Seance 9.3: Mesure de l'intensité de l'association entre deux variables catégorielles

Visseho Adjiwanou, PhD.

15 March 2022

Introduction

- Le chi-carré permet de mesurer à quelle distance se trouve un tableau d'un tableau où il n'existe pas d'association au sein de la population
- Cependant, sa valeur est affectée par:
  - le nombre de ligne et de colonne
  - la taille de l'échantillon
- Autrement dit, un tableau qui a plus de lignes et de colonne aura tendance à avoir un chi-carré élevé.
- Il permettrait ainsi de comparer plusieurs études dans des pays différents par exemple.
- Il existe plusieurs mesures d'intensité: elles dépendent du niveau de mesure des deux variables.

Introduction

Ces mesures varient :

- entre 0 et 1 pour les variables nominales
- entre -1 et 1 (en passant par 0) pour les variables ordinales ou d'intervalle/ratio
- 0 Indique une absence de relation
- -1 et 1 indiquent une relation parfaite
- Plus la valeur d'une association est élevée, plus la relation est forte
- Le signe indique le sens de la relation

Introduction

Deux types de mesure de l'association 1. Basé sur le chi-carré

2. Basé sur la **réduction proportionnelle de l'erreur (RPE)**

Introduction

Deux types de mesure de l'association 1. Basé sur le chi-carré

2. Basé sur la **réduction proportionnelle de l'erreur (RPE)**

- Nous disent dans quelle proportion on réduit les erreurs de prédiction des scores de la **variable dépendante** lorsque l'on connaît les scores de **la variable indépendante**.

Introduction

APPLYING CONCEPTS IN EVERYDAY LIFE

Measures of Association and Levels of Measurement

Peter Nardi
Pitzer College

Note that this table itself is set up with the dependent variables in the rows and the independent variable in the columns, as tables are commonly organized. Also, notice that the levels of measurement are themselves an ordinal scale.

If you want to use an interval/ratio level variable in a crosstab, you must first recode it into an ordinal-level variable.

		Independent Variable		
		Nominal	Ordinal	Interval/Ratio
Dependent Variable	Nominal	Crosstabs Chi-square Lambda	Crosstabs Chi-square Lambda	
	Ordinal	Crosstabs Chi-square Lambda	Crosstabs Chi-square Lambda Gamma Kendall's tau Sommers' d	
	Interval/Ratio	Means *t*-test ANOVA	Means t-test ANOVA	Correlate Pearson *r* Regression (*R*)

Mesures d'association pour les variables dont l'une au moins est nominale

Mesures basées sur le chi-carré pour les variables nominales: le C, le V et le ϕ

- Le chi-carré dépend:
  - de l'intensité de la relation
  - du nombre de cas (N)

Coefficient de contingence de Pearson : C

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

- La division de χ^2 par $(\chi^2 + N)$ élimine l'effet du nombre de cas

Coefficient de contingence de Pearson : C

Problèmes

- Dépend du nombre de colonnes (c) et de rangées du tableau (r): augmente avec une augmentation de c et r
- C est toujours inférieur à 1
  - Pour une tableau (2,2), ie 2 colonnes et 2 rangées, la plus grande valeur de C est 0.71
  - Pour une tableau (3,3), ie 3 colonnes et 3 rangées, la plus grande valeur de C est 0.82
- Ce plafond inférieur à 1 rend l'interprétation un peu malaisée
- Ce n'est donc pas la mesure idéale

Le V de Cramer

- Le V de Cramer est semblable au C, mais s'ajuste au r et au c
- Il peut atteindre 1.
- Sa formule est:

$$V = \sqrt{\frac{\chi^2}{N * Min(r-1)(c-1)}}$$

- Min(a, b) étant le minimum entre les deux nombres
- Exemple: Min(12, 9) = 9; Min(3, 7) = 3

Le ϕ

- Cas particulier où le nombre de colonne est égale 2 ou le nombre de rangées est égale 2
- Dans ce cas, Min(r-1)(c-1) $= 1$
- Et V devient ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Le ϕ

- Le ϕ a une limite supérieur à 1 pour des tableaux de plus de deux rangées et deux colonnes.
- Le ϕ sera souvent élevé au carré

Remarque sur C, V et ϕ

- C, V et ϕ sont des mesures symétriques d'association: leur valeur ne dépend pas de la variable considérée comme dépendante ou indépendante

