



THÉMATIQUE 3 - ANALYSE BIVARIÉE DE VARIABLES QUALITATIVES

Data Sciences Revision
MENTION MATHÉMATIQUES ET INFORMATIQUE
PARCOURS HPDA

12/09/2024

Rédigé par :

PAULY ALEXANDRE

alexandre.pauly@cy-tech.fr

L'objectif d'une analyse bivariée est de faire lien entre deux variables (ici qualitatives). Il est donc possible de faire le lien entre l'âge d'une personne et son ancienneté de chômage par exemple.

L'étude de ce croisement nous conduit à effectuer un test du Khi-2.

1 Mise en place d'une analyse d'indépendance

1.1 Préparation des données

Comme pour tout jeux de données, il faut démarrer par une analyse des données et un visionnage. Par contre, dans le cas d'un tableau qualitatif présenté comme celui de la **Figure 1**, il faut effectuer une transformation du tableau pour obtenir les effectifs. Effectifs nécessaires à la réalisation du test du Khi-2.

Attention : Les tests d'indépendance sur les variables s'effectuent entre 2 variables, il faut donc faire des sélections de variables ou alors faire le test pour chaque combinaison.

Le terme "femme" et "sans diplôme" apparaît 6 fois dans les données.

Cas	Genre	Niveau d'études le plus élevé
1	Homme	enseignement primaire
2	Femme	enseignement supérieur
3	Homme	Sans diplôme
4	Homme	enseignement primaire
5	Femme	enseignement secondaire
...

	Femme	Homme
Sans diplôme	6	7
enseignement primaire	13	16
enseignement secondaire	16	15
enseignement supérieur	8	11
Total	43	49

Figure 1: Transformation des données

Remarque : Transformer les données sous forme de tableau de contingence permet de facilement déterminer le nombre total d'individus pour chaque catégorie. Mais cela est surtout important pour l'analyse car des indicateurs comme la moyenne ou la médiane ne sont pertinents dans un croisement qualitatif x qualitatif.

1.2 Effectifs et fréquences observées

Dans le cas de variables qualitatives, on utilise uniquement des calculs d'effectifs ou de fréquences auquel on ajoute les effectifs (resp fréquences) totaux dans une colonne et

ligne en extrémité des tableaux. Ici, les indicateurs numériques (moyenne, médiane, ...) ne veulent rien dire puisqu'il n'y a pas d'échelle de mesure (même si les modalités sont codées 0, 1, 2, ...).

Table 1: Tableau des effectifs observés

X \ Y	F	H	Total
Bac+3	45	49	94
Bac+5	16	11	27
Bac+8	4	6	10
Total	65	66	131

Table 2: Tableau des fréquences observées

X \ Y	F	H	Total
Bac+3	0.34	0.37	0.71
Bac+5	0.12	0.08	0.2
Bac+8	0.03	0.04	0.07
Total	0.49	0.5	1

Exemples de lecture :

- $n_{3.} = 10$: Représente les effectifs des Bac+8 tous sexes confondus.
- $n_{.1} = 65$: Représente les effectifs des femmes tous niveaux de formation confondus.
- $f_{31} = 0.03$: Il y a 3% des salariés qui sont des femmes ayant Bac+8.
- $f_{3.1} = 0.076$: Il y a 7,6% des salariées qui ont un Bac+8.
- $f_{.1} = 0.496$: Il y a 49,6% de femmes dans l'entreprise.
- $f_{3|1} = \frac{4}{65} = 0.06$: Il y a 6% de Bac+8 parmi les femmes salariées.
- $f_{1|3} = \frac{4}{10} = 0.4$: Il y a 40% de femmes parmi les Bac+8.

1.3 Effectifs et fréquences théoriques

Les effectifs théoriques se calculent avec la formule suivante : $\frac{n_{i.} \times n_{.j}}{n_{..}}$

Table 3: Tableau des effectifs théoriques

X \ Y	F	H	Total
Bac+3	46.64	47.36	94
Bac+5	13.4	13.6	27
Bac+8	4.96	5.04	10
Total	65	66	131

Table 4: Tableau des fréquences théoriques

X \ Y	F	H	Total
Bac+3	0.36	0.36	0.72
Bac+5	0.1	0.1	0.2
Bac+8	0.04	0.04	0.08
Total	0.5	0.5	1

1.4 Tableau des différences

Le tableau des différences est calculé à partir de la formule suivante : $\frac{(\text{eff théorique} - \text{eff observé})^2}{\text{eff théorique}}$.

Table 5: Tableau des différences

$X \setminus Y$	F	H
Bac+3	0.06	0.06
Bac+5	0.5	0.5
Bac+8	0.19	0.18

C'est la somme de tous les éléments de ce tableau qui donnera la valeur du Khi-2.

1.5 Tableaux des profils

Les tableaux des profils servent à mieux comprendre les relations entre deux variables qualitatives. Ils permettent de comparer les distributions relatives des catégories de ces variables et sont souvent utilisés dans le cadre de l'analyse des correspondances.

1.5.1 Profils lignes

Les profils lignes permettent de voir, pour une ligne spécifique (une modalité de la première variable), comment les individus sont répartis selon les colonnes (les modalités de la seconde variable).

