RIBOULET Anna
21227318

19/12/2025

M1 GAED SCT - Analyse de données

Rapport d'activité, parcours débutant

Séance 2. Les principes généraux de la statistique

1. Quel est le positionnement de la géographie par rapport aux statistiques?

La relation de la géographie avec les statistiques est complexe. Historiquement, elle a souvent négligé les outils mathématiques, se considérant davantage comme une science humaine interprétative, jusqu'à la Nouvelle Géographie, un courant très concentré sur les aspects quantitatifs des phénomènes spatiaux. Dès lors, la géographie produisant des données en grandes quantités, il convient de maîtriser certains outils statistiques afin d'en réaliser l'analyse. La géographie, à l'aide des statistiques, permet de réaliser des analyses scientifiques multiscalaires en dégageant des lois et des grandes tendances de phénomènes locaux.

2. Le hasard existe-t-il en géographie?

Il existe deux positions philosophiques quant au hasard : la position déterministe affirme que le hasard n'existe pas, puisque tout phénomène a une cause identifiable. La position probabiliste affirme au contraire que le hasard traduit une ignorance des causes, mais reste un phénomène existant et explicable. Ainsi, en géographie, les phénomènes individuels sont impossibles à anticiper, mais on peut dégager des tendances collectives probables, par exemple dans le cas de l'étalement urbain. Il est dès lors possible de mesurer et modéliser le hasard.

3. Quels sont les types d'information géographique.

Il existe deux types d'information géographiques : les informations attributaires décrivent les caractéristiques d'un territoire (population, météo). Quant aux informations géométriques, elles concernent la forme et la structure des objets géographiques . C'est sur ces deux dimensions que s'appuie un Système d'Information Géographique (SIG) où on réalise des représentations graphiques de phénomènes (informations attributaires) par des polygones, des points, des lignes (informations géométriques)

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

La géographie a pour objectif d'observer des phénomènes et de structurer ces observations à l'aide de nomenclatures/métadonnées pertinentes. Les structures internes de ces données sont analysées par le biais d'outils statistiques qui permettent, à la fin, de visualiser ces structures par des représentations graphiques (cartes, diagrammes…). Ainsi, par l'analyse de données statistiques, on passe de la description à la compréhension des phénomènes spatiaux.

5. Quelles sont les différences entre la statistique descriptive et la statistique explicative?

La statistique descriptive et la statistique explicative constituent deux étapes complémentaires dans l'analyse des données géographiques.
La statistique descriptive vise avant tout à résumer, ordonner et représenter les données observées. Elle permet de mettre en évidence les grandes tendances d'un phénomène sans

chercher à en identifier les causes. Ses outils principaux sont les moyennes, médianes, écarts types, diagrammes, histogrammes ou encore les cartes thématiques.
En géographie, elle sert par exemple à décrire la population d'un territoire, à mesurer les densités, ou à cartographier la répartition des activités. En revanche, la statistique explicative (ou inférentielle) cherche à comprendre, modéliser et prévoir les phénomènes observés. Elle ne se limite pas à décrire : elle établit des relations de causalité entre variables — par exemple entre la répartition de la population et l'accès à l'emploi. Ses méthodes incluent les analyses de corrélation, les régressions simples ou multiples, les analyses factorielles ou les tests statistiques. En géographie, elle permet par exemple d'expliquer les causes de l'exode rural ou d'anticiper les dynamiques urbaines à partir de modèles probabilistes.

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

En géographie, on le choix d'une visualisation ou d'une représentation graphique des données dépend du type de variable et de l'analyse (comparaison, classement, corrélation…) qu'on souhaite réaliser :

- Variables catégorielles ou qualitatives : diagramme en barres ou en secteurs
- Variables quantitatives continues : histogramme ou boîte à moustache

- Corrélations et densités : nuages de points ou cartes de chaleur
- Répartition spatiale : cartes thématiques

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

On considère trois grands types de méthodes d'analyse de données :
- Descriptives : ACP, AFC, ACM, CAH qui permettent de visualiser les données.
- Explicatives : régression simple/multiple, régression logistique, analyse discriminante qui permettent d'établir des corrélation, des liens de causalité.
- De prévision : séries chronologiques, modèles temporels.

Par l'analyse de données, on passe d'une série statistique à un modèle spatial compréhensible et qu'il est possible d'analyser.

