Côme DUBOIS **2025-2026**
M1 Geoint

Parcours débutant

Remarques générales

1. Des manipulations parfois trop complexes pour un niveau débutant

Lorsqu'on choisit le parcours « débutants » et qu'on commence à la première séance, on s'attend à un cours d'initiation. Or, une partie du cours suppose d'avoir réalisé une première prise en main des principes et principaux outils de la programmation, ce qui nécessite à ce qu'il me semble plusieurs heures d'apprentissage préalable. De plus, une partie des données à traiter relève de notions mathématiques et statistiques que des étudiants issus d'une formation de licence où les mathématiques ont été malheureusement totalement absentes ignorent, en particulier dans le cas des étudiants recrutés après trois années de classes préparatoires littéraires (A/L).

2. Usage et limites des assistants IA

Les assistants IA sont très utiles, voire indispensables pour les débutants (et paraît-il même aux professionnels), à condition de comprendre les scripts qu'ils nous proposent : il est le plus souvent nécessaire de modifier le code proposé pour l'adapter à nos données et aux consignes. Par conséquent, l'usage d'instructions pour forcer l'IA à expliquer son raisonnement et à détailler les différentes lignes de codes et fonctions utilisées est utile pour assimiler un minimum de notions et éviter les erreurs. Lorsqu'on rencontre des erreurs, il est également très pratique de copier le message reçu dans le terminal et le transmettre à l'assistant IA pour comprendre quoi corriger.

3. Une documentation trop abondante sur le Github

La surabondance de documentation (PDF, tutoriels, forums) sur le Github me paraît contre-productive, car choisir parmi les centaines de notices explicatives proposées est chronophage et devient vite un casse-tête. Une amélioration possible serait d'effectuer un tri pour n'en garder qu'une par thème ou notion afin que les étudiants s'y retrouvent et ne se dispersent pas. Le Github bien qu'organisé en séances apparaît de fait labyrinthique du fait du très grand nombre de dossiers, sous-dossiers et fichiers.

<u>Séance 2 :</u> Les principes généraux de la statistique

Questions de cours

1. <u>Quel est le positionnement de la géographie par rapport aux statistiques ?</u>

La discipline considère souvent que les statistiques n'entrent pas dans le champ qu'elle couvre ; or la recherche géographique produit des données massives qui ne sont exploitables que grâce à la statistique. Il existe différentes écoles géographiques dont le rapport à la discipline de la statistique varie. Les tenants de l'école de l'analyse spatiale considèrent que les statistiques sont indispensables à la conception de modèles spatiaux. Face à cette école surtout présente aux États-Unis, l'école vidalienne ou héritant de la géographie vidalienne française considère au contraire que la géographie est une méthode de raisonnement qui ne saurait fabriquer de modèles mathématiques car la contingence est au cœur de la discipline, ce qui la place non parmi les sciences « dures » mais bien parmi les sciences humaines et sociales.

2. <u>Le hasard existe-t-il en géographie ?</u>

Malgré l'impossibilité de réaliser des modèles mathématiques certains grâce aux statistiques, ces dernières tendent à dégager des tendances globales qui permettent d'améliorer la compréhension des phénomènes géographiques. Par conséquent, le hasard produit des effets à l'intérieur de ces tendances : il peut produire un écart au sein du modèle ou du scénario le plus probable.

3. <u>Quels sont les types d'information géographique ?</u>

Il en existe deux types : les éléments de géographie humaine ou physique caractérisant un territoire, ou la morphologie des territoires.

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

Le géographe doit étudier les informations géographiques contenues dans des jeux massifs de données fournis par des organismes publics ou des scientifiques effectuant un travail de mesure. Il modifie la nomenclature de ces données pour l'adapter au niveau de détail requis par l'échelle de son étude. Il en étudie également les méta-données. La discipline géographique s'attache donc à l'analyse de la structure interne des données qui doit le renseigner sur les conditions de leur production et sur le phénomène étudié.

5. <u>Quelles sont les différences entre la statistique descriptive et la statistique explicative ?</u>

La statistique descriptive consiste à étudier des données pour en dégager des propriétés remarquables ou des paramètres caractéristiques par rapport à une distribution théorique connue, dans le but d'obtenir une image simplifiée et ordonnée de la réalité. Elle permet donc d'identifier des lois de probabilités associées aux valeurs des données, et de résumer les dimensions principales du phénomène étudié, notamment en les visualisant dans une classification.

La statistique explicative (ou mathématique) cherche à prédire des scénarios possibles en établissant des liens de cause à effet entre une variable à expliquer et une variable explicative.

