearman et de kendall qui vont être utilisés pour comparer ces deux variables ordinales Le résultat est un coefficient compris entre −1 et +1, qui exprime une intensité et un sens de la relation (positive, négative ou indépendante). Tandis que la concordance des classement en revanche, s'intéresse plus directement à la ressemblance effective entre les classements. Elle repose sur le comptage des paires

concordantes et discordantes : on regarde, deux à deux, si les objets sont dans le même ordre ou dans un ordre inversé selon les classements. L'objectif est de savoir si les classements sont globalement cohérents entre eux. Cette logique est centrale dans le coefficient τ de Kendall, dans le coefficient W pour plusieurs classements, ou encore dans le coefficient Γ de Goodman-Kruskal

Les tests de Spearman et de Kendall servent à comparer des classements et à mesurer la corrélation des rangs mais ne le font pas de la même manière. Le test de spearman compare directement les rangs obtenus dans deux classement (d'un même objet d'étude). Il regarde, pour chaque objet, l'écart entre son rang dans le premier classement et son rang dans le second. Plus ces écarts sont faibles, plus les classements se ressemblent. Le coefficient de Spearman mesure donc une relation globale entre les rangs. Le test de Kendall, lui, adopte une logique différente. Il ne compare pas directement les rangs, mais les paires d'objets. Il examine si deux objets sont dans le même ordre ou dans un ordre inversé d'un classement à l'autre. Il compte alors les paires concordantes et discordantes, puis mesure le degré d'accord entre les classements. Le coefficient de Kendall exprime donc une concordance, c'est-à-dire un niveau d'accord réel entre les ordres.

Les coefficients de Goodman-Kruskal et de Yule servent à mesurer l'association entre des variables catégorielles ordinales, en se fondant sur la comparaison des paires concordantes et discordantes. Le coefficient de Goodman-Kruskal est utilisé lorsque l'on compare deux variables ordinales ou deux classements. Il mesure le déséquilibre entre le nombre de paires concordantes et le nombre de paires discordantes. Autrement dit, il indique s'il existe un surplus de concordances par rapport aux discordances.

Γ varie entre −1 et +1 :

- Γ proche de +1 : association positive forte (les ordres vont dans le même sens),
- Γ proche de −1 : association négative forte (ordres inverses),
- Γ proche de 0 : absence d'association nette.

Il sert donc à apprécier la force et le sens d'une association ordinale, sans supposer de relation linéaire ni utiliser les valeurs numériques des rangs.

Le coefficient de Yule est un cas particulier du coefficient de Goodman-Kruskal. Il s'applique uniquement aux tableaux de contingence 2 × 2, c'est-à-dire lorsque l'on compare deux variables binaires (oui/non, présent/absent, positif/négatif). Il mesure également l'association entre les deux variables à partir des fréquences observées et s'interprète de la même manière :

- Q = +1 : association positive parfaite,
- Q = −1 : association négative parfaite,
- Q = 0 : absence d'association.

<u>Partie humanité numérique</u> :

La lecture de ce document m'a permis de me centrer davantage sur cette notion d'humanité numérique. Ces humanités numériques ne sont pas seulement des outils. En effet, elles ne se limitent pas à l'usage de logiciels ou de techniques informatiques. De plus, le numérique est à la fois un outil de recherche, une méthode et un objet d'étude, qui transforme les manières de produire, d'analyser et de diffuser les savoirs.

Ces humanités numériques visent à totalement changer la séparation traditionnelle entre les disciplines des sciences, techniques et humanités en montrant l'historicité de cette séparation et en favorisant les approches transdisciplinaires et collaboratives.

On note également l’importance centrale des données dans ces humanités numériques (bases de données, métadonnées, corpus numériques), permettant de travailler sur des volumes de données inaccessibles à la base, et renforçant la dimension quantitative dans les sciences humaines.
Elles permettent ainsi un réel renouvellement des méthodes en sciences humaines, en offrant de nouvelles possibilités d’analyse, un développement de méthodes hybrides mêlant quantitatif et qualitatif.
On note également des enjeux critiques et éthiques importants. En effet, certaines dépendances aux outils numériques peuvent être observées. De plus, les algorithmes ne sont pas neutres et peuvent introduire des biais (le chercheur doit conserver un regard critique sur les outils utilisés).
Un autre défaut à noter est celui de la fragilité de cette institutionnalisation. En effet, elles sont encore difficiles à définir comme discipline autonome. De plus, s’ajoute à cela l'ambiguïté de leur caractère transversal, étant à la fois une force et une faiblesse.
Les humanités numériques participent à une réflexion critique sur la société numérique, en interrogeant la production, la diffusion et l’accès aux savoirs. Pour finir, elles contribuent à repenser la place des humanités dans le monde actuel.