Statistics & Explanatory Data Analysis

Association measures & tests

dr Marcin Chlebus, dr Ewa Cukrowska - Torzewska

Uniwersytet Warszawski
Wydział Nauk
Ekonomicznych

Measures of Association

First variable	Second variable	Measures	Tests
Nominal	Nominal	2 x 2 - Phi n x m – Cramer’s V , Goodman & Kruskal lambda	Fischer Exact test Chi Square test G test
Nominal	Ordinal	Freeman’s theta	Cochran – Armitage test
Ordinal	Ordinal	Kendall’s Tau-b, Goodman and Kruskal's gamma, Somers’ D	Linear-by-linear test
Ordinal	Ordinal/Dichotomus (represent continuous latent variable)	Polychoric/tetrachoric correlatrion	
Continuous	Ordinal/Dichotomus (represent continuous latent variable)	Biserial/polyserial correlation	
Continuous	Ordinal/Dichotomus	Point biserial/polyserial correlation	
Continuous	Continuous	Pearson, Spearman & Kendall Correlation	Correlation tests

Phi, Cramer's V & G-K lambda – Nominal Data

DATA:
- Two nominal variables with two or more levels each. Usually expressed as a contingency table.
- Experimental units aren't paired.
- For *phi*, the table is 2 x 2 only.
- Equivalent of correlation for nominal data

PHI

$$phi = \frac{(n_{11}n_{22} - n_{12}n_{21})}{\sqrt{n_{1.}n_{2.}n_{.2}n_{.2}}} = \frac{10 * 40 - 20 * 200}{\sqrt{210 * 60 * 30 * 240}} = 0.38$$

	FEMALE	MALE	JOINTLY
GOOD	10	200	210
BAD	20	40	60

CRAMER'S V

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(w-1, k-1)}}$$

χ^2 – chi square independence test statistic
n – number of observations
w – number of categories in dependent variable
k – number of categories in independent variable

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}.$$

-1<phi<1
0<V<1
0<Lambda<1

GOODMAN-KRUSKAL LAMBDA

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1} = \frac{60\ (Bad) - (10(B\ if\ F) + 40(G\ if\ M)}{60(Bad)} = 0{,}17$$

ε_1 - prediction error only for dependent variable

ε_2 - prediction error for dependent & independent variable

Kendall Tau, G-K Gamma & Somers’ D – Ordinal data

P – concordant pairs (X1>X2) i (Y1>Y2) or(X1<X2) i (Y1<Y2)

Q – discordant pairs (X1>X2) i (Y1<Y2) or (X1<X2) i (Y1>Y2)

KENDALL TAU

$$\tau_a = 2\frac{P-Q}{N(N-1)} \qquad \tau_b = \frac{P-Q}{\sqrt{(P+Q+T(depvar))(P+Q+T(indepvar))}}$$

GOODMAN - KRUSKAL GAMMA

$$\gamma = \frac{P-Q}{P+Q}$$

SOMERS’ D

Somers’ D is not symmetric $D_{YX} \neq D_{XY}$
For X binary – Somers’D is equal to Gini coefficient

$$D_{YX} = \frac{\tau(X,Y)}{\tau(X,X)} \equiv \frac{P-Q}{P+Q+T(depvar)}$$

Association Tests – nominal vs ordinal variable

DATA:

- One variable is nominal (with 2 or more level), the other one is ordinal

HYPOTHESIS:

- H0: There is no association between the two variables (they are independent).
- H1 (2-sided): There is an association between the two variables

Cochran – Armitage test for trend (for 2 x k contingency table)

	Y=1	Y=2	Y=3	R TOT
X=1	N_{11}	N_{12}	N_{13}	$N_{1.}$
X=2	N_{21}	N_{22}	N_{23}	$N_{2.}$
C TOT	$N_{.1}$	$N_{.2}$	$N_{.3}$	N

$$T = \sum_{i=1}^{k} t_i (N_{1i} N_{2.} - N_{2i} N_{1.})$$

There is an extention for nominal variable with *more than 2 categories*

k – number of categories for ordinal variable

t_i - weigths for each category (t=(0,1,2) for linear trend

$N_{1i} N_{2.} - N_{2i} N_{1.}$ - can be seen as the difference between N_{1i} and N_{2i} after reweighting the rows to have the same total

Association Tests – ordinal vs ordinal variable

DATA:
- Two ordinal variables with two or more levels each.

HYPOTHESIS:
- H0: There is no association between the two variables (they are independent).
- H1 (2-sided): There is an association between the two variables

Linear-by-linear model

$$log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j$$

$u_1 \leq u_2 \leq \ldots \leq u_I$ - - ordered row scores
$v_1 \leq v_2 \leq \ldots \leq v_I$ - - ordered column scores
λ_i^X - row effect
λ_j^Y - column effect
$\beta u_i v_j$ - interactions between scores for row and column variable

$$\beta \neq 0 \; indicates \; assosiation \; - effect \; of \; Y \; depends \; on \; values$$
$$\beta > 0 \; indicates \; positive \; assosiation$$
$$\beta < 0 \; indicates \; negative \; assosiation$$

Correlations – Pearson, Spearman, Kendall

DATA:
- For Pearson correlation, two interval/ratio variables. Together the data in the variables are bivariate normal. The relationship between the two variables is linear.
- For Kendall correlation, two variables of interval/ratio or ordinal type.
- For Spearman correlation, two variables of interval/ratio or ordinal type.

PEARSON

1. Biased for small samples
2. When outlires are observed it may lead to wrong conclusions

$$\rho = corr(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^{N}(X_i - \bar{X})^2\right)}\sqrt{\left(\sum_{i=1}^{N}(Y_i - \bar{Y})^2\right)}}$$

SPEARMAN

1. Solution for outlires problem

$$\rho_s = corr(RX, RY)$$

KENDALL

1. Solution for outliers problem
2. It works better when many ranks are tied

$$\tau = 2\frac{P - Q}{N(N-1)}$$

P – concordant pairs (X1>X2) i (Y1>Y2) or(X1<X2) i (Y1<Y2)

Q – discordant pairs (X1>X2) i (Y1<Y2) or (X1<X2) i (Y1>Y2)

Association Tests – PEARSON, SPEARMAN, KENDALL

DATA:
- For Pearson correlation, two interval/ratio variables. Together the data in the variables are bivariate normal. The relationship between the two variables is linear.
- For Kendall correlation, two variables of interval/ratio or ordinal type.
- For Spearman correlation, two variables of interval/ratio or ordinal type.

PEARSON

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}} \sim t_{n-2}$$

Permutation, Exact and Fischer Transformation tests are also available

SPEARMAN

$$t = \rho_s\sqrt{\frac{n-2}{1-{\rho_s}^2}} \sim t_{n-2}$$

Permutation, Exact and Fischer Transformation tests are also available

KENDALL

$$Z_\tau = \frac{P-Q}{\sqrt{v}} \sim N(0,1)$$