Cours d'analyse de données en géographie
Niveau Master 1 - GEANDO
Séance 8. Étude de deux variables qualitatives

Maxime Forriez[1,a]

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

28 septembre 2025

1 Questions de cours

Les réponses comptent pour 20 % de la note finale du parcours « intermédiaires ».
Les réponses comptent pour 10 % de la note finale du parcours « confirmés ».

1. La corrélation entre deux variables qualitatives a-t-elle un sens ? Expliquez votre réponse.
2. Pourquoi pratiquer le test d'indépendance du χ^2 ?
3. Expliquez dans un court paragraphe ce qu'est l'analyse de la variance à simple entrée.
4. Qu'est-ce qu'un rapport de corrélation ? Quelle différences avec la correspondance ?
5. Qu'est-ce qu'une analyse factorielle ?
6. Expliquez en un court paragraphe ce qu'est l'analyse factorielle des correspondances.

2 Mise en œuvre avec Python

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

2.1 Objectifs

— Manipuler des variables qualitatives avec `Python`
— Calculer les principaux indicateurs de liaison avec la bibliothèque `scipy.stats`

2.2 Manipulations

Le fichier obtenu compte pour 20 % de la note finale du parcours « intermédiaires ».

Le fichier obtenu compte pour 15 % de la note finale du parcours « confirmés ».

Vous allez utiliser les données 2024 de l'Institut national de la statistique et des études économiques (I.N.S.E.E.) concernant la relation entre catégorie socioprofessionnelles et le sexe biologique https://www.insee.fr/fr/statistiques/2381478#figure1_radio2. J'ai simplifié le tableau initial. Si vous avez compris le cours, vous avez compris qu'il s'agit d'un tableau de contingence avec deux variables qualitatives. Normalement, vous le calculerez un tableau croisé dynamique avec la méthode de la bibliothèque `Pandas`, `crosstab()`. Toutefois, il arrive souvent, comme dans le cas présent, que le tableau de contingence n'ait pas être calculé. De fait, on ne peut pas utiliser les méthodes de `Pandas` directement, notamment pour calculer les marges.

1. Calculer les marges des lignes et des colonnes en vous servant des fonctions locales `sommeDesColonnes()` et `sommeDesLignes()`.
2. Faire une condition vérifiant si le total des marges des lignes et le total des marges des colonnes est identique.
3. Faire un test d'indépendance du χ^2 à partir de la bibliothèque `scipy.stats` et de sa méthode `chi2_contingency()`. Existe-t-il une liaison ?
4. Calculer l'intensité de liaison ϕ^2 de Pearson à partir des résultats précédents.
5. Dans votre rapport, faire un court paragraphe expliquant vos résultats.

2.3 Bonus

Les points bonus sont attribués par rapport aux chapitres d'approfondissement.

Avec le fichier `Echantillonnage-100-Echantillons.csv`, faire une analyse ANOVA.

Avec le tableau de données, calculer une A.F.C. et faisant le commentaire.