Analyse de données quantitatives (LPOLS1221)

Séance 3 : Tableaux croisés et test khi-carré

Les tableaux bivariés

Aussi appelés « tableaux de contingence », ils permettent d'avoir un premier aperçu des deux variables dont on veut étudier la relation : la variable indépendante (ou explicative, la « cause ») et la variable dépendante (ou expliquée, « l'effet »). Faire la distinction entre les deux permet de bien construire et analyser le tableau bivarié et donc en tirer les conclusions correctes (comparer les fréquences d'une modalité de la variable indépendante à l'autre).

Le test Khi-carré

Objectif : tester l'existence d'une relation entre deux variables catégorielles (vérifier si l'association observée dans l'échantillon est due au hasard ou si l'on peut généraliser les résultats à la population d'où l'échantillon a été tiré).

Conditions d'application : maximum 20% des effectifs théoriques (E.T.) peuvent présenter une valeur inférieure à 5 et tous les E.T. sont supérieurs à 1

Procédure : on teste l'hypothèse nulle (H_0) qui prévoit une **absence de relation** entre les **deux variables catégorielles**. Le test se base sur la comparaison des effectifs observés (E.O.) et des effectifs théoriques (E.T.) (observés s'il n'y avait aucune relation entre les deux variables)

$$\chi^2 = \sum \frac{(E.O. - E.T.)^2}{E.T.}$$

Interprétation : à partir de la p-valeur (du niveau de signification) associée à la valeur Khi-carré et du seuil d'erreur que l'on est prêt à accepter. La p-valeur représente le risque de se tromper en rejetant H_0 et donc en affirmant qu'il y a une relation entre les deux variables.

- Signification < seuil → faible probabilité de se tromper en affirmant qu'il y ait une relation non due au hasard entre deux variables
- Signification > seuil → probabilité trop élevée de se tromper

Imaginons un tableau croisé, avec des informations sur le genre et le niveau de salaire des individus :

Effectifs observés		**Niveau de salaire**		**Total**
		Elevé	**Faible**	
Genre	**Hommes**	60 50%	60 75%	120 60%
	Femmes	60 50%	20 25%	80 40%
Total		120	80	200

Effectifs théoriques : si l'échantillon comprend 60% d'hommes et 40% de femmes, et qu'il n'y a pas de différences de salaire, on devrait logiquement conserver les proportions 60/40 à l'intérieur des deux modalités de niveau de salaire. On obtiendrait ceci :

Effectifs théoriques		**Niveau de salaire**		**Total**
		Elevé	**Faible**	
Genre	**Hommes**	72 60%	48 60%	120 60%

	Femmes	48 40%	32 40%	80 40%
Total		120	80	200

La valeur du Khi-carré est obtenue à partir des différences entre E.T. et E.O. :

$$\chi^2 = \sum \frac{(E.O. - E.T.)^2}{E.T.} = \frac{(60-72)^2}{72} + \frac{(60-48)^2}{48} + \frac{(60-48)^2}{48} + \frac{(20-32)^2}{32}$$
$$= \frac{144}{72} + \frac{144}{48} + \frac{144}{48} + \frac{144}{32} = 2 + 3 + 3 + 4.5 = 12.5$$

A l'aide d'une table des valeurs de Khi-carré, il est possible d'identifier une valeur seuil, à partir du degré de liberté et de la probabilité d'erreur que l'on est prêt à accepter en rejetant l'hypothèse nulle. Dans notre cas, la valeur observée du Khi-carré est de 12.5.