6. <u>Quelles sont les types de visualisation de données en géographie ? Comment choisir celles-ci ?</u>

Il existe deux principaux types de visualisation des données en géographie : l'histogramme, pour visualiser des variables quantitatives continues, et les diagrammes en secteurs pour visualiser des variables qualitatives. D'autres types de visualisation secondaires peuvent être utilisés comme les polygones de fréquence, les courbes cumulatives etc. Le choix du type de visualisation dépend donc du type de variable à visualiser, mais également sur l'objectif que l'on se donne : résumer les données, comparer des catégories ou analyser une distribution.

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

Il existe des méthodes descriptives où il s'agit de visualiser et classer les données, et des méthodes explicatives où l'on cherche à relier deux variables afin d'expliquer l'une par l'autre.

8. Comment définiriez-vous : (a) population statistique ? (b) individu statistique ? (c) caractères statistiques ? (d) modalités statistiques ? Quels sont les types de caractères ? Existe-t-il une hiérarchie entre eux ?

a) Une population statistique correspond à un ensemble mathématique, comme le nombre d'habitant dans un territoire. En géographie, c'est l'ensemble des unités spatiales ou des attributs.

b) Un individu statistique est un élément de la population statistique ; dans le cas des données géographiques cet individu est localisable et fréquemment composé lui-même d'un ensemble d'éléments de niveau inférieur.

c) Les caractères statistiques sont les caractéristiques de l'individu pris parmi la population statistique étudiée.

d) Les modalités statistiques sont les valeurs prises par un caractère statistique. Elles sont incompatibles et exhaustives car il s'agit de caractériser l'appartenance ou la non-appartenance d'un individu à une modalité.

Il existe deux grands types de caractères (ou variables) : qualitatifs (éventualité non chiffrée échappant à la mesure et pour laquelle on ne peut établir que des fréquences sans pouvoir lui associer une loi de probabilité) et quantitatifs (éventualité chiffrée à partir de laquelle on peut calculer des paramètres et qu'on peut associer à des lois de probabilité). On distingue aussi les variables qualitatives nominales (qui décrivent des états et qualifient textuellement les données) des variables qualitatives ordinales (qui décrivent des relations pour ordonner et classer des données) ; et les variables quantitatives discrètes ou discontinues (qui décrivent des listes finies et isolées de valeurs pour *compter* les données) des variables quantitatives continues (qui décrivent des valeurs prises dans un intervalle pour *mesurer* quelque chose par les données).

9. Comment mesurer une amplitude et une densité ?

L'amplitude est la longueur $b - a$ avec a la valeur minimale de la classe et b la valeur maximale. La densité est le rapport entre l'effectif et l'amplitude de la classe décrivant une modalité.

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

Les formules de Sturges et de Yule donnent une valeur approximative du nombre de classes étudiées.

11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?

Un effectif associé à une valeur d'une variable correspond au nombre d'apparitions de cette variable dans la population.

La fréquence cumulée est la somme des effectifs associés aux valeurs du caractère inférieures ou égales au nombre de modalités.

Une distribution statistique correspond au nombre d'apparitions de chaque valeur dans un ensemble de données. Elle permet d'établir une loi de probabilité associée à cet ensemble.

Manipulations sur Python

Question 12 et 13 : Je n'ai pas réussi à créer de diagrammes, la documentation existante et les propositions de l'assistant IA proposant des scripts trop difficiles.

Séance 3 : Paramètres statistiques élémentaires

Questions de cours

1. Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ?

Le caractère quantitatif est le plus général puisque les caractères qualitatifs sont individuels et impossibles à mesurer, et peuvent concerner un nombre infini de caractères.

2. Que sont les caractères quantitatifs discrets et les caractères quantitatifs continus ?

Les caractères quantitatifs discrets décrivent des listes finies ou isolées de valeurs tandis que les caractères quantitatifs continus décrivent des valeurs prises au sein d'un intervalle. On les distingue par le fait que les premières peuvent être comptées, alors que les secondes ne peuvent être que mesurées.

3. Paramètres de position

- Pourquoi existe-t-il plusieurs types de moyenne ? Les différents types de moyenne correspondent aux différents types de variable à partir desquelles est calculée la moyenne.
- Pourquoi calculer une médiane ? On calcule une médiane afin de diviser la population statistique en deux sous-groupes de probabilité équipropable. Comme elle n'est pas influencée par les valeurs aberrantes et qu'elle est déterminée par le classement des valeurs, elle permet de résumer efficacement des distributions dissymétriques.
- Quand est-il possible de calculer un mode ? On peut calculer un mode lorsqu'une valeur a une fréquence maximale.

4. Paramètres de concentration : quel est l'intérêt de la médiale et de l'indice de C. Gini ?

La médiale, qui consiste à partager les valeurs globales en deux parties égales représentant chacune 50 % des valeurs globales, permet de mesurer la concentration de certaines valeurs. L'indice de Gini permet de visualiser la médiale et ses conséquences notamment en termes d'inégalité de richesses.

