Categorical Data Analysis
Lecture Note 4

Instructor: Seokho Lee

Hankuk University of Foreign Studies

4. Two-Way Contingency Tables with More Structure

In many cases, the researcher is not interested in just determining that two categorical variables are independent, but would like to model a structure that incorporates some type of dependence. Interesting datasets are following:

- Tables with ordered categories
- Square tables
- Matched pairs

In this lecture we look at the first case and leave other two for future studies.

4.1. Testing Independence for Ordinal Data

The methods for two-way tables in Chapter 3 are most appropriate for nominal data. We now look at methods that are more powerful when data are ordinal.

4.1.1. Review of Linear Correlation

When *X* and *Y* are both **numerical** random variables, a measure of their linear relationships is the population covariance:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

The corresponding sample measure of their relationship is the sample covariance:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

A measure not affected by the units of measurement is the population correlation coefficient:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

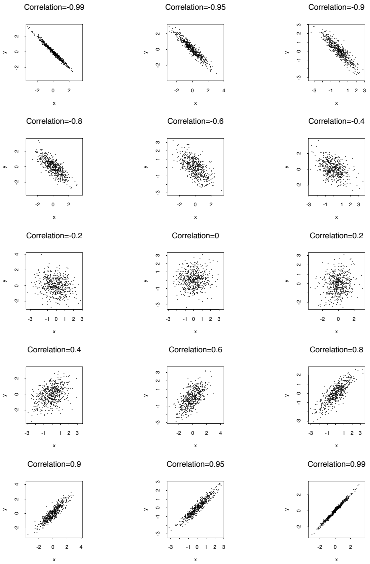
The corresponding sample measure of linear relations is the sample (Pearson's) correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Remarks:

1. $-1 \leq \rho \leq 1$
2. $\rho = \pm 1$ if all the distribution of (X, Y) is concentrated on a straight line
3. ρ near zero indicates no linear relationship
4. $\rho > 0$ indicates that Y has a tendency to increase as X increases
5. $\rho < 0$ indicates that Y has a tendency to decrease as X increases
6. r has a similar interpretation for the scatter plot of n (x, y) pairs

Random Samples from Bivariate Normal Distributions



Correlation=-0.99

Correlation=-0.95

Correlation=-0.9

Correlation=-0.8

Correlation=-0.6

Correlation=-0.4

Correlation=-0.2

Correlation=0

Correlation=0.2

Correlation=0.4

Correlation=0.6

Correlation=0.8

Correlation=0.9

Correlation=0.95

Correlation=0.99

Testing for a Linear Relationship

For statistical inference we assume that (X, Y) has a bivariate normal distribution with parameters $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$.

Consider testing the hypotheses:

$$H_0 \ : \ \rho = 0 \text{ no linear relationship}$$

$$H_a \ : \ \rho \neq 0 \text{ some linear relationship}$$

Test Statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Rejection Region:

$$|t| > t_{\alpha/2}(n-2)$$

4.1.2. Rank Correlation Coefficients

Pearson's correlation coefficient measures the linearity of the relationship between two numerical variables. A rank correlation coefficient measures whether one variable tends to increase (or decrease) as the other increases, even when the relationship is not linear.

- **Spearman's rank correlation:** Compute Pearson's r using the ranks of the observations.
    - Obtain the ranks of $x_1, \cdots, x_n : r_1, \cdots, r_n$
    - Obtain the ranks of $y_1, \cdots, y_n : s_1, \cdots, s_n$
    - Compute $r_S =$ Pearson's correlation coefficient computed on the pairs: $(r_1, s_1), \cdots, (r_n, s_n)$
    - $r_S = 1$ indicates that $x_i < x_j$ whenever $y_i < y_j$
    - $r_S = -1$ indicates that $x_i < x_j$ whenever $y_i > y_j$
- **Kendall's τ:** Determine the proportion of pairs where they are concordant or discordant.
    - A pair of observations (x_i, y_i), (x_j, y_j) is concordant if $(x_i - x_j)(y_i - y_j) > 0$
    - A pair of observations (x_i, y_i), (x_j, y_j) is discordant if $(x_i - x_j)(y_i - y_j) < 0$
    - Let $C =$ the number of concordant pairs
    - Let $D =$ the number of discordant paris
    - Kendall's $\tau = \frac{C-D}{n(n-1)/2}$
    - $\tau = 1$ indicates that $x_i < x_j$ whenever $y_i < y_j$