- Les degrés de liberté sont calculés comme le nombre de modalités en colonne moins 1, multiplié par le nombre de modalités en ligne moins 1. Dans ce cas :
$$dl = (2-1) * (2-1) = 1$$
- On définit un risque (une probabilité, α) que l'on est prêt à accepter de se tromper en rejetant H_0. Des seuils couramment utilisés sont est par exemple de 5% (0.05) ou de 1% (0.01). On rejette l'hypothèse nulle dès lors que la valeur du Khi-carré est supérieure à cette valeur.

dl	0,05	0,02	0,01	0,001
1	3,841	5,412	6,635	10,827
2	5,991	7,824	9,210	13,815
3	7,815	9,837	11,345	16,266
4	9,488	11,668	13,277	18,467
5	11,070	13,388	15,086	20,515

Dans notre cas, avec un seuil de 5% et 1 degré de liberté, si la valeur du Khi-carré dépasse 3.841, alors on peut rejeter H_0. D'ailleurs, on observe que quel que soit le α, la valeur du Khi-carré est plus élevée que le seuil. Cela signifie donc qu'on peut affirmer avec une probabilité d'erreur inférieure à 0.1% de se tromper qu'il y a une association entre niveau de salaire et genre.

La valeur du Khi-carré et sa p-valeur ne donnent aucun indice sur la force de l'association. Pour cela on dispose du ***V de Cramer*** (le coefficient de contingence), qui varie entre 0 et 1 :

$$V = \sqrt{\frac{\chi^2}{N * \min(r-1; c-1)}} = \sqrt{\frac{12.5}{200 * 1}} = 0.25$$

Pour les tableaux 2x2, le V de Cramer aura la même valeur qu'un autre coefficient calculé par SPSS, le coefficient Phi.

A titre indicatif, un coefficient situé autour de 0.1 sera faible, autour de 0.3 il indiquera une association moyenne et autour de 0.5 il sera fort. Des coefficients bien plus élevés que 0.5 suggèrent que les deux variables mesurent le même phénomène (concepts redondants).
Dans notre cas (0.25), on peut considérer que la force de l'association est moyenne.

Manipulations SPSS

Analyse (Analyze) ➔ Statistiques descriptives (Descriptive Statistics) ➔ Tableaux croisés (Crosstabs). Dans l'onglet « Statistics », sélectionner :

- Khi-carré (chi-square)
- Phi et V de Cramer
- les pourcentages selon la variable indépendante

Applications

Utilisez la base de données de l'Eurobaromètre 2017 pour réaliser ces exercices.

Nous nous intéressons à l'image que les individus ont de l'Union Européenne.

1. A partir de la variable **qa9** (l'image que le répondant a de l'UE), créez-en une nouvelle, qui s'appellera **opinion_ue** et qui regroupe les avis des répondants en trois modalités :
   - 1-2 → opinion positive
   - 3 → opinion neutre
   - 4-5 → opinion négative

   N'oubliez pas de correctement renseigner toutes les valeurs manquantes.

   Testez ensuite s'il y a une association entre le fait d'avoir confiance dans le gouvernement national (**qa8a_7**) et l'opinon sur l'Union Européenne (**opinion_UE**) sur la population totale de l'enquête.

   Quelles seraient les stratégies possibles pour le traitement des non réponses (DK) de la variable **qa8a_7** ? Discutez les avantages et désavantages de ces stratégies.

2. Sélectionnez les observations pour la Belgique (**country**) et testez la même association. Est-ce qu'il est nécessaire de refaire toutes les manipulations ou bien il y a moyen d'éviter cela, si l'on a déjà effectué ces manipulations ? Si vous regardez les résultats, est-ce que les conclusions tirées sur l'échantillon dans son ensemble restent valables ?

3. Sur l'entièreté de la base, sélectionnez uniquement les répondants en âge de travailler (âgés entre 15 et 64 ans - **d11**). Ensuite, appliquez une transformation de la variable âge (**d11**) en créant une nouvelle (**age_cat**), pour distinguer trois catégories : les étudiants (moins de 25 ans), travailleurs adultes (entre 25 et 54) et les travailleurs âgés (55 et plus). Utilisant cette nouvelle variable, testez si la catégorie d'âges influe sur l'opinion par rapport à l'Union Européenne (**opinion_UE**).

4. Est-ce qu'il y a une relation entre l'opinion sur l'Union Européenne (**opinion_UE**) et l'orientation politique (gauche/droite - **d1r1**) ?

5. Est-ce qu'il y a une association entre l'orientation politique (gauche/droite - **d1r1**) et le milieu de résidence (**d25**) au sein de la population française ? Et en ce qui concerne la population italienne ? Commentez.