5. Paramètres de dispersion

- Pourquoi calculer une variance à la place de l'écart à la moyenne ? Pourquoi la remplacer par l'écart type ? On calcule une variance à la place de l'écart à la moyenne car elle est la meilleure caractéristique de dispersion du fait qu'elle garde les écarts positifs et permet donc de mesurer la dispersion. Il est plus pratique de la remplacer par l'écart-type car il permet d'obtenir une mesure de la dispersion dans les mêmes unités que les données.
- Pourquoi calculer l'étendue ? L'étendue, différence entre la plus grande valeur observée et la plus petite, permet de savoir facilement à quel point les données sont dispersées.
- À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ? Un quantile sert à partager une série statistique ordonnée en plusieurs parties égales. Le quantile le plus utilisé est le quartile qui partager la série en 4 parts égales.
- Pourquoi construire une boîte de dispersion ? Comment l'interpréter ? On construit une boîte de dispersion pour représenter graphiquement un caractère quantitatif afin de comparer visuellement plusieurs séries statistiques. Le rectangle s'étend du premier au troisième quantile et il contient un trait qui marque la médiane ; les segments tracés de part et d'autre du rectangle sont les « moustaches » qui vont de la valeur minimale et de la valeur maximale au rectangle.

6. Paramètres de forme

- Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ? Un moment absolu correspond à la moyenne, tandis qu'un moment centré correspond à la variance. On les utilise pour établir des lois de probabilité.

- Pourquoi vérifier la symétrie d’une distribution et comment faire ? On vérifie la symétrie d’une distribution pour évaluer la manière plus ou moins équilibrée dont les données sont réparties autour de la moyenne. Cela nous permet de savoir si la médiane est plus fiable que la moyenne en raison d’une distribution asymétrique qui s’explique par la présence d’une donnée extrême. On peut donc comparer la moyenne et la médiane pour vérifier la symétrie : si elles sont proches la distribution est probablement symétrique, si elles sont éloignées la distribution est probablement asymétrique.

Manipulations sur Python

Organigramme de la démarche suivie sur le fichier island-index :

DÉBUT
  │
  ▼
Lire le fichier CSV contenant les données des îles
  │
  ▼
Sélectionner la colonne "Surface (km2)"
  │
  ▼
Définir les bornes de catégories (intervalle en km²) :
[0–10],]10–25],]25–50], ..., ≥ 10 000
  │
  ▼
Appliquer pd.cut() pour assigner chaque île à une tranche
  │
  ▼
Compter le nombre d’îles dans chaque catégorie avec value_counts()
  │
  ▼
Afficher ou sauvegarder les résultats (ex. : impression, CSV, graphique)
  │
  ▼
FIN

Séance 4 : Les distributions statistiques

Questions de cours

1. Quels critères mettre en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues ?

Le choix est lié à quatre critères principaux : la nature du phénomène étudié, la forme de la distribution empirique, la connaissance et à l'interprétation des principales caractéristiques de l'ensemble des données, le nombre de paramètres des lois.

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

La loi normale, car elle caractérise une variable aléatoire qui subit un grand nombre de facteurs indépendants et non synchrones, ce qui correspond notamment aux phénomènes naturels géographiques.

Manipulations sur Python

Bloqué à la première instruction concernant la visualisation d'une loi de Dirac, les forums d'usager Python et les assistants IA (chatGPT et Copilot VS Code) proposent tous un code très long et complexe que je ne comprends pas.

Séance 5 : Les statistiques inférentielles

Questions de cours

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

L'échantillonnage consiste à prélever dans une population totale (« mère ») une partie de celle-ci au hasard avec une taille *n* fixée ; c'est un sous-ensemble issu d'une variable aléatoire de cette population mère.

On ne peut utiliser la population en entier lorsque celle-ci est composée d'un très grand nombre d'individus.

Il existe deux méthodes d'échantillonnage. La méthode aléatoire procède par tirage au sort à partir d'une base de sondage (liste d'individus). La méthode non aléatoire tente de fabriquer un modèle réduit d'une population mère en utilisant d'autres procédés que le tirage au sort : échantillonnage systématique et technique des quotas (qui respecte la proportion d'éléments distinctifs de sa population totale).

2. Comment définir un estimateur et une estimation ?

L'estimateur est une fonction ou statistique qui intervient dans la théorie de l'estimation. Il est construit de façon à ce que sa valeur soit proche de la vraie valeur du paramètre de la loi de probabilité. Le but de la théorie de l'estimation est de choisir, parmi toutes les statistiques possibles, le meilleur estimateur, c'est-à-dire celui qui donnera une estimation la plus proche possible du paramètre quel que soit l'échantillon.