8. Comment définiriez-vous :

(a) Une **population statistique** est un ensemble au sens mathématique : il s'agit d'une population finie d'objets sur laquelle peut porter une étude statistique.

(b) Un **individu statistique** est un élément individuel considéré au sein d'une analyse statistique.

(c) Les **caractères statistiques**, qu'on appelle aussi variables statistiques, correspondent à ce qui est observé ou mesuré sur les individus d'une population.

(d) Une **modalité statistique** est l'une des variantes qu'un élément d'une population peut présenter dans le cadre d'une variable (le mois de naissance par exemple).

Quels sont les types de caractères? Existe-t-il une hiérarchie entre eux?

Un caractère peut être qualitatif (département de naissance d'un individu) nominal ou ordinal ou quantitatif discret ou continu (taille, poids…).

9. Comment mesurer une amplitude et une densité?

L'amplitude (A) correspond à l'étendue d'une classe statistique où a = borne inférieure et b = borne supérieure. La densité (d) établit un rapport entre l'effectif et l'amplitude. Elle permet d'établir des comparaisons des effectifs entre classes d'amplitudes différentes. Pour les mesurer, il faut d'abord discrétiser les données, donc construire des classes statistiques. On calcule l'amplitude b - a et la densité par le rapport de l'effectif à l'amplitude tel que $d = \frac{n_i}{b-a}$.

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

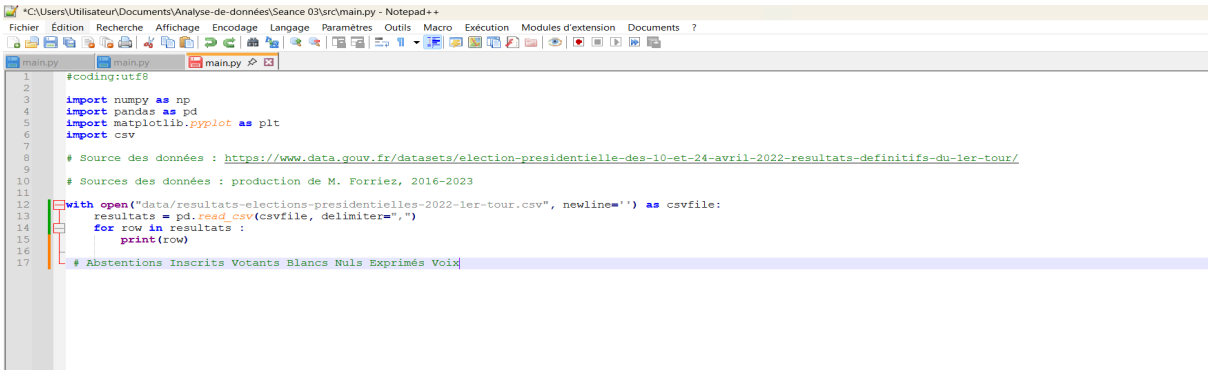
Elles servent à déterminer le nombre optimal de classes (K) pour produire un histogramme, afin de ne pas réaliser un découpage trop grossier qui aboutit à une perte d'informations, ou trop fin et impossible à structurer.

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

L'effectif (n_i) correspond au nombre d'occurrences d'une modalité. Quant à la fréquence (f_i) on la définit comme une part relative de cette modalité dans la population totale. Pour l'obtenir, on divise chaque effectif par l'effectif total avant de multiplier le résultat par 100. La fréquence cumulée est la somme progressive des fréquences jusqu'à une modalité donnée. On l'obtient en divisant la fréquence de chaque intervalle par le nombre total d'observation.
La distribution statistique correspond à la répartition des effectifs (ou fréquences) selon les modalités d'un caractère. Elle permet d'identifier les tendances globales et de relier la statistique descriptive à la probabilité théorique

Séance 3. Les paramètres statistiques élémentaires

Code



J'ai écrit ce code destiné à ouvrir les données contenues dans le fichier csv et à les print dans le terminal. J'ai rencontré beaucoup de problèmes à cette étape notamment avec les indentations :


