STA4030: Categorical Data Analysis
Two-way Tables: Ordinal Data

Instructor: Bojun Lu

School of Data Science
CUHK(SZ)

October 13, 2020

Agenda

1. 5.1 Introduction
2. 5.2 Ordinal Measure of Association: Gamma
3. 5.3 Ordinal Measure of Association: Correlation
4. 5.4 A Test for ρ

5.1 Introduction

5.1.1 More Powerful Approaches

The tests of independence/association and measures of association we have discussed so far treat ordinal categorical data just the same as nominal categorical data.

This is a missed opportunity because, as we have noted before, there exist tests for ordinal data which are more powerful than the tests designed for nominal data.

In this chapter we will look at two ordinal measures of association and one test for a linear trend in the data.

5.1 Introduction

When we tested for independence before, X^2 and G^2 allowed for *any* kind of statistical dependence.

They need $(r-1)(c-1)$ degrees of freedom to do this - the alternate hypothesis of "some sort of dependence" needs that many parameters to describe any possible pattern.

Most ordinal tests require only one degree of freedom, because they are testing a particular type of association that can be summarized in one parameter.

5.1 Introduction

Previously, we focused on 2×2 tables, especially for measures of association.

If a categorical variable has just two categories, it is automatically an ordinal variable, as the mere fact the categories are different means that we could arbitrarily designate one as “high” and the other “low”.

Thus the advantage of ordinal tests only comes to the fore when we have $r \times c$ tables, with at least one of r, c larger than 2.

5.1 Introduction

5.1.2 Example: Job Satisfaction and Income

In the following table the variables are income and job satisfaction, measured for black males in a nationwide (U.S.) sample. Both classifications are ordinal.

Income (thousand dollars)	Job Satisfaction			
	Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
<15	1	3	10	6
15-25	2	3	10	7
25-40	1	6	14	12
>40	0	1	9	11

5.1 Introduction

The ordinary chi-squared statistics for testing independence are $X^2 = 6.0$ and $G^2 = 6.8$ with $df = 9$ (P = 0.74 and 0.66).

These statistics show little evidence of association between job satisfaction and income, but they ignore the ordering of rows and columns.

Would a more powerful test change our conclusion?

5.2 Ordinal Measure of Association: Gamma

5.2.1 Concordant and Discordant Pairs

When X and Y are ordinal, a monotone trend association is common. A single parameter can describe this trend. Some measures are based on classifying each pair of subjects as concordant or discordant.

A pair is *concordant* if the subject ranked higher on X also ranks higher on Y.
A pair is *discordant* if the subject ranking higher on X ranks lower on Y.

The pair is *tied* if the subjects have the same classification on X and/or Y.

5.2 Ordinal Measure of Association: Gamma

Two points: (X_i, Y_i), (X_j, Y_j)

1. Concordant pair:

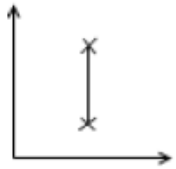
$$(X_i - X_j)(Y_i - Y_j) > 0$$

2. Discordant pair:

$$(X_i - X_j)(Y_i - Y_j) < 0$$

3. Tied in X:

$$(X_i - X_j)(Y_i - Y_j) = 0$$



4. Tied in Y:

$$(X_i - X_j)(Y_i - Y_j) = 0$$

5. Tied in X and Y:

$$(X_i - X_j)(Y_i - Y_j) = 0$$

5.2 Ordinal Measure of Association: Gamma

5.2.2 Example: Job Satisfaction and Income (Again)

- Job satisfaction has the following categories,
very dissatisfied (VD), little dissatisfied (LD),
moderately satisfied (MS), very satisfied (VS).
Consider a pair of subjects, one in the cell ($<$ 15, VD), and the other in the cell (15-25, LD). This pair is concordant.
- The subject in the cell ($<$ 15, VD) forms concordant pairs when matched with each of the three subjects classified in (15-25, LD). Thus these two cells provide $1 \times 3 = 3$ concordant pairs.
- The subject in the cell ($<$ 15, VD) is also part of a concordant pair when matched with each of the other (10+7+6+14+12+1+9+11) subjects ranked higher on both variables.

