Cours d'analyse de données en géographie
Niveau Master 1 - GEANDO
Séance 5. Les statistiques inférentielles

Maxime Forriez[1,a]

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

22 septembre 2025

1 Questions de cours

Les réponses comptent pour 30 % de la note finale du parcours « débutants ».
Les réponses comptent pour 30 % de la note finale du parcours « intermédiaires ».
Les réponses comptent pour 10 % de la note finale du parcours « confirmés ».

1. Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier ? Quelles sont les méthodes d'échantillonnage ? Comment les choisir ?
2. Comment définir un estimateur et une estimation ?
3. Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?
4. Qu'est-ce qu'un biais dans la théorie de l'estimation ?
5. Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien avec la notion de données massives [1] ?
6. Quels sont les enjeux autour du choix d'un estimateur ?
7. Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?
8. Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?
9. Que pensez-vous des critiques de la statistique inférentielle ?

Attention, questions de cours dans l'exercice : surligné en jaune

2 Mise en œuvre avec Python

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

1. En anglais : *big data*

2.1 Objectifs

— Manipuler harmonieusement les fonctions natives avec les méthodes `Pandas`

— Comprendre les trois théories permettant de valider un résultat en analyse de données

2.2 Manipulations

Le fichier obtenu compte pour 40 % de la note finale du parcours « débutants ».
Le fichier obtenu compte pour 30 % de la note finale du parcours « intermédiaires ».
Le fichier obtenu compte pour 10 % de la note finale du parcours « confirmés ».

N.B. Il est très compliqué de trouver des données « réelles ». La plupart d'entre elles ne sont pas publiées. Il faut vous approprier les trois méthodes pour bien comprendre leur intérêt. De fait, les données utilisées sont obtenues par simulation aléatoire. Comme vous l'avez lu dans le topo de la séance, les méthodes sont quasiment infinies.

1. Théorie de l'échantillonnage

   Problème à résoudre. Vous réalisez une enquête d'opinion. On suppose que la population mère est composée de 2 185 individus. Si vous leur posez une question quelconque, nécessitant une opinion tranchant de type : « Pour », « Contre » ou « Sans opinion ». On suppose que, pour la population mère, il existe 852 personnes « Pour », 911 personnes « Contre » et 422 personnes « Sans opinion ».

   Exercice simulant une situation réelle. Vous n'avez pas accès à la totalité de l'information précédente. Ici, on parle de 2 185 individus, mais imaginez que votre population soit celle de la France entière ou du monde. Vous n'avez pas accès à ce que pense tout le monde. Il faut de fait l'échantillonner. Avec des outils de simulation, il est facile de générer 100 échantillons. aléatoires.

   — Dans le dossier `src`, introduire le dossier `data` le fichier `Echantillonnage-100-Echantillons.csv` disponible dans la `Seance-05` du `GitHub`

   — Ouvrir le fichier en utilisant la fonction locale `ouvrirUnFichier()`. Elle prend en paramètre une chaîne de caractères matérialisant l'adresse et le nom du fichier. Le fichier contient le résultat d'un tirage au hasard sans remise dans la population initiale de 100 échantillons. Sur chaque ligne, il est compté le nombre d'individus ayant telle ou telle opinion.

   — Pour chaque colonne, calculer la moyenne obtenue. Le nombre de personnes devant être entier, il faut arrondir votre calcul avec aucune décimale en respectant la règle de l'arrondi avec la fonction native `round()`.

   — Pour comparer l'échantillon avec sa population mère, il faut calculer les fréquences. Pour ce, calculer la somme des trois moyennes obtenues, puis diviser ce résultat à l'ensemble de vos moyennes. Calculer également suivant le même principe les fréquences de la population mère afin de comparer les deux résultats. Arrondir les fréquences obtenues à deux décimales.

— Vous devez constater un écart entre les valeurs observées en moyennant les échantillons et les valeurs réelles de la population mère. Calculer l'intervalle de fluctuation de chacune des fréquences à un seuil de 95 %, soit $z_C = 1,96$.

— Dans votre rapport, expliquer le lien entre l'intervalle de fluctuation et les valeurs réelles de la population mère. Que pouvez-vous en conclure par rapport aux échantillons utilisés pour le calcul ?1

2. Théorie de l'estimation

Problème à résoudre. Dans le cas d'espèce, et c'est un des cas les plus fréquents en sciences humaines et sociales, on ne possède qu'**un** échantillon de la population mère. Comment avoir confiance en lui ? Ici, la méthode consiste à construire des intervalles de confiance.

— Prendre le premier échantillon de la liste précédente en utilisant la méthode `Pandas iloc(0)`, le paramètre 0 correspondant à votre première ligne de données. Il faudra utiliser des fonctions natives, donc convertissez l'objet `Pandas` en castant une `list()`.

— Calculer la somme de la ligne, puis comme avec le cas précédent, les fréquences en utilisant l'effectif total de l'échantillon isolé.

— L'intervalle de confiance ne dépend que de la taille de l'échantillon, c'est-à-dire la somme précédente. Calculez cet intervalle pour chaque opinion.

— Dans votre rapport, vous interpréterez le résultat obtenu et vous le comparerez avec le résultat précédent.

N.B. N'hésitez pas à comparer le résultat avec plusieurs lignes de l'échantillon afin de valider votre jugement.

3. Théorie de la décision

Problème à résoudre. Intervalles de fluctuation ou intervalles de confiance peuvent être insuffisants pour valider un résultat. En fonction de la nature des variables (quantitatives ou qualitatives), les statisticiens ont inventé et inventent encore de nombreux tests statistiques. Le test de Shapiro-Wilks permet de tester si une distribution suit la loi normale.

— Vous disposez de deux fichiers `Loi-normale-Test-1.csv` et `Loi-normale-Test-2.csv`. Il s'agit d'une série de nombres aléatoires traduisant une distribution statistique. En utilisant la méthode de `scipy.stats shapiro()`. Laquelle est une distribution normale ?

— Vous expliquerez dans votre rapport pourquoi.

N.B. Il vous faudra essentiellement connaître les types de tests en fonction des situations.

2.3 Bonus

Lors du test de la décision, l'une des lois n'est pas normale. En vous aidant de la séance précédente, quelle est sa distribution ?