C:\Windows\System32\cmd.e
time="2025-12-15T15:44:40+01:00" level=warning msg="Found orphan containers ([seance03-python-run-e951c814214d seance03-python-run-76631f67f13a seance03-python-run-f2c6f1ad215f seance03-python-run-e74bfdc70186]) for this project. If you removed or renamed this service in your compose file, you can run this command with the --remove-orphans flag to clean it up."
File "main.py", line 15
print(row)
^
IndentationError: expected an indented block
C:\Users\Utilisateur\Documents\Analyse-de-données\Seance 03> docker-compose run python
time="2025-12-15T15:44:56+01:00" level=warning msg="C:\\Users\\Utilisateur\\Documents\\Analyse-de-données\\Seance 03\\docker-compose.yml: the attribute `version` is obsolete, it will be ignored, please remove it to avoid potential confusion"
time="2025-12-15T15:44:56+01:00" level=warning msg="Found orphan containers ([seance03-python-run-ba35715f429e seance03-python-run-e951c814214d seance03-python-run-76631f67f13a seance03-python-run-f2c6f1ad215f seance03-python-run-e74bfdc70186]) for this project. If you removed or renamed this service in your compose file, you can run this command with the --remove-orphans flag to clean it up."
File "main.py", line 15
print(row)
^
IndentationError: expected an indented block
C:\Users\Utilisateur\Documents\Analyse-de-données\Seance 03> docker-compose run python
time="2025-12-15T15:45:36+01:00" level=warning msg="C:\\Users\\Utilisateur\\Documents\\Analyse-de-données\\Seance 03\\docker-compose.yml: the attribute `version` is obsolete, it will be ignored, please remove it to avoid potential confusion"
time="2025-12-15T15:45:36+01:00" level=warning msg="Found orphan containers ([seance03-python-run-48810e589468 seance03-python-run-ba35715f429e seance03-python-run-e951c814214d seance03-python-run-76631f67f13a seance03-python-run-f2c6f1ad215f seance03-python-run-e74bfdc70186]) for this project. If you removed or renamed this service in your compose file, you can run this command with the --remove-orphans flag to clean it up."


C:\Users\Utilisateur\Documents\Analyse-de-données\Seance 03> docker-compose run python
time="2025-12-15T16:49:50+01:00" level=warning msg="C:\\Users\\Utilisateur\\Documents\\Analyse-de-données\\Seance 03\\docker-compose.yml: the attribute `version` is obsolete, it will be ignored, please remove it to avoid potential confusion"
time="2025-12-15T16:49:50+01:00" level=warning msg="Found orphan containers ([seance03-python-run-b5e42ee08f53 seance03-python-run-d1fa28ed69d7 seance03-python-run-7db8aa4ce6c9 seance03-python-run-40c19a65b5c6 seance03-python-run-dbf9910ce821 seance03-python-run-e39fde0cddab seance03-python-run-300818546bfa seance03-python-run-161027da57ac seance03-python-run-014480c135c9 seance03-python-run-3f00957a26a8 seance03-python-run-4cf8f457177c seance03-python-run-f68f96c396b2 seance03-python-run-3bb0c127dc19 seance03-python-run-47484507cfc seance03-python-run-c9e675fff2d3 seance03-python-run-606952ccf565 seance03-python-run-0e1fc70d0c50 seance03-python-run-48810e589468 seance03-python-run-ba35715f429e seance03-python-run-e951c814214d seance03-python-run-76631f67f13a seance03-python-run-f2c6f1ad215f seance03-python-run-e74bfdc70186]) for this project. If you removed or renamed this service in your compose file, you can run this command with the --remove-orphans flag to clean it up."
Code du département
Libellé du département
Inscrits
Abstentions
Votants
Blancs
Nuls
Exprimés
Sexe
Nom
Prénom
Voix
Sexe.1
Nom.1
Prénom.1
Voix.1
Sexe.2
Nom.2
Prénom.2
Voix.2

```
C:\Users\Utilisateur\Documents\Analyse-de-données\Seance 03> docker-compose
run python
time="2025-12-15T17:22:35+01:00" level=warning msg="C:\\Users\\Utilisateur\\
Documents\\Analyse-de-données\\Seance 03\\docker-compose.yml: the attribute
`version` is obsolete, it will be ignored, please remove it to avoid potenti
al confusion"
time="2025-12-15T17:22:35+01:00" level=warning msg="Found orphan containers
([seance03-python-run-5606b2fee05d seance03-python-run-921f16b9a0a8 seance03
-python-run-b5e42ee08f53 seance03-python-run-d1fa28ed69d7 seance03-python-ru
n-7db8aa4ce6c9 seance03-python-run-40c19a65b5c6 seance03-python-run-dbf9910c
e821 seance03-python-run-e39fde0cddab seance03-python-run-300818546bfa seanc
e03-python-run-161027da57ac seance03-python-run-014480c135c9 seance03-python
-run-3f00957a26a8 seance03-python-run-4cf8f457177c seance03-python-run-f68f9
6c396b2 seance03-python-run-3bb0c127dc19 seance03-python-run-4748d4507cfc se
ance03-python-run-c9e675fff2d3 seance03-python-run-606952ccf565 seance03-pyt
hon-run-0e1fc70d0c50 seance03-python-run-48810e589468 seance03-python-run-ba
35715f429e seance03-python-run-e951c814214d seance03-python-run-76631f67f13a
 seance03-python-run-f2c6f1ad215f seance03-python-run-e74bfdc70186]) for thi
s project. If you removed or renamed this service in your compose file, you
can run this command with the --remove-orphans flag to clean it up."
0       97541.0
1      101089.0
2       58497.0
3       29290.0
4       25357.0
         ...
102    142121.0
103      2272.0
104      4125.0
105     15812.0
106    931455.0
Name: Abstentions, Length: 107, dtype: float64
```