5.2 Ordinal Measure of Association: Gamma

Similarly, the three subjects in the (<15, LD) cell are part of concordant pairs when matched with the (10+7+14+12+9+11) subjects ranked higher on both variables.

The total number of concordant pairs, denoted by C, equals

$$\begin{aligned} C = {} & 1(3+10+7+6+14+12+1+9+11) \\ & + 3(10+7+14+12+9+11) + 10(7+12+11) \\ & + 2(6+14+12+1+9+11) + 3(14+12+9+11) \\ & + 10(12+11) + 1(1+9+11) + 6(9+11) + 14(11) = 1331. \end{aligned}$$

5.2 Ordinal Measure of Association: Gamma

The total number of discordant pairs, denoted by D, equals

$$\begin{aligned} D = {} & 3(2+1+0) + 10(2+3+1+6+0+1) \\ & + 6(2+3+10+1+6+14+0+1+9) \\ & + 3(1+0) + 10(1+6+0+1) + 7(1+6+14+0+1+9) \\ & + 6(0) + 14(0+1) + 12(0+1+9) = 849. \end{aligned}$$

$C > D$, suggesting a tendency for low income to occur with low job satisfaction and high income with high job satisfaction.

5.2 Ordinal Measure of Association: Gamma

5.2.3 Gamma

Given that a pair is untied on both variables:

$\Pi_c/(\Pi_c + \Pi_d)$ — the probability of concordance
$\Pi_d/(\Pi_c + \Pi_d)$ — the probability of discordance

The difference between these probabilities is

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d},$$

called *gamma*. The sample version is $\hat{\gamma} = \frac{C-D}{C+D}$.

Gamma treats the variables symmetrically. i.e., it is unnecessary to identify one classification as a response variable. Also, like correlation, gamma has the range $-1 \leq \gamma \leq 1$.

5.2 Ordinal Measure of Association: Gamma

A reversal in the category orderings of one variable causes a change in the sign of γ. Whereas the absolute value of the correlation is 1 when the relationship between X and Y is perfectly linear, only monotonicity is required for $|\gamma| = 1$, with $\gamma = 1$ if $\Pi_d = 0$ and $\gamma = -1$ if $\Pi_c = 0$. Independence implies that $\gamma = 0$, but the converse is not true. e.g., a U-shaped join distribution can have $\Pi_c = \Pi_d$ and hence $\gamma = 0$.

For the Job Satisfaction and Income data, $C = 1331$ and $D = 849$. Hence,

$$\hat{\gamma} = (1331 - 849)/(1331 + 849) = 0.221.$$

Only a weak tendency exists for job satisfaction to increase as income increases. Of the untied pairs, the proportion of concordant pairs is 0.221 higher than the proportion of discordant pairs.

5.3 Ordinal Measure of Association: Correlation

5.3.1 Pearson's ρ

You are already very familiar with one measure of association for ordinal data: ρ, (Pearson's) correlation.

Recall that ρ describes the strength of a *linear* trend in the population and is defined for interval data. Therefore, to calculate estimator r for ρ for a given contingency table, both X and Y must be ordinal and we must assign *scores* to their categories.

Let $u_1 \leq u_2 \leq \cdots \leq u_I$ and $v_1 \leq v_2 \leq \cdots \leq v_J$ denote scores for X and Y, respectively. The scores should reflect the distances between categories, with greater distances between categories regarded as farther apart.

5.3 Ordinal Measure of Association: Correlation

Let

$\bar{u} = \sum_i u_i p_{i+}$ — the sample mean of the row scores.

$\bar{v} = \sum_j v_j p_{+j}$ — the sample mean of the column scores.

$\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{ij}$ — sample covariance of X and Y.

Then $r = \dfrac{\sum_{i,j}(u_i-\bar{u})(v_j-\bar{v})p_{ij}}{\sqrt{\left[\sum_i (u_i-\bar{u})^2 p_{i+}\right]\left[\sum_j (v_j-\bar{v})^2 p_{+j}\right]}}$ is the sample correlation between X and Y.