Remarque sur C, V et ϕ

- C, V et ϕ sont des mesures symétriques d'association: leur valeur ne dépend pas de la variable considérée comme dépendante ou indépendante
- Puisque C, V et ϕ sont basés sur le χ^2, le test de signification statistique du χ^2 s'applique aussi à eux:

Remarque sur C, V et ϕ

- C, V et ϕ sont des mesures symétriques d'association: leur valeur ne dépend pas de la variable considérée comme dépendante ou indépendante
- Puisque C, V et ϕ sont basés sur le χ^2, le test de signification statistique du χ^2 s'applique aussi à eux:
- Si le χ^2 pour un tableau est statistiquement significatif, les mesures d'association reposant sur lui le seront aussi.

Remarque sur C, V et ϕ

- C, V et ϕ sont des mesures symétriques d'association: leur valeur ne dépend pas de la variable considérée comme dépendante ou indépendante
- Puisque C, V et ϕ sont basés sur le χ^2, le test de signification statistique du χ^2 s'applique aussi à eux:
- Si le χ^2 pour un tableau est statistiquement significatif, les mesures d'association reposant sur lui le seront aussi.
- Si le χ^2 pour un tableau n'est pas statistiquement significatif, les mesures d'association reposant sur lui ne le seront pas aussi.

Le coefficient de prédiction de Guttman

- Encore appelé le lambda (λ)
- Mesure d'association pour les variables nominales
- Mesure l'intensité d'une relation en calculant dans quelle proportion on peut réduire les erreurs que l'on commet en prédisant le score de la variable dépendante d'un cas aussitôt que l'on connait la valeur de la variable **IN**dépendante de ce cas.

Le coefficient de prédiction de Guttman

- Encore appelé le lambda (λ)
- Mesure d'association pour les variables nominales
- Mesure l'intensité d'une relation en calculant dans quelle proportion on peut réduire les erreurs que l'on commet en prédisant le score de la variable dépendante d'un cas aussitôt que l'on connait la valeur de la variable **IN**dépendante de ce cas.
- Il ne repose pas sur le χ^2

Le coefficient de prédiction de Guttman

- Encore appelé le lambda (λ)
- Mesure d'association pour les variables nominales
- Mesure l'intensité d'une relation en calculant dans quelle proportion on peut réduire les erreurs que l'on commet en prédisant le score de la variable dépendante d'un cas aussitôt que l'on connait la valeur de la variable **IN**dépendante de ce cas.
- Il ne repose pas sur le χ^2
- N'est pas symétrique: le choix de la variable dépendante et indépendante est capital

Le coefficient de prédiction de Guttman

- Encore appelé le lambda (λ)
- Mesure d'association pour les variables nominales
- Mesure l'intensité d'une relation en calculant dans quelle proportion on peut réduire les erreurs que l'on commet en prédisant le score de la variable dépendante d'un cas aussitôt que l'on connait la valeur de la variable **IN**dépendante de ce cas.
- Il ne repose pas sur le χ^2
- N'est pas symétrique: le choix de la variable dépendante et indépendante est capital
- Se calcule à partir des fréquences bivariées

Le coefficient de prédiction de Guttman

- Encore appelé le lambda (λ)
- Mesure d'association pour les variables nominales
- Mesure l'intensité d'une relation en calculant dans quelle proportion on peut réduire les erreurs que l'on commet en prédisant le score de la variable dépendante d'un cas aussitôt que l'on connait la valeur de la variable **IN**dépendante de ce cas.
- Il ne repose pas sur le χ^2
- N'est pas symétrique: le choix de la variable dépendante et indépendante est capital
- Se calcule à partir des fréquences bivariées
- Se calcule à partir du mode de chaque rangée

Exemple

- N'a pas besoin de la fréquence marginale colonne (des modalités de la variable indépendante)

	q2_new	totalement d'accord	d'accord	Ne sait pas	En désaccord	Totalement en désaccord	Total
sexe							
Homme		208	304	12	418	419	1361
Femme		308	332	14	476	368	1498
Total		516	636	26	894	787	2859

Figure 2: lambda

Exemple

1. On ne considère pas la colonne totale
2. On détermine les valeurs modales de chaque ligne

- Pour Homme, c'est 419 (Totalement en désaccord)
- Pour Femme, c'est 476 (En désaccord)
- Pour Total, c'est 894 (En désaccord)

Exemple

3. Si nous ne connaissions que la variable dépendante, que prédirions-nous des individus de l'échantillon en terme de leur opinion sur la sexualité des jeunes filles?