Je n'ai réussi à isoler qu'une seule variable quantitative. Je n'ai pas trouvé comment en traiter plusieurs à la fois.

```
Traceback (most recent call last):
  File "/usr/local/lib/python3.6/site-packages/pandas/core/indexes/base.py",
 line 2898, in get_loc
    return self._engine.get_loc(casted_key)
  File "pandas/_libs/index.pyx", line 70, in pandas._libs.index.IndexEngine.
get_loc
  File "pandas/_libs/index.pyx", line 101, in pandas._libs.index.IndexEngine
.get_loc
  File "pandas/_libs/hashtable_class_helper.pxi", line 1675, in pandas._libs
.hashtable.PyObjectHashTable.get_item
  File "pandas/_libs/hashtable_class_helper.pxi", line 1683, in pandas._libs
.hashtable.PyObjectHashTable.get_item
KeyError: ('Abstentions', 'Votants')

The above exception was the direct cause of the following exception:

Traceback (most recent call last):
  File "main.py", line 14, in <module>
    print(resultats["Abstentions","Votants"])
  File "/usr/local/lib/python3.6/site-packages/pandas/core/frame.py", line 2
906, in __getitem__
    indexer = self.columns.get_loc(key)
  File "/usr/local/lib/python3.6/site-packages/pandas/core/indexes/base.py",
 line 2900, in get_loc
    raise KeyError(key) from err
KeyError: ('Abstentions', 'Votants')
```

Questions de cours

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

Le caractère le plus général semble être le caractère qualitatif, dans la mesure où le quantitatif permet de rentrer dans les détails des cas particuliers, que l'on peut mesurer de façon précise. Au contraire, le qualitatif permet de décrire ou définir des états, donc des qualités, au sens plus général.

2. Quels sont les caractères quantitatifs discrets et les caractères quantitatifs continus? Pourquoi les distinguer?

Un caractère quantitatif discret est une donnée qui peut prendre seulement certaines valeurs qui se situent toutes dans un intervalle donné. Au contraire, un caractère quantitatif continu est une donnée qui peut prendre toutes les valeurs dans cet intervalle. Par conséquent, il y a un nombre infini de possibilités de valeurs pour cet intervalle.

3. Paramètres de positions

- Pourquoi existe-t-il plusieurs types de moyennes?

En statistiques, il existe plusieurs types de moyennes car toutes les moyennes ne donnent pas la même importance à tous les types de valeurs, qui compteront de manière plus ou moins lourde dans le calcul. Il existe la moyenne arithmétique où l'on additionne les valeurs, la moyenne géométrique qui permet de calculer des phénomène multiplicatifs (des taux de croissance par exemple), la moyenne harmonique qui calcule des rapports entre des phénomènes, et enfin la moyenne pondérée qui est calculée en accordant une importance relative aux différentes valeurs. Ainsi, en fonction du phénomène que l'on cherche à observer, les données ne vont pas se combiner de la même façon : certains phénomènes ont des données qui s'additionnent, d'autres où on cherche à ce que la moyenne soit pondérée par une donnée spécifique…

- Pourquoi calculer une médiane?

