K300 (4392) Statistical Techniques (Fall 2007)

Lecture Note: Chi-square Test

Instructor: Hun Myoung Park

1. Categorical Data Analysis

A categorical variable is measured in either nominal or ordinal level of measurement. They are discrete but not all discrete variables are not categorical variables. Event count data (e.g., the number of patients arrived at a prompt care in an hour) are discrete but not categorical. In a discrete variable, there is a limited number of values between any arbitrary values (there are only two categories—sophomore and junior—between freshmen and senior; no other category between male and female).

Why should categorical variables be treated in a different way compared to interval and ratio scaled variables? This is because we cannot use arithmetic operators in categorical variables. DO NOT compute the mean of department name (e.g, SPEA, Kelley School, SLIS, and IU School of Medicine) or divide Junior by Sophomore. When you assign some numbers to these categories (e.g., 1=SPEA, 2=Kelley… 12=Junior, 24=Sophomore…), you should know that these numbers are not interpretable. In other word, you should get the same result when you assign totally different numbers (e.g., -100=SPEA, 3.15=Kelley… .007=Junior, 1M=Sophomore…). You just can check equality in a nominal variable and compare the order in an ordinal variable.

Level of Measurement and Categorical Data

	Order	Difference	Type	Operator
Nominal			Discrete	=
Ordinal	X		Discrete	=, <, >
Interval/Ratio	X	X	Continuous	=, <, >, +, -, *, /

2. Frequency Table, Contingency Table.

The best way to summary the information of categorical variables is to draw a frequency table for a categorical variable or a contingency table of two categorical variables. These tables contain the number of observations (cases) that falls into each category. Of course, observations should be *mutually exclusive* and *collectively exhaustive*. Any one observation should not fall into more than one category and all observations should be fall into any categories. For example, if you fall into both `female` and `male`, you are violating *mutual exclusiveness*. If you are neither `male` nor `female` and thus do not fall into any category, this is the violation of *collective exhaustiveness*.
See the following frequency table.

```
       male |      Freq.     Percent        Cum.
------------+-----------------------------------
     female |         14       63.64       63.64
       male |          8       36.36      100.00
------------+-----------------------------------
      Total |         22      100.00
```

We know that there are 14 female students in the class (first column), who account for 63.6 percent of the total 22 students (second column). The last column reports the cumulative percentage (cumulative relative frequency)

Now consider two categorical variables. This contingency table (cross-table) appeared in the midterm exam. In a cell, the first number is the frequency and the second number is the row percentage of the frequency. For example, 1,526 is the number of observations who use both the Internet and email, while 90.73 is the percent of the row total (90.73=1,526/1,682)

```
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
```

Internet	Emails (Yes) 1	2	Total
(Yes) 1	1,526	156	1,682
	90.73	9.27	100.00
2	44	833	877
	5.02	94.98	100.00
Total	1,570	989	2,559
	61.35	38.65	100.00

Similarly, you may put column percentage and cell percentage. For instance, 97.20 in the first cell is 1,526/1,570 and 59.63 is 1,526/2,559.

```
+------------------+
| Key              |
|------------------|
|     frequency    |
|  row percentage  |
| column percentage |
|  cell percentage |
+------------------+
```

row	col 1	2	Total
1	1,526	156	1,682
	90.73	9.27	100.00
	97.20	15.77	65.73
	59.63	6.10	65.73
2	44	833	877
	5.02	94.98	100.00
	2.80	84.23	34.27
	1.72	32.55	34.27

```
   Total |     1,570         989 |     2,559
         |     61.35       38.65 |    100.00
         |    100.00      100.00 |    100.00
         |     61.35       38.65 |    100.00
```

3. Chi-square Test for Goodness-of-fit: Frequency Table

Goodness-of-fit test checks whether sample is drawn from a particular distribution. $\sum_{i=1}^{r}\frac{(n_i - E_i)^2}{E_i} \sim \chi^2[(r-1)]$, where n_i and E_i are observed and expected (hypothesized) frequencies of cell *i*. In the gender above, the expected frequency is 11 (50 percent of the total) when assuming no gender discrimination and barrier in enrollment. This test compares observed and expected frequencies and examines the difference. If the difference is large (observed frequency is far away from the expected frequency), we can reject the null hypothesis of good goodness-of-fit. See the following example.