- Chaque répondant est plus susceptible de répondre **En désaccord** qu'autre chose.

Exemple

3. Si nous ne connaissions que la variable dépendante, que prédirions-nous des individus de l'échantillon en terme de leur opinion sur la sexualité des jeunes filles?

- Chaque répondant est plus susceptible de répondre **En désaccord** qu'autre chose.
- En prédisant donc que les 2859 répondants répondraient **en désaccord**, nous aurons 894 fois raisons et 1965 fois tort (2859 - 894)

Exemple

3. Si nous ne connaissions que la variable dépendante, que prédirions-nous des individus de l'échantillon en terme de leur opinion sur la sexualité des jeunes filles?

- Chaque répondant est plus susceptible de répondre **En désaccord** qu'autre chose.
- En prédisant donc que les 2859 répondants répondraient **en désaccord**, nous aurons 894 fois raisons et 1965 fois tort (2859 - 894)
- On ne peut pas faire mieux pour prédire dans quelle catégorie de la variable dépendante placée chaque cas, si on n'a pas d'information additionnelle

Exemple

4. Utiliser les informations de la variable indépendante: maintenant supposons que pour chaque répondant, nous connaissons le score de la variable indépendante, le sexe.

- Si nous savons qu'un répondant est de sexe masculin, nous minimisons nos erreurs en plaçant cet homme et tous les autres dans la catégorie (Totalement en désaccord). Nous ne commettrons que 942 erreurs (tous les hommes qui ont répondu autre chose)

Exemple

4. Utiliser les informations de la variable indépendante: maintenant supposons que pour chaque répondant, nous connaissons le score de la variable indépendante, le sexe.
   - Si nous savons qu'un répondant est de sexe masculin, nous minimisons nos erreurs en plaçant cet homme et tous les autres dans la catégorie (Totalement en désaccord). Nous ne commettrons que 942 erreurs (tous les hommes qui ont répondu autre chose)
   - Si nous savons qu'un répondant est femme, nous minimisons nos erreurs en plaçant cette femme et toutes les autres femmes dans la catégorie (En désaccord). De ce fait, nous commettons seulement 1022 erreurs

Exemple

5. Décompte totale
   - Erreur si aucune information : 1965

Exemple

5. Décompte totale
   - Erreur si aucune information : 1965
   - Erreur si information sur la variable **IN**dépendante: $1022 + 942 = 1964$

Exemple

5. Décompte totale

- Erreur si aucune information : 1965
- Erreur si information sur la variable **IN**dépendante: $1022 + 942 = 1964$
- Ainsi, l'information additionnelle réduit l'erreur de 1

Exemple

6. Calcul de Lambda

$$lambda = \frac{\text{Erreurs si VI est inconnue - Erreurs si VI est connue}}{\text{Erreurs si VI est inconnue}}$$

- lambda $1/1965 = 0{,}0005$, ce qui est très faible.

Le coefficient de prédiction de Guttman

Remarques

- On peut le calculer aussi en inversant la VI et la VD: On obtiendrai dans ce cas une valeur différente

Le coefficient de prédiction de Guttman

Remarques

- On peut le calculer aussi en inversant la VI et la VD: On obtiendrai dans ce cas une valeur différente
- On peut prendre alors la moyenne des deux valeurs pour avoir une valeur symétrique

Le coefficient de prédiction de Guttman

Remarques

- On peut le calculer aussi en inversant la VI et la VD: On obtiendrai dans ce cas une valeur différente
- On peut prendre alors la moyenne des deux valeurs pour avoir une valeur symétrique
- Lambda est dite de **réduction proportionnelle de l'erreur (RPE)**

Le coefficient de prédiction de Guttman

Remarques

- On peut le calculer aussi en inversant la VI et la VD: On obtiendrai dans ce cas une valeur différente
- On peut prendre alors la moyenne des deux valeurs pour avoir une valeur symétrique
- Lambda est dite de **réduction proportionnelle de l'erreur (RPE)**
- Ces mesures sont faciles à interpréter

Le coefficient de prédiction de Guttman

Remarques

- On peut le calculer aussi en inversant la VI et la VD: On obtiendrai dans ce cas une valeur différente
- On peut prendre alors la moyenne des deux valeurs pour avoir une valeur symétrique
- Lambda est dite de **réduction proportionnelle de l'erreur (RPE)**
- Ces mesures sont faciles à interpréter
- Nous disent dans quelle proportion on réduit les erreurs de prédiction des scores de la variable dépendante lorsque l'on connaît les scores de la variables indépendante.