La médiane est un quantile qui permet de diviser un ensemble en deux parties égales. Cela la rend simple à interpréter. Ce faisant, elle permet d'avoir une idée générale des valeurs moyennes d'un ensemble, puisqu'elle va être peu influencée par les valeurs extrêmes.

- Quand est-il possible de calculer un mode?

On peut calculer un mode, qui est la valeur la plus fréquente d'une série, pour tout type de données : cela permet de repérer la ou les valeurs dominantes. C'est une méthode qui se révèle cependant moins utile pour les données dispersées où il n'existe pas vraiment de valeur dominante. Il est impossible de calculer un mode pour des données continues, puisqu'une valeur ne se répète pas plusieurs fois (on peut alors regrouper les données pour fabriquer un mode).

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

La médiale permet de trouver un intermédiaire entre la moyenne, très sensible aux valeurs extrêmes, et la médiane, qui ne les prend pas en considération. Ainsi, elle utilise toutes les données sans pour autant être influencée de manière démesurée par les valeurs extrêmes. Quant à l'indice de Gini, c'est un indice dont les valeurs sont comprises entre 0 et 1 qui permet de mesurer les inégalités au sein d'une population. Ainsi, plus l'indice est élevé, plus il indique une inégalité importante. Au contraire, plus il est proche de zéro, moins les inégalités sont marquées au sein d'une population.

5. Paramètres de dispersion

- Pourquoi calculer une variance à la place de l'écart à la moyenne? Pourquoi la remplacer par l'écart type?

L'écart à la moyenne, la variance et l'écart type ont pour objectif de calculer la dispersion des données d'une série. Or, la somme des écarts à la moyenne est toujours égale à zéro, ce qui ne permet pas de mesurer une dispersion. Par ailleurs, la variance permet de mieux prendre en compte les valeurs extrêmes qui sont plus éloignées de la moyenne. On peut calculer l'écart type à partir de la racine carré de la variance, qui est plus pratique à utiliser parce qu'il s'exprime dans la même unité que les données de la moyenne.

- Pourquoi calculer l'étendue?

L'étendue se calcule par une soustraction l'écart entre la valeur minimale et la valeur maximale d'une distribution. La calculer permet d'avoir une indication sur la répartition des données. A partir de l'étendue, on peut calculer la concentration :

$$C = \frac{m_l - m_e}{\omega} = \frac{\Delta M}{\omega}$$

- A quoi sert-il de créer un quantile? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s)?

Les quantiles permettent de comprendre la répartition des données en les organisant. Ce faisant, on peut identifier les asymétries, les valeurs extrêmes, découper la distribution en parties égales… On peut ainsi analyser tout ensemble de données, même si il est irrégulier. Les quantiles les plus utilisés sont la médiane en 2 parties (surtout en statistiques), les quartiles en 4 parties (permettent d'analyser les dispersions, surtout en sciences sociales) et les déciles en 10 parties, que l'on mobilise surtout en économie.

- Pourquoi construire une boîte de dispersion? Comment l'interpréter?

Une boîte de dispersion, que l'on appelle aussi boîte à moustache, est un type de graphique qui permet de représenter visuellement une distribution (souvent en quartiles). Elle permet de communiquer visuellement la structure d'un ensemble en représentant les dispersions, si elle est plus ou moins large (données dispersées) ou étroites (données concentrées), les asymétries (selon

la position de la médiane), les valeurs extrêmes ou aberrantes puisqu'elles sont représentées par des points hors de la boîte de dispersion…

6. Paramètres de forme

- Quelle différence faites-vous entre les moments centrés et les moments absolus? Pourquoi les utiliser?

Les moments sont des indicateurs de la dispersion d'une variable. Le moment centré utilise l'écart à la moyenne pour calculer la dispersion. Ils décrivent la variabilité (la variance est un moment centré) ou encore l'asymétrie. Ils permettent d'analyser de manière détaillée la distribution des valeurs. Les moments absolus mesurent la dispersion sans utiliser le signe de l'écart à la moyenne et accordent plus d'importance à la taille de l'écart. En les calculant, on obtient donc l'écart moyen, de manière plus facile à interpréter.

- Pourquoi vérifier la symétrie d'une distribution et comment faire?

Pour vérifier la symétrie d'une distribution, il convient de calculer le coefficient de distribution. Si il est nul ou presque nul, on sait qu'il existe une symétrie de distribution par rapport à la moyenne. Si on se rend compte que la distribution est asymétrique, il peut être préférable de calculer la médiane en plus de la moyenne pour connaître le centre. Si la distribution est particulièrement asymétrique, l'écart-type et les comparaisons peuvent ne pas être très fiables pour interpréter l'ensemble.