Example: The following data represent a tutor's scores of five clinical psychology students as to their suitability for their career and their knowledge of psychology:

Career (x)	2	0	4	3	11
Knowledge (y)	12	13	14	17	19

Spearman's rank correlation:

1. Obtain ranks:

Career (r)	4	5	2	3	1
Knowledge (s)	5	4	3	2	1

2. Compute Pearson's correlation between these ranks:

$$\sum_{i=1}^{5}(r_i - \bar{r})^2 = \sum_{i=1}^{5} r_i^2 - 5\bar{r}^2 = 55, \qquad \sum_{i=1}^{5}(s_i - \bar{s})^2 = 55$$

$$\sum_{i=1}^{5}(r_i - \bar{r})(s_i - \bar{s}) = \sum_{i=1}^{5} r_i s_i - 5\bar{r}\bar{s} = 53 - 5(3)(3) = 8$$

$$r_S = \frac{\sum_{i=1}^{5}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{5}(r_i - \bar{r})^2 \cdot \sum_{i=1}^{5}(s_i - \bar{s})^2}} = \frac{8}{\sqrt{55^2}} = 0.1455$$

Kendall's τ:

1. Observe that $C = 8$ and $D = 2$ (why?)

2. Compute

$$\tau = \frac{C - D}{5 \cdot 4/2} = 0.6$$

3. Kendall's τ follows an asymptotic normal distribution with mean zero and variance $2(2n + 5)/9n(n - 1)$. So you can do inference on this when the sample size is large.

4.1.3. Tests and Measures of Association for $I \times J$ Contingency Tables

- The previous tests and measures of association treated both categorical variables X and Y as nominal
- If either the rows or columns are ordinal, test statistics that use the order in X and Y are more appropriate
- If the data (X and Y) have the interval numerical scale (scores that are evenly spaced), the Pearson correlation coefficient is an appropriate measure of association
- If the data do not have an obvious scale, but both X and Y are ordinal in nature, the following measures are appropriate:
  - The Spearman rank coefficient is obtained by replacing the variables with their ranks in the Pearson correlation coefficient
  - Other measurements are based on the number of concordant and discordant pairs in the data

    A pair is **concordant** if the category ranking higher on the row variable also ranks higher on the column variable.

    A pair is **discordant** if the category ranking higher on the row variable also ranks lower on the column variable.
  - The gamma, Kendall's Tau-b, Stuart's Tau-c, and Somer's D statistics are all based on concordant and discordant pairs.

    These measures take values between -1 and 1

    These measures differ mainly in their strategies for adjusting for ties and sample sizes

- To test for linear trends, one must assign **scores** to the categories.

  The scores should be **monotone**, that is, have the same ordering as the category levels.

  The scores represent the distances between the categories. For Spearman's correlation, the scores are 1,2,3, etc.

  Let $u_1 \leq u_2 \leq \cdots \leq u_n$ denote the scores for the rows.