Independence between X and $Y \Rightarrow \rho = 0$. The larger $|\rho|$ is, the farther the data fall from independence in the linear dimension.

5.3 Ordinal Measure of Association: Correlation

5.3.2 Example: Job Satisfaction and Income (Again, again)

To calculate r for this data, we need to assign scores to the categories. Let

$(v_1, v_2, v_3, v_4) = (1, 2, 3, 4)$ — scores for job satisfaction;

$(u_1, u_2, u_3, u_4) = (7.5, 20, 32.5, 60)$ — scores for income that approximate midpoints of categories.

We then have $r = 0.200$ suggesting a weak positive linear trend between job satisfaction and income.

5.4 A Test for ρ

5.4.1 Inference

Estimating ρ via r is informative, we would like some formal tests to support or contradict our impression of a linear trend. The necessary hypotheses are H_0: X and Y are independent vs H_1: $\rho \neq 0$ (two-sided)

and an appropriate test statistic is

$$M^2 = (n-1)r^2.$$

as for large n, M^2 has approximately a chi-squared distribution with $df = 1$. Large values contradict independence, so, the P-value is the right-tail probability above the observed value.

5.4 A Test for ρ

For a one-sided test, for instance to test whether there is a positive linear trend in the data, the hypotheses are

H_0: X and Y are independent vs H_1: $\rho > 0$ (one-sided)

and the appropriate test statistic is

$$M = \sqrt{n-1}r.$$

as for large n, M has approximately a standard normal distribution.

5.4 A Test for ρ

5.4.2 Example: Job Satisfaction and Income (Again[3])

We have already calculated the sample correlation. We quickly find the test statistic M^2:

$$r = 0.200 \quad \text{and} \quad M^2 = (96-1)(0.200)^2 = 3.81,$$

which shows some evidence of association ($P = 0.051$).

For the one-sided alternative $H_1 : \rho > 0$ (positive trend), $M = \sqrt{n-1}r = 1.95$, and $P = 0.026$. The evidence is stronger in support of a positive linear trend.

When a positive or negative trend exists, analyses designed to detect that trend can provide much smaller P-values (or much stronger evidences) than analyses that ignore it.

5.4 A Test for ρ

5.4.3 Example: Alcohol Use and Infant Malformation

The following table refers to a prospective study of maternal drinking and congenital malformations. After the first 3 months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on the presence or absence of congenital sex organ malformations.

Alcohol Consumption	Malformation		Total	Percentage Present
	Absent	Present		
0	17,066	48	17,114	0.28
< 1	14,464	38	14,502	0.26
1-2	788	5	793	0.63
3-5	126	1	127	0.79
≥ 6	37	1	38	2.63

5.4 A Test for ρ

Explanatory variable – Alcohol consumption, measured as the average number of drinks per day (ordinal).
Response variable – Malformation (absent or present, nominal).

When a variable is nominal but has only two categories, statistics (such as M^2 or M) that treat the variable as ordinal are still valid. E.g., one can treat “absent" as “low" and “present" as “high." Then, a choice of two scores, such as 0 for “absent" and 1 for “present," yields the same value of M^2 or M.

The percentage of malformation cases has roughly an increasing trend across the levels of alcohol consumption, which suggests a possible tendency for malformations to be more likely at higher levels of alcohol consumption.

5.4 A Test for ρ

Since the table has a mixture of very small, moderate, and extremely large counts, even though the sample size is large, in such cases the actual sampling distributions of X^2 or G^2 may not be close to chi-squared. Here, $df = 4$, $G^2 = 6.2$ ($P = 0.19$) and $X^2 = 12.1$ ($P = 0.02$) provide mixed signals.

To use M^2, we assign scores (midpoints of the categories) to alcohol consumption: $v_1 = 0$, $v_2 = 0.5$, $v_3 = 1.5$, $v_4 = 4.0$, $v_5 = 7.0$. Then, $r = 0.0142$, $M^2 = (32,573)(0.0142)^2 = 6.6$ ($P = 0.01$), suggesting strong evidence of a nonzero correlation. While $M = 2.56$ has $P = 0.005$ for testing $H_1 : \rho > 0$.