

Explore data TD-TP : Analyse bivariée - Durée : 3h

Qualitatif \times Qualitatif

Exercice 1. Niveau de diplôme en fonction du sexe

Le tableau suivant donne la répartition des salariés d'une entreprise selon le niveau de formation X et le sexe Y .

| X Y | F | H |
|-------|----|----|
| Bac+3 | 45 | 49 |
| Bac+5 | 16 | 11 |
| Bac+8 | 4 | 6 |

- 1) Quels calculs est-il possible de faire avec ce type de tableau ?
- 2) Que représentent $n_{3\bullet}$ et $n_{\bullet 1}$? Que représente $f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}$? Calculer f_{31} et interpréter les résultats.
- 3) Donner les distributions marginales des effectifs et des fréquences du caractère X (resp. Y)
- 4) Que représente $f_{i|j}$ et $f_{i\bullet}$? Faire une phrase interprétant les fréquences suivantes : $f_{3|1}$, $f_{3\bullet}$ et $f_{\bullet 1}$.
- 5) Établir le tableau des profils lignes et des profils colonnes. Que peut-on en déduire ?
- 6) Comparer le tableau des effectifs théoriques avec celui des effectifs observés. Calculer la distance du chi-deux. Peut-on conclure sur l'indépendance de X et Y ?

Exercice 2.

On a interrogé 410 individus d'une population et on leur a proposé deux questions :

- Dans quelle ville habitez-vous ?
- Quel type de programme regardez-vous le plus souvent à la télévision dans la liste suivante : actualités, reportages, films, séries, dessins animés, variétés, sports, jeux ?

Les réponses sont données dans le tableau suivant :

| Ville | Actu | Report | Films | Séries | dess.anim | Variétés | Sport | Jeux |
|-------|------|--------|-------|--------|-----------|----------|-------|------|
| V1 | 10 | 6 | 18 | 10 | 3 | 1 | 11 | 4 |
| V2 | 5 | 6 | 5 | 7 | - | 4 | 6 | 2 |
| V3 | 2 | 2 | 3 | 4 | 1 | 4 | 5 | 6 |
| V4 | 2 | 5 | 8 | 9 | 1 | 1 | 5 | 5 |
| V5 | 8 | 2 | 12 | 7 | 1 | 6 | 7 | 1 |
| V6 | 1 | 2 | 4 | 1 | - | 5 | 5 | 3 |
| V7 | 2 | 1 | 3 | 3 | 1 | 4 | 2 | 4 |
| V8 | 5 | 4 | 4 | 6 | 3 | 3 | 2 | 2 |
| V9 | 6 | 2 | 13 | 8 | - | - | 6 | 2 |
| V10 | 3 | 1 | 2 | 2 | 2 | 4 | 4 | 6 |
| V11 | 8 | 4 | 6 | 5 | 1 | 3 | 6 | 1 |
| V12 | 5 | 3 | 9 | 7 | 2 | 3 | 8 | 3 |

Nous souhaitons savoir si le type de programme regardé diffère selon le type de la ville. Nous regroupons certains modalités entre elles pour définir deux variables X et Y ayant les modalités suivantes :

- x_1 : Capitale et banlieue regroupant V1, V5, V9 et V12.
- x_2 : Villes moyennes regroupant V2, V4, V8 et V11.
- x_3 : Petites villes regroupant V3, V6, V7 et V10.
- y_1 : Regroupant actualités et reportages.
- y_2 : Regroupant films, séries et dessins animés.
- y_3 : Regroupant variétés, sport et jeux

1. Dresser le tableau de contingence croisant les variables X et Y . Quel est le croisement le plus fréquent ?
2. Dresser le tableau des effectifs théoriques sous l'indépendance puis celui des contingences simples. Quelle est la contingence simple la plus importante ?
3. On donne $\chi^2 = 26.45$. calculer la contribution de chacune des deux contingences mentionnées ci-dessus. Commenter.

Exercice 3. Durée du Chômage en fonction de l'âge et le sexe.

Étudier le fichier DureeChomageData.xls

Exercice 4. Couleur des cheveux de garçons et de fille

Cet exemple historique est dû à Fisher. Il étudie la couleur des cheveux de garçons et de filles d'un discret écossais.

| | Blond | Roux | Châtain | Brun | Noir de jais |
|--------|-------|------|---------|------|--------------|
| Garçon | 592 | 119 | 849 | 504 | 36 |
| Fille | 544 | 97 | 677 | 451 | 14 |

Nous souhaitons savoir si la couleur des cheveux est indépendante du sexe (avec une erreur de 5%). Si la couleur dépend du sexe, nous aimerions avoir une idée des couleurs qui sont les plus dépendantes du sexe.

Étapes

1. Saisir les données

- `tab <- matrix(c(592,544,119,97,849,677,504,451,36,14), ncol=5)`
- `rownames(tab) <- c("Garçon", "Fille")`
- `colnames(tab) <- c("Blond", "Roux", "Châtain", "Brun", "Noir de jais")`

2. Visualiser les données

- `par(mfrow=c(2,1))`
- `couleur <- c("Gold", "OrangeRed", "Goldenrod", "Brown", "Black")`
- `barplot(tab[1,], main="Garçon", col=couleur)`
- `barplot(tab[2,], main="Fille", col=couleur)`

3. Calculer les profils lignes et les profils colonnes

- a) `round(100 * tab / sum(tab), 1)`
- `tab <- as.matrix(tab)`
- `round(100 * prop.table(tab, margin=1), 1)`

4. Construire le test du χ^2

- `resultat <- chisq.test(tab)`

5. Calculer les constructions au χ^2 .

- `round(100 * resultat$residuals2 / resultat$stat, 1)`
- `round(resultat$residuals, 3)`

Exercice 5. Propriétés

L'objectif de cette exercice est de démontrer quelques propriétés vues dans le cours.

1. L'égalité des profils lignes entraîne l'égalité des profils colonnes

On rappelle que:

- le profil de la i -ème ligne est:

l'ensemble des fréquences des modalités de Y dans la modalité x_i de X
c'est-à-dire les $\frac{n_{ij}}{n_{i\bullet}}$ pour j variant de 1 à s .

- le profil de la j -ème colonne est:

l'ensemble des fréquences des modalités de X dans la modalité y_j de Y
c'est-à-dire les $\frac{n_{ij}}{n_{\bullet j}}$ pour i variant de 1 à r .

a) Vérifier que l'égalité des profils lignes dans un tableau de contingence se traduit ainsi:

$$\text{pour tout } j \text{ fixé et pour tout } i \text{ variant de 1 à } r, \text{ on a : } \frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n}$$

b) En déduire que l'on a alors l'égalité des profils colonnes.

2. La somme des contingences simples est nulle

Vérifier sur une ligne ou une colonne quelconque que la somme des contingences simples $n_{ij} - n_{ij}^*$ est nulle.

3. $\chi^2_{\max} = n \times \min(r - 1, s - 1)$

a) En développant le carré dans l'expression de χ^2 , montrer que l'on a:

$$\chi^2 = n \times \left(\sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} \right) - n.$$

b) En remarquant qu'une fréquence est toujours inférieure ou égale à 1, montrer que l'on a:

$$\chi^2 \leq n(s - 1) \quad \text{et} \quad \chi^2 \leq n(r - 1)$$

c) en déduire l'inégalité

$$\chi^2 \leq n \min(s - 1, r - 1)$$

d) on considère un tableau de contingence tel que, sur chaque ligne, une seule case contient des effectifs non nul. En reprenant l'écriture du χ^2 obtenue en 3 a), montrer qu'alors

$$\chi^2 = n(s - 1)$$