Warsaw Summer School 2023, OSU Study Abroad Program

Contingency table
Nonparametric relations

Frequency cross-tabulation

- **Cross tabulation is the process of creating a contingency table from the bivariate frequency distribution of a statistical variables.**

- **Frequency f_{ij}**

Matrix form of cross-tabulation

Variable 1: Education	Variable 2: Political Party				
	PO 1	PIS 2	PLS 3	SLD 4	Total
Elementary 1		f12			f1.
Basic vocational 2					f2.
High school 3			f33		f3.
Some college 4					f4.
College 5					f5.
Total	f.1	f.2	f.3	f.4	f..

Basics

Observed data versus benchmarks for comparisons.

Assumptions for benchmarks:

(1) margins are given,

(2) values in the cells are limited $f_{.j}, f_{i.} \geq f_{ij} \geq 0$

Two basic benchmarks

a) Maximum association (the highest strength of the relationship)

b) No association (statistical independence)

Observed data

Religious Affiliation	Political Party Preference			
	Dem	Rep	Other	Total
• Cath	45	35	20	100
• Prot	35	55	10	100
• Total	80	90	30	200

Maximum strength of the relationship

Maximum strength: Preference to the largest row and column values

Religious Affiliation	Political Party Preference			
	Dem	Rep	Other	Total
• Cath	<u>80</u>	0	20	100
• Prot	0	<u>90</u>	10	100
• Total	80	90	30	200

Statistical independence

Religious Affiliation	Political Party Preference			
	Dem	Rep	Other	Total
• Cath				100
• Prot				100
• Total	80	90	30	200

Statistical Independence & Dependence

Statistical independence and dependence pertains to ideas that are applied to the relationships between events.

Two events are independent if the occurrence of one has no effect on the probability that the other will occur.

If two variables are independent of each other, knowing the value of one variable tells us nothing about the value of the other variable.

Expected frequencies for independence:

$$e_{ij} = (Tr * Tc) / Tt \quad = \quad (f_{i.} * f_{.j}) / f_{..}$$

where

Tr, $f_{i.}$ = total row

Tc, $f_{.j}$ = total column,

Tt, $f_{..}$ = overall total

Testing hypotheses

Could we test the hypothesis that our data are significantly different from the benchmark of statistical independence?

Yes. For this purpose we have to compute the overall distance between our data and the data of the benchmark of statistical independence.

Computing the index of dissimilarity is a good start but we do not know the sampling distribution of the values of this index.

For chi-square χ^2 statistics the sampling distribution is known. Then, we have to learn χ^2.

Chi square

Pearson chi square is the sum of squared differences between observed and expected frequencies, divided by the expected frequencies.

$$\chi^2 = \Sigma\, [(f_{ij} - e_{ij})^2 / e_{ij}]$$

Appendix Table provides critical values of χ^2 for a given df

For α =.05

and df = 1 $\chi^2 = 3.841$

and df = 2 $\chi^2 = 5.991$

and df = 5 $\chi^2 = 11.070$

	A	B	C	D	E	F	G	H	I	J
2										
3	**COMPUTING PEARSON CHI-SQUARE FOR A 2-BY-2 TABLE**									
4										
5			**Party * Sex Crosstabulation**							
6				**Sex**						
7				**Male**	**Female**	**Total**		(O - E)	$(O - E)^2$	$(O - E)^2/E$
8	**Party**	**Rep**	**Count**	10	5	15		2.5	6.25	0.833333
9			**Expected**	7.5	7.5			-2.5	6.25	0.833333
10								-2.5	6.25	0.833333
11		**Dem**	**Count**	5	10	15		2.5	6.25	0.833333
12			**Expected**	7.5	7.5					
13								SUM =		3.333333
14	**Total**		**Count**	15	15	30				
15			**Expected**	15	15			**This is the Pearson**		
16									**chi-square.**	
17			Note: Expected = (15*15)/30 = 7.5							

Testing

Chi-square for contingency tables serves to answers our initial question: Could we reject the null hypothesis that our data are randomly distributed within imposed margins?

Correction for small n

If for some f_{ij} the values are 0 or very close to 0, then

$$\chi^2 = \Sigma \left[(|f_{ij} - e_{ij}| - .5)^2 / e_{ij} \right]$$

Association

Association refers to coefficients which gauge the strength of a relationship.

Some measures are based on chi square statistic. χ^2 depends on the strength of the relationship and sample size. Various coefficients intend to eliminate the effect of the sample size.

Phi

The simplest measure, for 2x2 table, <u>Phi</u>.

Phi eliminates sample size by dividing chi-square by n, the sample size, and taking the square root.

$$\text{Phi} = \sqrt{(\chi^2 / n)}$$

Phi measures the strength of the relationship in terms of the concentration of cases on the diagonal.

Phi

Phi

10	**5**
5	**10**

expected 15x15/30 = 7.5 $\chi^2 = 3.33$

Phi = $\sqrt{(3.33/30)} = .33$

Chi square measures

Besides Phi, we have two basic chi-square-based measures of association: Cramer V, and Contingency Coefficient.

<u>Cramer's V</u> is the most popular of the chi-square-based measures of nominal association because it gives good values from 0 to 1 regardless of table size.

V equals the square root of chi-square divided by a product of:

- n [sample size]

- m [the smaller of (rows - 1) or (columns - 1)]

- $V = \sqrt{(\chi^2 / n*m)}$

Pearson C

The Contingency Coefficient, Pearson's C.

- **Intended to adapt Phi to tables larger than 2-by-2.**

C is equal to the square root of chi-square divided by chi-square plus n, the sample size:

$$C = \sqrt{\chi^2/(\chi^2 + n)}$$

C has a maximum approaching but never totally reaching 1.0 only for large tables and some researchers recommend it only for 5-by-5 tables or larger.

PER

There are also different measures of association for nominal variables.

These different measures are not based on χ^2 . Their logic: Proportional Reduction in Error (PER)

Lambda

Lambda, also known as *Goodman-Kruskal lambda*:

The proportion reduction in errors in predicting the dependent variable (DV) given knowledge of the modes of the independent variable (IV).

Lambda

Lambda = $[(SUM(d_i) - Td]/(n - Td)$

where $SUM(d_i)$ is a sum the largest f_{ij} of IV

Td = the largest marginal value of DV

n = sample size

Lambda

	Dem	Rep	Total
A	80	40	120
B	9	1	10
C	1	9	10
Total	90	50	140

Lambda = [(80 + 9 + 9) - 90]/(140-90) = .16

A, B, and C refers to different localities (Knowing these localities increases the guessing the distribution of Dem and Prot by 16%.

Other PER measures

Other measures based on PER:

(1) Goodman and Kruskal Tau – similar to lambda, but avoids a major shortcoming of lambda - giving 0 if there is some relationship between variables.

(2) The Uncertainty Coefficient, or Theil's U, also called the *entropy coefficient*, gives a proportionate reduction in error in terms of information theory.

Properties of PER measures

All PER measures have two important properties:

- **(1) they are directional: IV and DV should be specified. Computer programs give solutions for both possible arrangements of IV and DV (each called asymmetrical)**
- **(2) the sampling distribution of them is known and the test of significance is provided**

Ordinal level

Association for ordinal variables:

Gamma, Somers d, Kendall’s coefficients