Table 6: Tableau des profils lignes

$X \setminus Y$	F	H	Total
Bac+3	0.48	0.52	1
Bac+5	0.6	0.41	1
Bac+8	0.4	0.6	1
Fréq marginale	0.5	0.5	1

Exemple de lecture : Il y a 50% de femmes dans l'entreprise. Ce pourcentage descend à 40% chez les bac+8 et augmente à 60% chez les bac+5. Cela laisse penser que la répartition des femmes dépend du niveau de formation.

Attention : On compare les profils lignes avec les fréquences marginales des sexes (=profil ligne moyen). Attention, on reporte la ligne des fréquences marginales obtenue dans le tableau des fréquences et on ne calcule pas la somme ou la moyenne des profils lignes!!!!

1.5.2 Profils colonnes

Les profils colonnes permettent d'observer, pour une colonne donnée, comment les individus sont répartis selon les lignes.

Table 7: Tableau des profils lignes

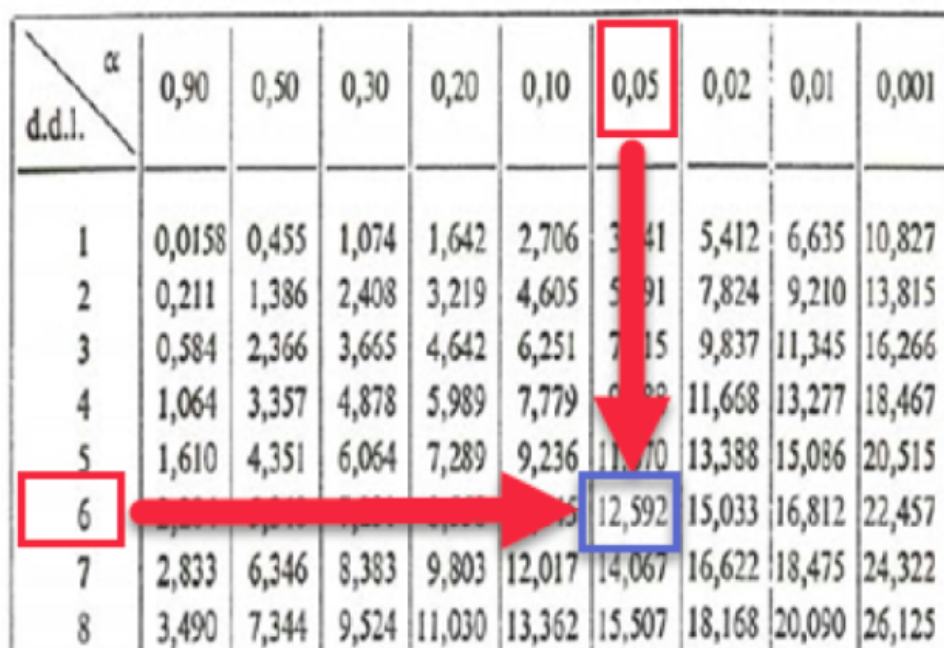
X \ Y	F	H	Fréq marginale
Bac+3	0.69	0.74	0.72
Bac+5	0.25	0.17	0.21
Bac+8	0.06	0.1	0.1
Total	1	1	1

Exemple d'analyse : Concernant les profils colonnes, il y a beaucoup moins d'écart entre la répartition total des niveaux de formation et la répartition par sexe. Il ne semble pas y avoir de dépendance entre ces variables.

2 Test du khi-2

Le test du Khi-2 est un test d'indépendance entre deux caractères. Il faut donc un tableau de contingence, comme précédemment expliqué. Ce test suppose deux hypothèses : **(H0)** : Hypothèse dite nulle, ce qui suppose que les caractères sont indépendants. Et l'hypothèse **(H1)** : Hypothèse dite alternative, qui vient contredire la précédente.

La valeur du Khi-2 est obtenue par la somme du tableau des différences. Pour prendre une décision, il faut comparer cette valeur à un seuil déterminé en fonction du degré de liberté ((nombre de colonnes - 1) × (nombre de lignes - 1)). Ci-dessous sont présentes quelques valeurs du seuil en fonction du ddl et du risque d'erreur (souvent $\alpha = 5\%$).



The table shows critical values for the Chi-squared test. A red box highlights the 0.05 significance level in the header. A red arrow points from this box down to the value 12.592, which is also highlighted with a blue box. Another red box highlights the 6 degrees of freedom in the first column, with a red arrow pointing from it to the same value 12.592.

α \ d.d.l.	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,715	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,296	7,231	8,558	10,597	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125

Figure 2: Table du Khi-2

Ainsi, la règle de décision est la suivante :

- $\text{Khi2} > \text{seuil} \Rightarrow$ Les variables sont fortement liées.
- $\text{Khi2} < \text{seuil} \Rightarrow$ Les variables sont indépendantes.

Attention : Ce test ne peut s'appliquer dans tous les cas. Ses conclusions sont valables uniquement si l'effectif total est supérieur à 30 et si aucun effectif théorique n'est inférieur à 5. Cette dernière restriction n'est pas toujours satisfaite, ce qui n'empêche pas le calcul, mais ne permet de tirer aucune conclusion de ce travail puisque les conditions de ce test ne sont pas respectées.

References

- [1] t-Test, khi-deux, ANOVA, Régression, Corrélation... (s. d.). Calculatrice de statistiques en ligne : test t, khi-deux, régression, corrélation, analyse de variance, <https://datatab.fr/tutorial/cross-table>.
- [2] (s. d.). ARISTERI.COM - Site Bernard Andruccioli., https://aristeri.com/pages/statistiques/test_khi2/fiche_khi2.pdf.