Séance 4. Les distributions statistiques

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

Les variables discrètes sont dénombrables, quand les variables continues correspondent à des intervalles. Pour choisir entre l'une et l'autre des distributions statistiques, on prend en compte le degré de précision (si la valeur est exacte, on choisira une valeur discrète). Le choix est ainsi basé sur la possibilité d'obtenir des données les plus précises possibles avec l'instrument de mesure. On choisira une variable continue si on mesure une grandeur.

2. Expliquez quelles sont selon vous les lois les plus utilisées en géographie?

En géographie, on mobilise régulièrement les lois suivantes :

La loi de Gauss ou loi normale : elle permet de modéliser des phénomènes naturels issus de plusieurs événements aléatoires. Ainsi, elle paraît extrêmement utile non seulement en géographie environnementale et des risques (modéliser des risques naturels, météorologiques ou telluriques…) mais également en géographie humaine et sociale, dans l'objectif de comprendre les comportements individuels en apparence imprévisibles et aléatoires.

La loi de Zipf : cette loi statistique est particulièrement présente dans le cadre de la géographie urbaine. Elle permet d'appréhender la fréquence ou l'importance de certains phénomènes, même inégalement répartis.

Loi de Poisson : c'est une loi de probabilité discrète mobilisée notamment par la géographie des risques. Elle permet d'appréhender un nombre d'évènements se produisant dans un intervalle de temps fixé, comme des crues.

Séance 5. Les statistiques inférentielles

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

En statistiques, réaliser une étude sur une population entière peut s'avérer trop coûteux ou trop long/laborieux. On a alors recours à l'échantillonnage, qui est la sélection d'une partie d'une population que l'on souhaite étudier, afin d'en réaliser un modèle réduit. Un échantillon est dit représentatif lorsqu'il respecte la structure de la population dans son ensemble, et facilite la collecte de données. On peut choisir un échantillon selon des méthodes probabilistes ou aléatoires, et des méthodes non aléatoires. La méthode probabiliste peut s'avérer pertinente si on a une bonne connaissance de la population étudiée. Néanmoins, si celle-ci est constituée d'une variété de groupes différents, il peut être difficile d'obtenir un échantillon représentatif. Lorsque l'enquêteur mobilise des méthodes non aléatoires, il constitue l'échantillon par un quota ou un faux hasard. Lorsqu'il s'agit d'un échantillon aléatoire, on peut réaliser un tirage systématique, un tirage avec stratification ou un tirage aérolaire. Finalement, le choix d'une méthode aléatoire ou non repose sur le sujet de l'enquête et ses besoins en termes de représentativité. Ainsi, si on cherche à étudier un phénomène à travers l'axe de l'âge, avoir un échantillon représentatif des différentes catégories d'âge paraît le plus important.

2. Comment définir un estimateur et une estimation?

Un estimateur est un outil théorique utilisé pour approcher une valeur inconnue d'un paramètre d'une population. C'est une fonction des données qui se rapporte à une variable aléatoire. Quant à l'estimation, il s'agit de la valeur numérique que l'on obtient lorsqu'on calcule l'estimateur à partir de l'échantillon de population que l'on a choisi d'observer.

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

Un intervalle de fluctuation correspond à un intervalle dans lequel l'une des grandeurs observées est caractérisée par une très haute probabilité. Un intervalle de confiance intervient dans des études avec un échantillon, et permet de juger de la fiabilité de cet échantillon : le pourcentage affirme que l'étude a n% de chances de contenir les mêmes valeurs qu'une interrogation exhaustive, sur la population totale. Dès lors, un intervalle de fluctuation concerne plutôt les paris et les probabilités, alors qu'un intervalle de confiance permet de juger de la fiabilité d'un échantillonnage.

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

Parfois également qualifié d'erreur d'estimation, un biais correspond à l'écart entre l'espérance de l'enquêteur et la valeur réelle à estimer dans la population étudiée.

