

Seance 9.1: Chi-carré

Visseho Adjivanou, PhD.

16 March 2022

Rappel des objectifs de l'analyse descriptive

- 1 Y a-t-il une relation entre deux variables ?

Rappel des objectifs de l'analyse descriptive

- 1 Y a-t-il une relation entre deux variables ?
- 2 Quelle est l'intensité de cette relation?

Rappel des objectifs de l'analyse descriptive

- 1 Y a-t-il une relation entre deux variables ?
- 2 Quelle est l'intensité de cette relation?
- 3 Quelles sont la direction et la forme de cette relation?

Rappel des objectifs de l'analyse descriptive

- 1 Y a-t-il une relation entre deux variables ?
- 2 Quelle est l'intensité de cette relation?
- 3 Quelles sont la direction et la forme de cette relation?
- 4 Pouvons-nous **généraliser** la relation à la population de laquelle est tiré l'échantillon?

Rappel des objectifs de l'analyse descriptive

- 1 Y a-t-il une relation entre deux variables ?
- 2 Quelle est l'intensité de cette relation?
- 3 Quelles sont la direction et la forme de cette relation?
- 4 Pouvons-nous **généraliser** la relation à la population de laquelle est tiré l'échantillon?
- 5 La relation est-elle vraiment causale?

Rappel des objectifs de l'analyse descriptive

- 1 Y a-t-il une relation entre deux variables ?
- 2 Quelle est l'intensité de cette relation?
- 3 Quelles sont la direction et la forme de cette relation?
- 4 Pouvons-nous **généraliser** la relation à la population de laquelle est tiré l'échantillon?
- 5 La relation est-elle vraiment causale?
- 6 Quelles sont les variables intermédiaires qui relient les variables indépendantes et dépendante?

Rappel

			Type de la variable dépendante		
Type de la variable indépendante			Catégorielle		Continue
			Nominale	Ordinale	Intervalle/ratio
	Catégorielle	Nominale	Tableau croisé Test du chi-carré V de Cramer Diagramme de barre empilée	Tableau croisé	Analyse de la variance Test t État carré Diagramme boxplot
		Ordinale	Tableau croisé	Tableau croisé	idem
	Continue	Intervalle/ratio	Transformer la VI en catégorielle		Corrélation / régression Test t R carré Diagramme de nuage de points

Figure 1

Rappel

- 2 Quelle est l'intensité de cette relation?

Intensité de la relation

- Une relation peut-être nulle

	Y a-t-il une vie après la mort?			
Sexe	Oui	Non	Total	N
Homme	82,0	18,0	100,0	759
Femme	83,0	17,0	100,0	978

Figure 2

Intensité de la relation

- Une relation peut-être modérée

	Possédez-vous une arme à feu			
Sexe	Oui	Non	Total	N
Homme	48,0	52,0	100,0	848
Femme	35,0	65,0	100,0	1066

Figure 3

Intensité de la relation

- ou forte

Sexe	Peur de se promener dans son quartier?			
	Oui	Non	Total	N
Homme	26,0	74,0	100,0	846
Femme	55,0	45,0	100,0	1057

Figure 4

Introduction

- Nous sommes arrivés à ces différentes conclusions en regardant l'écart entre les pourcentages.
- Mais, que faire si les variables dépendante et indépendantes ont plusieurs modalités?
- Il serait bien pratique d'avoir un nombre unique qui donnerait une évaluation complète de l'intensité de la relation.
- Il permettrait ainsi de comparer plusieurs études dans des pays différents par exemple.
- Il existe plusieurs mesures d'intensité

Chi-carré ou chi-deux

Chi-carré

- Le chi-carré, χ^2 , est un nombre qui compare les fréquences observées dans un tableau bivarié aux fréquences auxquelles on devrait s'attendre (les fréquences anticipées) s'il n'y avait pas du tout de relation entre les deux variables dans la population.

Chi-carré

Voici un tableau observé entre le sexe et le fait de fumer la cigarette

Est-ce que tu fumes?	Homme	Femme	Total
Oui	570	215	785
Non	450	600	1050
Total	1020	815	1835

- Comment devrait être ce tableau pour ne pas avoir de relation?
- On va utiliser les informations à la marge pour trouver les éléments de ce tableau théorique.

