Séance 8 – Étude de deux variables qualitatives

Honoka OYAMADA

Partie 1 – Questions de cours

1. Corrélation entre deux variables qualitatives

La notion de corrélation au sens strict s'applique aux variables quantitatives. Pour les variables qualitatives, on parle plutôt de liaison ou d'association. On cherche à savoir si la répartition des modalités d'une variable dépend de celles d'une autre.

2. Intérêt du test du χ^2

Le test d'indépendance du χ^2 permet de vérifier si deux variables qualitatives sont indépendantes. Il compare les effectifs observés à des effectifs théoriques attendus sous hypothèse d'indépendance.

3. Analyse de la variance (ANOVA)

L'ANOVA à simple entrée permet de comparer les moyennes d'une variable quantitative entre plusieurs groupes définis par une variable qualitative, afin de déterminer si les différences observées sont statistiquement significatives.

4. Rapport de corrélation

Le rapport de corrélation mesure la part de variance expliquée par une variable explicative. Contrairement à la correspondance, il s'applique à des variables quantitatives expliquées par des variables qualitatives.

5. Analyse factorielle

Une analyse factorielle est une méthode de réduction de dimension qui vise à résumer l'information contenue dans un grand nombre de variables à l'aide de quelques axes synthétiques.

6. Analyse factorielle des correspondances (AFC)

L'AFC est une méthode factorielle adaptée aux tableaux de contingence. Elle permet d'analyser simultanément les relations entre les modalités des variables en projetant lignes et colonnes dans un espace factoriel.

Partie 2 – Manipulation avec Python

À partir du tableau de contingence Catégorie socioprofessionnelle × Sexe, les marges de lignes et de colonnes ont été calculées. Les totaux sont identiques, ce qui confirme la cohérence des données.

Le test du χ^2 d'indépendance donne une p-value très faible, ce qui conduit à rejeter l'hypothèse d'indépendance : il existe une liaison statistiquement significative entre le sexe et la catégorie socioprofessionnelle.

L'intensité de liaison φ^2 de Pearson vaut environ 0,088. Cette valeur indique une liaison faible : la dépendance est significative mais l'effet reste de faible ampleur.

Bonus

L'ANOVA réalisée sur le fichier Echantillonnage-100-Echantillons.csv montre une différence très significative entre les groupes (p-value extrêmement faible).

L'analyse factorielle des correspondances montre que l'essentiel de l'inertie est porté par le premier axe ($\approx$ 8,8 %). La structure du tableau est donc essentiellement unidimensionnelle.

Partie 2.2 – Manipulations détaillées

1. Calcul des marges

Le tableau fourni par l'INSEE est déjà un tableau de contingence croisant la catégorie socioprofessionnelle et le sexe. Les marges des colonnes (Femmes, Hommes) et des lignes (catégories) ont été calculées à l'aide des fonctions locales sommeDesColonnes() et sommeDesLignes(). Ces fonctions parcourent explicitement le tableau, car les méthodes Pandas classiques ne peuvent pas être utilisées directement dans ce cas.

2. Vérification de la cohérence des totaux

Une condition logique a été appliquée afin de vérifier que la somme des marges des lignes est égale à la somme des marges des colonnes. Les deux totaux étant identiques, la cohérence globale du tableau est confirmée et les données peuvent être utilisées pour un test statistique.

3. Test d'indépendance du χ^2

Le test du χ^2 d'indépendance a été réalisé à l'aide de la fonction chi2_contingency() de la bibliothèque scipy.stats. La p-value obtenue est extrêmement faible, ce qui conduit à rejeter l'hypothèse nulle d'indépendance. Il existe donc une liaison statistiquement significative entre le sexe et la catégorie socioprofessionnelle.

4. Intensité de liaison φ^2 de Pearson

L'intensité de la liaison a été mesurée à l'aide du coefficient φ^2 de Pearson, calculé comme le rapport entre la statistique du χ^2 et l'effectif total. La valeur obtenue ($\varphi^2 \approx 0{,}088$) indique une liaison faible : bien que la dépendance soit statistiquement significative, l'ampleur de la relation entre les deux variables reste limitée.

5. Interprétation synthétique

En conclusion, l'analyse montre que la répartition des femmes et des hommes varie selon les catégories socioprofessionnelles. Cette dépendance est statistiquement significative, mais d'intensité modérée. Les résultats confirment l'intérêt d'utiliser à la fois le test du χ^2 pour détecter une liaison et le coefficient φ^2 pour en mesurer l'intensité.