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 est calculée à partir d'un échantillon. Par opposition, un paramètre est une grandeur que l'on calcule sur l'ensemble de la population. Ainsi, une statistique qui travaille sur la population totale s'appelle un paramètre statistique. Les mégadonnées, ou données massives, correspondent à un ensemble de données vaste, varié et volumineux dont la taille augmente au fil du temps. On peut supposer qu'avec des capacités de stockage toujours plus importantes (bien que coûteuses pour l'environnement) on puisse réaliser de plus en plus d'études basées sur des paramètres statistiques et non sur des échantillons, grâce à un accès toujours plus important à des données en grand nombre.

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

Le choix d'un estimateur articule plusieurs paramètres : la précision du calcul, sa robustesse (selon la sensibilité aux valeurs extrêmes, par exemple), la cohérence avec le modèle statistique mobilisé, la taille de l'échantillon… Cela dépend de la variance de la série.

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

La méthode d'estimation des paramètres est choisie en fonction du modèle, de la taille de l'échantillon, de la robustesse souhaitée et des objectifs d'inférence. Il en existe de nombreuses, notamment l'estimation paramétrique, qui est une façon d'estimer un paramètre en supposant que les données suivent une loi appartenant à une famille paramétrée connue. Parmi la variété de méthodes d'estimation paramétriques, la plus utilisée est la méthode du maximum de vraisemblance (MV), qui donne généralement des résultats efficaces et cohérents en mettant en avant les données observées les plus probables. On utilise aussi la méthode d'estimation maximum à posteriori ou la méthode des moindres carrés. Dans les autres méthodes, on peut également mentionner l'estimation bayésienne, l'estimation régularisée, l'estimation robuste…

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

Il existe de nombreux types de tests statistiques. Les plus courants sont les tests qui s'appliquent à des moyennes ou des proportions : le test de Student (comparer une moyenne à une valeur ou une autre moyenne), le test Z qui concerne les grands échantillons et le test sur une proportion (qui compare une proportion observée à une proportion théorique. Il existe ensuite les tests de comparaison : le test ANOVA qui compare plusieurs moyennes, et les tests non paramétriques comme Mann-Whitney ou Kruskal-Wallis (entre autres). Les test d'association ou

d'indépendance qui font des liens entre deux variables qualitatives (test de khi-deux ou de Fisher pour les petits effectifs). Il existe ensuite les tests sur la distribution, soit de normalité soit d'homogénéité des variances. Enfin, les tests de corrélation permettent d'établir des relations linéaires entre les variables quantitatives normales ou des relations monotones.

Les tests servent à décider si les effets que l'on observe sont dûs ou non au hasard, permettent de tester des hypothèses, de comparer des situations ou d'appuyer des décisions scientifiques, bien qu'ils soient toujours incertains.

Pour construire un test statistique, on formule une hypothèse, on choisit une statistique de test et on détermine la loi statistique que l'on va appliquer (de Student, de Fisher…) ainsi qu'un seuil de risque. Enfin, on calcule la statistique à partir des données récoltées dans l'échantillon.

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

La statistique inférentielle est la méthode statistique qui permet de tirer des conclusions sur une population à partir d'un échantillon, en tenant compte des incertitudes de l'échantillonnage. Par conséquent, leur caractère clef est l'incertitude relative au processus d'échantillonnage, que l'on considère à l'aide de l'intervalle de confiance. Ainsi, la statistique inférentielle peut faire l'objet de critiques quant à l'illusion d'objectivité qu'elle peut donner, en ayant une vision trop mécaniste des tests de certitude. Selon la taille de l'échantillon, on peut avoir de fortes variations (disparition des particularités dans de vastes groupes, surinterprétation de phénomènes microscopiques dans de petits échantillons). Bien que ces critiques soient pertinentes, les alternatives me paraissent soit impossibles à mettre en place (réaliser une étude sur la totalité d'une population paraît bien trop long et coûteux) ou mauvaises : l'utilisation des mégadonnées, qui implique des dépenses énergétiques toujours plus importantes, paraît irresponsable en contexte de crise climatique et environnementales, autant qu'illusoire puisque non durable.

Séance 6. Les statistiques d'ordre des variables qualitatives

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

Une statistique ordinale concerne des données qualitatives ordinales. C'est-à-dire que les données sont catégorielles, donc des modalités non numériques, mais qu'elles peuvent tout de même être classées (ordonnées) selon un ordre logique (un niveau de satisfaction). Les données ordinales s'opposent aux données qualitatives nominales, c'est-à-dire sans ordre. On peut seulement les distinguer par classes ou catégories (la commune de naissance par exemple). Elles utilisent donc les variables qualitatives ordinales. On peut les utiliser pour matérialiser une hiérarchie spatiale, par exemple en faisant des catégories urbaines hiérarchiques (village/ville/métropole), des niveaux d'altitudes ou encore de vulnérabilité aux risques. On peut ainsi produire des classements ou des comparaisons des territoires que l'on peut représenter visuellement (graphiques, cartes).