Le coefficient de prédiction de Guttman

Remarques

- On peut le calculer aussi en inversant la VI et la VD: On obtiendrai dans ce cas une valeur différente
- On peut prendre alors la moyenne des deux valeurs pour avoir une valeur symétrique
- Lambda est dite de **réduction proportionnelle de l'erreur (RPE)**
- Ces mesures sont faciles à interpréter
- Nous disent dans quelle proportion on réduit les erreurs de prédiction des scores de la variable dépendante lorsque l'on connaît les scores de la variables indépendante.
- Lambda varie de 0 à 1 quel que soit la taille des tableaux

Le coefficient de prédiction de Guttman

Remarques

- Inconvénient: il peut donner comme résultat 0, dans des situations où il existe vraiment une relation entre les variables:
- C'est le cas où une catégorie de la **variable dépendante** a une fréquence beaucoup plus élevée que les autres

Le coefficient de prédiction de Guttman

Remarques

- Inconvénient: il peut donner comme résultat 0, dans des situations où il existe vraiment une relation entre les variables:
- C'est le cas où une catégorie de la **variable dépendante** a une fréquence beaucoup plus élevée que les autres
- A cause de cela, il est **peu utilisé** en science sociales où beaucoup de variables sont suffisamment asymétriques.

Mesures d'association pour les variables ordinales

Gamma de Goodman et Kruskal

- Repose sur la prédiction des scores de la variable dépendante à partir de la connaissance de la variable indépendante
- Similaire au calcul du lambda, mais tient compte de l'ordre entre les valeurs des variables

$$G = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées}}$$

Gamma de Goodman et Kruskal

- Repose sur la prédiction des scores de la variable dépendante à partir de la connaissance de la variable indépendante
- Similaire au calcul du lambda, mais tient compte de l'ordre entre les valeurs des variables

$$G = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées}}$$

- Varie de -1 à 1

Gamma de Goodman et Kruskal

- Repose sur la prédiction des scores de la variable dépendante à partir de la connaissance de la variable indépendante
- Similaire au calcul du lambda, mais tient compte de l'ordre entre les valeurs des variables

$$G = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées}}$$

- Varie de -1 à 1
- L'ordre des variables est important:

Gamma de Goodman et Kruskal

- Repose sur la prédiction des scores de la variable dépendante à partir de la connaissance de la variable indépendante

- Similaire au calcul du lambda, mais tient compte de l'ordre entre les valeurs des variables

$$G = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées}}$$

- Varie de -1 à 1
- L'ordre des variables est important:
- VI (ordonnée du haut vers le bas par ordre décroissant)

Gamma de Goodman et Kruskal

- Repose sur la prédiction des scores de la variable dépendante à partir de la connaissance de la variable indépendante
- Similaire au calcul du lambda, mais tient compte de l'ordre entre les valeurs des variables

$$G = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées}}$$

- Varie de -1 à 1
- L'ordre des variables est important:
- VI (ordonnée du haut vers le bas par ordre décroissant)
- VD (ordonnée de gauche vers la droite par ordre croissant)

Gamma de Goodman et Kruskal

- Repose sur la prédiction des scores de la variable dépendante à partir de la connaissance de la variable indépendante
- Similaire au calcul du lambda, mais tient compte de l'ordre entre les valeurs des variables

$$G = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées}}$$

- Varie de -1 à 1
- L'ordre des variables est important:
- VI (ordonnée du haut vers le bas par ordre décroissant)
- VD (ordonnée de gauche vers la droite par ordre croissant)
- Gamma plus élevé que les autres mesures ordinales : les égalités ne sont pas prises en compte

Exemple

- Voir Labo

Le D_{yx} de Somers

- Semblable au Gamma mais tient compte des pairs qui présentent des égalités
- Formule:

$$D_{xy} = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées} + E_x}$$

- où Ex est obtenue en multipliant la fréquence de la cellule noire par chacune ds fréquences des cellules ombragées (cellule égales)