L'estimation consiste donc à donner une valeur approchée d'un paramètre à partir des résultats obtenus sur un échantillon aléatoire extrait de la population.

3. Comment distinguer l'intervalle de fluctuation et l'intervalle de confiance ?

L'intervalle de fluctuation décrit ce qu'on s'attend à observer par hasard dans un échantillon si l'hypothèse sur la population est vraie ; il sert donc à tester une hypothèse à partir d'un paramètre connu. L'intervalle de confiance estime quant à lui où se trouve le vrai paramètre de la population dans un échantillon, et sert donc à estimer un paramètre inconnu.

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

Un biais correspond à la différence entre l'espérance de l'estimateur et la valeur à estimer dans la population. Autrement dit, un biais mesure la différence systématique entre la valeur moyenne d'un estimateur, et la vraie valeur du paramètre que cet estimateur cherche à estimer.

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

Une statistique travaillant sur la population totale est un paramètre. A l'ère du Big Data, il devient possible de calculer ce paramètre directement, en raison de la masse de données disponibles qui peuvent désormais concerner l'ensemble de la population.

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

Le choix d'un estimateur est guidé par l'objectif d'obtenir une valeur qui soit la plus proche possible du paramètre réel afin de représenter au mieux la population et tirer les conclusions les plus fiables à partir de l'échantillon.

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

Il existe deux méthodes d'estimation d'un paramètre : la méthode des moments et la méthode du maximum de vraisemblance. Le choix de l'une ou l'autre repose sur la nature des données, la simplicité des calculs et la qualité statistique de l'estimateur en termes de biais, de précision et d'efficacité.

Manipulations sur Python

Théorie de l'échantillonnage :

Les fréquences de la population mère devraient, en théorie, tomber à l'intérieur des intervalles de fluctuation. Si ce n'est pas le cas, l'échantillon n'est probablement pas parfaitement représentatif de la population.

Théorie de l'estimation :

L'intervalle de confiance indique la zone où se situerait, avec 95 % de probabilité, la vraie fréquence de la population mère si on répétait l'expérience plusieurs fois. Plus la taille de l'échantillon est grande, plus l'intervalle est étroit. Si on compare les intervalles avec ceux de l'exercice précédent : les fréquences réelles de la population mère tombent à l'intérieur des intervalles, l'échantillon est donc représentatif.

Séance 6 : Statistiques d'ordre des variables qualitatives

Questions de cours

1. Qu'est-ce qu'une statistique ordinale ? A quelle autre statistique catégorielle s'oppose-t-elle ? Quel type de variables utilise-t-elle ? En quoi cela peut-il matérialiser une hiérarchie spatiale ?

Une statistique ordinale est une statistique qui se fonde sur le classement des observations et leur rang dans un ordre qui peut être croissant ou naturel. La statistique ordinale s'oppose à la statistique nominale qui traite des qualités ou catégories sans ordre particulier ou hiérarchie entre les valeurs possibles. Elle utilise des variables qualitatives ordinales, qui possèdent un ordre logique entre elles et que l'on peut classer du plus petit au plus grand sans pouvoir cependant calculer leur moyenne car ce ne sont pas des données numériques. Elles peuvent être un niveau de satisfaction, une taille de vêtement, une densité… Les statistiques ordinales peuvent donc matérialiser une hiérarchie spatiale comme celle des villes classées en fonction de leur population ou leur attractivité.

2. Quel ordre est à privilégier dans les classifications ?

Il faut privilégier l'ordre croissant (ou naturel), qui est la référence pour établir les statistiques d'ordre, car il permet de comparer correctement les rangs et d'identifier les valeurs aberrantes ou extrêmes.

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

La corrélation des rangs mesure le degré d'association (ou similarité) entre deux classements par la comparaison statistique de leurs rangs. Elle indique donc si deux variables ordinales évoluent dans le même sens. La concordance de classements permet quant à elle de savoir si plusieurs classements aboutissent ou non à un ordre similaire des objets étudiés. La différence entre les deux repose donc essentiellement sur le nombre de classements pris en considération dans la mesure de leur similarité.

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

Le test de Spearman évalue le degré de corrélation entre deux classements en mesurant l'écart quantitatif entre leurs rangs. Le test de Kendall compare quant à lui les classements en comptant les paires concordantes et discordantes, donc en mesurant un niveau d'accord basé sur la cohérence des ordres.

5. A quoi servent les coefficients de Goodman-Kruskal et de Yule ?

Ces deux coefficients servent à mesurer l'intensité de l'association entre deux variables catégorielles ordonnées. Le coefficient de Goodman-Kruskal indique si deux classements évoluent dans le même sens ou en sens inverse. Le coefficient de Yule sert à savoir si la relation entre deux modalités est positive, nulle ou négative.