	Observed	Expected	Difference	Difference^2	Difference^2/Expected
Female	14	11	3	9	.8182
Male	8	11	-3	9	.8182

$$\sum_{i=1}^{r}\frac{(n_i - E_i)^2}{E_i} = \frac{(14-11)^2}{11} + \frac{(8-11)^2}{11} = 1.636 \sim \chi^2[2-1]$$

TS 1.636 is smaller than the CV of 3.841 (df=1), so do not reject the null hypothesis of goodness-of-fit. Note that the degrees of freedom is *r-1*.The distribution of male and female fits the hypothesized distribution (uniform distribution); Having 14 female students (64 percent) in the class is likely even when assuming equal chance that male and female enroll the class.

Example 11-1 on page 565-567.

	Cherry	Strawberry	Orange	Lime	Grape
Observed	32	28	18	14	10
Expected	20	20	20	20	20

$$\sum_{i=1}^{r}\frac{(n_i - E_i)^2}{E_i} = \frac{(32-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(10-20)^2}{20} = 18 \sim \chi^2[5-1]$$

Since TS 18 is greater than the CV 9.488, reject the null hypothesis.

4. Chi-square Test for Independence: Contingency Table

Karl Pearson chi-square, χ^2, test examines independency of two categorical variables on the basis of a contingency table (cross-table). This is the typical type of chi-square test. The null hypothesis of this test is that two variables are independent. $H_0 = (\theta_{11}, \theta_{12}, ..., \theta_{1c}) = (\theta_{21}, \theta_{22}, ..., \theta_{2c}) = ... = (\theta_{r1}, \theta_{r2}, ..., \theta_{rc})$. The degrees of freedom are (*r-1*)(*c-1*), where *r* and *c* represent the number of categories in the row and column, respectively. *The chi-square test is not reliable when the expected frequency of any cell is less then five.*

$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2[(r-1)(c-1)]$, where $E_{rc}=\frac{n_{r*}n_{*c}}{N}$ and $E_{rc}\geq 5$

See the example on pages 577-580. We have a 2 by 3 contingency table for observed frequencies as follows.

row	col 1	2	3	Total
1	100	80	20	200
2	50	120	30	200
Total	150	200	50	400

We need to have a corresponding contingency table for expected frequency.

row	col 1	2	3	Total
1	75.0	100.0	25.0	200.0
2	75.0	100.0	25.0	200.0
Total	150.0	200.0	50.0	400.0

The expected frequency of row *r* and column *c* is $E_{rc}=\frac{n_{r*}n_{*c}}{N}$, where n_{r*} is the sum of *r*th row and n_{*c} is the sum of *c*th column. For example, the expected frequency of row 1 and column 1 is 75, $E_{11}=\frac{n_{1*}n_{*1}}{N}=\frac{200\times 150}{400}=75$. Similarly, $E_{23}=\frac{n_{2*}n_{*3}}{N}=\frac{200\times 50}{400}=25$. Let us put observed and expected frequencies together.

row	col 1	2	3	Total
1	100	80	20	200
	75.0	100.0	25.0	200.0
2	50	120	30	200
	75.0	100.0	25.0	200.0
Total	150	200	50	400
	150.0	200.0	50.0	400.0

Now let us compute the chi-square statistic using $\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2[(r-1)(c-1)]$.

$$\frac{(100-75)^2}{75}+\frac{(80-100)^2}{100}+\frac{(20-25)^2}{25}+\frac{(50-75)^2}{75}+\frac{(120-100)^2}{100}+\frac{(30-25)^2}{25}=26.67.$$

The degrees of freedom is 2=(2-1)×(3-1). The critical value for the 2 degrees of freedom at the .05 level is 5.991. Since TS 26.67 is lager than CV 5.991, reject the null hypothesis of independence of two categorical variables.