  Let $v_1 \leq v_2 \leq \cdots \leq v_n$ denote the scores for the columns.

row\column scores	v_1	v_2	$\cdots$	v_J	Total
u_1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
u_2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

We use the alternative formula for Pearson's correlation between X and Y:

$$r = \frac{\sum\sum_{i,j=1}^{n} u_i v_j n_{ij} - (\sum_{i=1}^{n} u_i n_{i+})(\sum_{j=1}^{n} v_j n_{+j})/n}{\sqrt{\left(\sum_{i=1}^{n} u_i^2 n_{i+} - \frac{(\sum_{i=1}^{n} u_i n_{i+})^2}{n}\right)\left(\sum_{j=1}^{n} v_j^2 n_{+j} - \frac{(\sum_{j=1}^{n} v_j n_{+j})^2}{n}\right)}}$$

- A statistic for testing the null hypothesis of independence against the two-sided alternative of nonzero correlation is

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

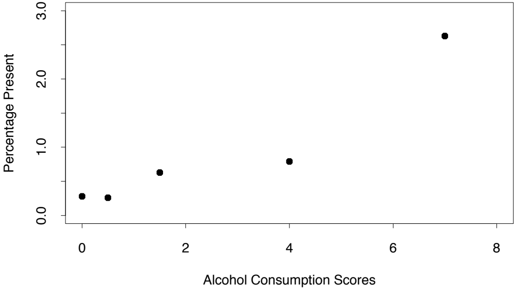
  - For large samples, M^2 has approximately a chi-squared distribution with $df = 1$. We reject independence for large value of M^2
  - This statistic is sensitive to positive (or negative) trends in the relationship of two ordinal categorical variables. When there is truly a “linear” trend, M^2 will tend to have more powerful than the Pearson and likelihood ratio chi-squared test, which are designed for a general alternative with $df = (r-1)(c-1)$
- Choice of Scores – For most data sets, the choices of scores has little effect on the results. However, if the data are unbalanced, the choice of monotone scores can affect the value of M^2.
  - Integer Scores: These are useful when the ordered categories can be considered as equally spaced
  - Rank-based Scores: For a two-way table, the use of ranks and midranks produces Spearman’s rank correlation coefficient
  - It is sometimes useful to carry out a *sensitive analysis* to see how sensitive the conclusion are to the choice of scores

Example: Infant Mortality Data from Agresti, Chapter 2

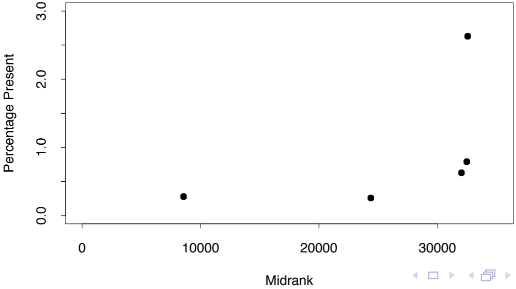
A prospective study was carried out of maternal drinking and congenital malformations. A survey was completed on alcohol consumption at three weeks of pregnancy. Following childbirth observations were made concerning the presence or absence of congenital sex organ malformations.

Alcohol Consumption	Malformation Absent	Malformation Present	Total
0	17066	48	17144
1	14464	38	14502
$1-2$	788	5	793
$3-5$	126	1	127
≥ 6	37	1	38

Infant Malformation and Mother's Alcohol Consumption



Plot Using Midranks



4.1.4. Trend Tests for $I \times 2$ and $2 \times J$ Tables

We suppose X is an explanatory variable and Y is a response variable.

- $2 \times J$ Tables: Tables like this represent two groups where the rows represent two treatments.
  The M^2 statistic is directed toward detecting differences in the row mean scores. When we use the midrank scores, the test is sensitive to testing differences in average ranks for the two rows. This is a form of the *Wilcoxon or Mann-Whitney test*.
- $I \times 2$ Tables: The response variable is binary. Here we look for a trend in the proportion of success across the rows. This test is called the *Cochran-Armitage trend test*.

4.1.5. Some ordinal measures of association on SAS output

- Gamma $= \hat{\gamma} = \frac{C-D}{C+D}$
- Kendall's Tau-b (τ_b) is similar except it uses a correction for ties
- Stuart's Tau-c (τ_c) also makes an adjustment for table size
- Somer's D (C|R) is an asymmetric modification of Tau-b where the column variable is the dependent or response variable
- Somer's D (R|C) is an asymmetric modification of Tau-b where the row variable is the dependent or response variable
- Gamma will have the largest value of these measures