Cours d'analyse de données en géographie
Niveau Master 1 - GEANDO
Séance 2. Les principes généraux de la statistique

Maxime Forriez[1,a]

[1] Institut de géographie, 191, rue Saint-Jacques, Bureau 105, 75 005 Paris,
[a]maxime.forriez@sorbonne-universite.fr

20 septembre 2025

1 Questions de cours

Les réponses comptent pour 10 % de la note finale du parcours « débutants ».

1. Quel est le positionnement de la géographie par rapport aux statistiques ?
2. Le hasard existe-t-il en géographie ?
3. Quels sont les types d'information géographique .
4. Quels sont les besoins de la géographie au niveau de l'analyse de données ?
5. Quelles sont les différences entre la statistique descriptive et la statistique explicative ?
6. Quelles sont les types de visualisation de données en géographie ? Comment choisir celles-ci ?
7. Quelles sont les méthodes d'analyse de données possibles ?
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 ?
9. Comment mesurer une amplitude et une densité ?
10. À quoi servent les formules de Sturges et de Yule ?
11. Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?

2 Mise en œuvre avec Python

La sous-partie « Bonus » vous permet d'obtenir des points supplémentaires.

2.1 Objectifs

— Manipuler un fichier C.S.V.

— Faire des sorties graphiques

— Utiliser les bibliothèques `Pandas` (données) et `Matplotlib` (graphiques)

N.B. `pd` et `plt` sont des alias qui remplacent respectivement `pandas` et `matplotlib.pyplot`.

— Calculer des effectifs

— Calculer des fréquences

— Faire des graphiques (diagrammes en bâton et circulaires, et histogrammes)

2.2 Manipulations

Le fichier obtenu compte pour 10 % de la note finale du parcours « débutants ».

1. Dans le dossier `src`, créer un dossier `data` et y introduire le fichier `resultats-elections-presidentielles-2022-1er-tour.csv` disponible dans la `Seance-02` du `GitHub`
2. Dans le dossier `src`, introduire le fichier `main.py` de la séance disponible dans la `Seance-02` du `GitHub`
3. Ouvrir le fichier `main.py` dans votre éditeur de code (`Notepad++` ou `VS Code`)
4. Repérer l'instruction `with`. Elle permet par l'intermédiaire d'une variable appelée `fichier` de lire le fichier C.S.V. grâce à la méthode `read_csv(...)` de la bibliothèque `Pandas`.

   N.B. Bien que `Pandas` puisse lire les fichiers `Excel`, il faut vous habituer à utiliser des formats textuels comme le format C.S.V.
5. Afficher sur le terminal exécutant le conteneur la variable `contenu` avec la méthode de `Pandas` `DataFrame(...)`
6. Calculer avec la fonction native `len(...)` le nombre de lignes et de colonnes du tableau de données et les afficher sur le terminal
7. Faire le point sur la nature statistique des variables en utilisant le lien vers les métadonnées fourni en commentaire. Faire une liste sur le type de chaque colonne (`int`, `float`, `str` ou `bool`).
8. Afficher sur le terminal le nom des colonnes, c'est-à-dire la première ligne avec la méthode `Pandas` `head()`
9. À l'aide du nom des colonnes affiché, sélectionner le nombre des inscrits
10. Calculer avec la méthode `Pandas` `sum(...)` les effectifs de chaque colonne et les placer dans une liste (à l'aide d'une boucle). Afficher le résultat sur le terminal. Normalement, le résultat sera bizarre. Mettre une condition pour calculer les effectifs des colonnes contenant des données quantitatives en utilisant la liste faisant le point sur le type de variables.

    Indice. Utiliser la méthode native `append(...)`

11. À l'aide de `Matplolib` et d'une boucle, faire des diagrammes en barres avec le nombre des inscrits et des votants pour chaque département. Créer les fichiers images des diagrammes.

    N.B. 1. Vous allez créer de nombreux fichiers. Créer un nouveau sous-dossier pour stocker vos images.

    N.B. 2. Privilégier le format `*.png`, qui est un format léger non propriétaire

12. À l'aide `Matplolib` et d'une boucle, faire des diagrammes circulaire avec les votes blancs, nuls, exprimés et l'abstention pour chaque département. Créer les fichiers images des diagrammes. Respecter les mêmes remarques.
13. À l'aide `Matplolib`, faire l'histogramme de la distribution des inscrits

Je rappelle la différence entre un diagramme en bâtons et un histogramme. Le premier sert à représenter des données, le second, à représenter une distribution statistique. Cela signifie sur la totalité des surfaces des rectangles vaut 1 (cf. cours sur les distributions).

2.3 Bonus

Sans remarque pour vous aider (conditions réelles), faire les diagrammes circulaires visualisant, pour chaque département, le nombre de voix par candidat.

Calculer le diagramme circulaire pour l'ensemble de la France.