Chi-carré

- Autrement dit, nous voulons les valeurs de A, B, C, D pour qu'il n'existe pas de relation entre le sexe et le fait de fumer

Est-ce que tu fumes?	Homme	Femme	Total
Oui	A	C	785
Non	B	D	1050
Total	1020	815	1835
Pourcentage (ligne)	55.6	44.4	100.0

- Vous voulez devinez?
- Voici un code simple qui vous donne en temps réel le résultat de votre proposition
- https://docs.google.com/spreadsheets/d/19nNUrw7hBCa4rgBwRh9bMBECO74gqOJzvSWmVx_sIXg/edit?usp=sharing

Chi-carré

- Il existe une formule simple pour trouver ces valeurs

$$f_a = \left(\frac{\text{Total de la colonne}}{N} \right) * (\text{Total de la rangée})$$

- fréquence anticipée d'une cellule: f_a ;

Chi-carré

- On peut ainsi facilement calculer la distance entre ces deux tableaux
- Il s'agit du chi-carré dont la formule est:

$$\chi^2 = \sum \frac{(f_o - f_a)^2}{f_a}$$

- Fréquence observée d'une cellule: f_o et
- N est le nombre total de cas
- \sum veut dire faire la somme de chaque calcul

Chi-carré

- Le Chi-deux n'est qu'une mesure de la distance entre deux tableaux, un tableau observé tiré de l'échantillon, et un tableau théorique
- Plus cette distance va être grande, plus nous pouvons affirmer qu'il y a de fortes chances que dans la population d'où est tirée cet échantillon, il existe une relation non nulle entre les deux variables

Chi-carré

- Le Chi-deux n'est qu'une mesure de la distance entre deux tableaux, un tableau observé tiré de l'échantillon, et un tableau théorique
- Plus cette distance va être grande, plus nous pouvons affirmer qu'il y a de fortes chances que dans la population d'où est tirée cet échantillon, il existe une relation non nulle entre les deux variables
- Plus cette distance est faible, plus, on ne peut affirmer qu'il existe une relation entre les deux variables au sein de la population

Réponse à notre exemple

Est-ce que tu fumes?	Homme	Femme	Total
Oui	570	215	785
Non	450	600	1050
Total	1020	815	1835

- Il faut d'abord trouver les valeurs anticipées f_a (qui sont les ABCD)
- Comme vous aimer maintenant R , on va utiliser un chunk pour le faire

Réponse à notre exemple

- Comme vous aimer maintenant R, on va utiliser un chunk pour le faire

```
A <- 785*1020/1835  
B <- 1050*1020/1835  
C <- 785*815/1835  
D <- 1050*815/1835  
  
c(A, B, C, D)
```

```
## [1] 436.3488 583.6512 348.6512 466.3488
```

Tableau anticipé

Est-ce que tu fumes?	Homme	Femme	Total
Oui	436.35	348.65	785
Non	583.65	466.35	1050
Total	1020	815	1835

- Remettez ces valeurs dans le fichier google sheet pour voir si cela fonctionne

Maintenant calculons le chi-carré

- Tableau anticipé: Pas d'association entre les deux variables

Est-ce que tu fumes?	Homme	Femme	Total
Oui	436.35	348.65	785
Non	583.65	466.35	1050
Total	1020	815	1835

- Tableau réellement observé : votre échantillon

Est-ce que tu fumes?	Homme	Femme	Total
Oui	570	215	785
Non	450	600	1050
Total	1020	815	1835

Maintenant calculons le chi-carré

La formule du chi-carré

$$\chi^2 = \sum \frac{(f_o - f_a)^2}{f_a}$$

```
Ao <- 570
```

```
Bo <- 450
```

```
Co <- 215
```

```
Do <- 600
```

```
chi_carre <- (Ao - A)^2/A + (Bo - B)^2/B + (Co - C)^2/C + (Do - D)^2/D
chi_carre
```

```
## [1] 161.0784
```

Maintenant calculons le chi-carré

- Ainsi la valeur du chi-carré vaut 161.09, qui est une très grande valeur. N'est-ce pas?
- Qui sais pourquoi je n'ai pas utilisé directement les chiffres dans ma formule?
- Refaites les mêmes calculs avec les autres exemples cités en haut.
- Que concluons-nous alors sur la relation au sein de la population d'où est tiré l'échantillon?