2. Quel ordre est à privilégier dans les classifications?

Pour des raisons de compréhension, il paraît plus pertinent d'utiliser un ordre croissant, qui est plus intuitif et rend les lecture et les compréhensions plus simples. Dans ces ordres croissants existent des ordres fondés sur une hiérarchie spatiale ou fonctionnelle préétablie. On va parler de gradient, de hiérarchie, de niveau d'intensité ou de degré d'intégration.

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

Bien que ces deux notions cherchent à comparer des ordres, elles ne mesurent pas exactement la même chose au sein de ces comparaisons. La corrélation des rangs mesure le sens et l'intensité d'une relation monotone entre deux variables ordinales, quand la concordance des classement mesure le degré de similarité exacte entre deux classements de données. Ainsi, la concordance des classements semble plus exigeante et plus précise qu'une simple indication sur la relation entre deux variables.

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

Ce sont deux tests de corrélations non paramétriques qui tendent à mesurer la relation entre deux variables ordinales. Leur logique est différente : le test de Spearman mesure la force et le sens d'une relation monotone, et est utilisé pour calculer la corrélation des rangs. Le test de Kendall compare toutes les paires d'observations peu importe la nature de leur relation, et est basé sur la concordance des classements. Il fonctionne mieux pour les petites séries/petits échantillons, dans la mesure où il est plus précis.

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

Il s'agit de deux mesures d'associations qui permettent de travailler avec des variables qualitatives ordinales. Ils permettent de mesurer l'intensité de l'association entre deux variables. Ils servent à savoir dans quelle mesure connaître l'une des deux variables permet de mieux prévoir l'autre. On peut ainsi quantifier des relations non numériques, ce qui s'avère particulièrement utile dans le cas des sciences sociales (et donc de la géographie).

Réflexion personnelle sur les sciences et les humanités numériques en fonction des exercices de votre parcours.

Les humanités numériques correspondent à un domaine d'enseignement et de recherche au croisement de l'informatique et des lettres, des arts, des sciences humaines et des sciences sociales, visant à produire et à partager des savoirs, des méthodes et de nouveaux objets de connaissance à partir d'un corpus de données numériques, selon le Ministère de la Culture. Ainsi, intégrer des moyens de traiter et quantifier des données non numériques, comme on l'a vu précédemment, semble faire partie intégrante de la démarche des humanités numériques, qui paraît très pertinente. C'est selon moi une démarche dans laquelle la géographie s'inscrit depuis ses débuts, dans la mesure où elle a toujours été une discipline difficile à caractériser, qui tend à

la pluridisciplinarité et au décloisonnement des différentes disciplines universitaires. Les humanités numériques me paraissent donc un pas de plus dans l'avènement d'exigences pluridisciplinaires de plus en plus importantes dans le cadre de l'université.

Cependant, si ces attentes perdurent et se développe dans l'enseignement supérieure, il convient de préparer correctement les élèves du secondaire, où les enseignements semblent, au contraire, de plus en plus cloisonnés (avec la réforme des spécialités qui implique par exemple que de nombreux élèves ne suivent plus d'enseignement mathématique après la classe de Seconde). Dans le cas contraire, l'apparition des humanités numériques paraît être une barrière de plus mise en place pour opérer une sélection et rendre inaccessible l'enseignement supérieur.

J'ai aussi eu l'occasion de discuter avec un ami qui suit un cursus de programmation. Il m'a expliqué qu'il est plutôt réticent à se lancer dans une spécialisation d'analyse de données, dans la mesure où ce secteur lui paraît plutôt menacé par les récents développements de l'intelligence artificielle, qui peut selon lui assurer cette tâche plus vite et mieux.

Ainsi, si les humanités numériques me paraissent excessivement bénéfiques pour le décloisonnement du monde universitaire et l'enrichissement de la pensée au sens large, elles doivent être suivies de réglementations et d'adaptations de l'enseignement, afin de lui être réellement bénéfiques.