Le D_{yx} de Somers

- Semblable au Gamma mais tient compte des pairs qui présentent des égalités
- Formule:

$$D_{xy} = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées} + E_x}$$

- où Ex est obtenue en multipliant la fréquence de la cellule noire par chacune ds fréquences des cellules ombragées (cellule égales)
- Il s'agit des informations en colonne

Le D_{yx} de Somers

- Semblable au Gamma mais tient compte des pairs qui présentent des égalités
- Formule:

$$D_{xy} = \frac{\text{Semblables - Opposées}}{\text{Semblables} + \text{Opposées} + E_x}$$

- où Ex est obtenue en multipliant la fréquence de la cellule noire par chacune ds fréquences des cellules ombragées (cellule égales)
- Il s'agit des informations en colonne
- On peut calculer aussi le Dyx en inversant les places des VI et VD.

Le tau-b et le taux-c

- Le tau-b de Kendall

$$tau - b = \sqrt{D_{yx} D_{xy}}$$

$$tau{-}b = \frac{\text{Semblables - Opposées}}{\sqrt{(\text{Semblables} + \text{Opposées} + E_x)(\text{Semblables} + \text{Opposées} + E_x)}}$$

Le tau-b et le taux-c

- Le tau-b de Kendall
- Même caractéristiques que le gamma et le Dxy

$$tau - b = \sqrt{D_{yx} D_{xy}}$$

$$tau{-}b = \frac{\text{Semblables - Opposées}}{\sqrt{(\text{Semblables} + \text{Opposées} + E_x)(\text{Semblables} + \text{Opposées} + E_x)}}$$

Le tau-b et le taux-c

- Le tau-b de Kendall
- Même caractéristiques que le gamma et le Dxy
- Symétrique comme gamma et a une interprétation de **réduction proportionnelle de l'erreur(RPE)**

$$tau - b = \sqrt{D_{yx} D_{xy}}$$

$$tau{-}b = \frac{\text{Semblables - Opposées}}{\sqrt{(\text{Semblables} + \text{Opposées} + E_x)(\text{Semblables} + \text{Opposées} + E_x)}}$$

Le tau-b et le taux-c

- Le tau-b de Kendall
- Même caractéristiques que le gamma et le Dxy
- Symétrique comme gamma et a une interprétation de **réduction proportionnelle de l'erreur(RPE)**
- Formule:

$$tau - b = \sqrt{D_{yx} D_{xy}}$$

$$tau{-}b = \frac{\text{Semblables - Opposées}}{\sqrt{(\text{Semblables} + \text{Opposées} + E_x)(\text{Semblables} + \text{Opposées} + E_x)}}$$

Le tau-b et le taux-c

- Le tau-b de Kendall
- Même caractéristiques que le gamma et le Dxy
- Symétrique comme gamma et a une interprétation de **réduction proportionnelle de l'erreur(RPE)**
- Formule:

$$tau - b = \sqrt{D_{yx} D_{xy}}$$

$$tau{-}b = \frac{\text{Semblables - Opposées}}{\sqrt{(\text{Semblables} + \text{Opposées} + E_x)(\text{Semblables} + \text{Opposées} + E_x)}}$$

- tau-b est inférieur à 1.00. peut atteindre 1.00 que dans les tableaux carrés

Le tau-b et le taux-c

- Le tau-b de Kendall
- Même caractéristiques que le gamma et le Dxy
- Symétrique comme gamma et a une interprétation de **réduction proportionnelle de l'erreur(RPE)**
- Formule:

$$tau - b = \sqrt{D_{yx} D_{xy}}$$

$$tau - b = \frac{\text{Semblables - Opposées}}{\sqrt{(\text{Semblables} + \text{Opposées} + E_x)(\text{Semblables} + \text{Opposées} + E_x)}}$$

- tau-b est inférieur à 1.00. peut atteindre 1.00 que dans les tableaux carrés
- difficile à interpréter

Le tau-b et le tau-c

- Le tau-c résout ce problème et fonctionne sur les tableaux rectangulaires aussi
- varie entre -1.00 et 1.00
- Formule:

$$tau - c = \frac{2Min(r, c) * (Semblables - Opposées)}{N^2 Min(r - 1, c - 1)}$$