Here is summary of Chi-square test for independence

1. Null hypothesis: variable 1 is independent of variable 2, alternative hypothesis: variable 1 is not independent of variable 2.
2. Determine the significance level (.05), compute degrees of freedom, and get the corresponding critical value from the table.
3. Compute chi-square statistic. Compute expected frequency. Examine if all expected frequencies are lager than 5. Otherwise, stop the chi-square test.
4. Reject the null hypothesis if chi-square is larger than the critical value
5. Draw conclusion.

Karl Pearson chi-square test does not tell anything about the extent that row and column are connected. It just examines whether two variables are independent or not. If two variables are not independent of each other, compute measures of association. If both categorical variables are ordinal, compute gamma. Otherwise (if one or two variables are nominal), compute lambda.

5. Measure of Association: Gamma γ for Ordinal Variables

Goodman and Kruskal's Gamma γ is an association indicator of ordinal scale variables. This symmetric measure ranges from -1 (negative relationship) to 1 (positive relationship). Zero means independence of two ordinal variables. The Gamma means percentage reduction in errors of predicting the rank of variable 2 when knowing variable 1.

$\gamma=\frac{C-D}{C+D}$, where C and D respectively represent the sum of product of concordant and discordant pairs. Concordant (positive relationship) and discordant (negative relationship) pairs are determined by the indices of the cells without allowing ties. Both indices (row and column) of concordant pairs are increasing or decreasing (e.g., $n_{11}n_{22}$), while discordant pairs have one increasing and one decreasing indices (e.g., $n_{13}n_{21}$). Tied pairs have the same index in either row or column (e.g., $n_{11}n_{13}$ and $n_{13}n_{23}$)

In a 2 by 2 contingency table, $C = n_{11}n_{22}$ and $D = n_{12}n_{21}$. The gamma for a 2 by 2 contingency table is also called Yule's Q.

In a 2 by 3 table, $C = n_{11}(n_{22}+ n_{23})+ n_{12}n_{23}$ and $D = n_{13}(n_{21}+n_{22})+n_{12}n_{21}$.

In a 3 by 3 table, $C = n_{11}(n_{22}+n_{23}+n_{32}+n_{33}) + n_{12}(n_{23}+n_{33}) + n_{21}(n_{32}+ n_{33})+n_{22}n_{33}$, and $D = n_{13}(n_{21}+n_{22}+n_{31}+n_{32}) + n_{12}(n_{21}+n_{31}) + n_{23}(n_{31}+n_{32}) + n_{22}n_{31}$.

Example 11-5 on pages 580-581. Education and location (urban, suburban, and rural) can be considered as ordinal variables.

row	col 1	2	3	Total
1	15	12	8	35
	11.5	13.9	9.5	35.0
2	8	15	9	32
	10.5	12.7	8.7	32.0
3	6	8	7	21
	6.9	8.4	5.7	21.0
Total	29	35	24	88
	29.0	35.0	24.0	88.0

Pearson chi2(4) = 3.0055 Pr = 0.557
gamma = 0.1914 ASE = 0.146

Let us first compute chi-squar statistic for independence of two ordinal variables.

$$\frac{(15-11.5)^2}{11.5}+\frac{(12-13.9)^2}{13.9}+\frac{(8-9.5)^2}{9.5}+\frac{(8-10.5)^2}{10.5}+\frac{(15-12.7)^2}{12.7}+\frac{(9-8.7)^2}{8.7}+$$
$$\frac{(6-6.9)^2}{6.9}+\frac{(8-8.4)^2}{8.4}+\frac{(7-5.7)^2}{5.7}=3.01$$

Degrees of freedom is 4=(3-1)(3-1). The critical value is 9.488. Do not reject the null hypothesis; two variables are independent. In fact, it is useless to compute the measure of association for two independent varialbes. But let us do for an exercise.

row	col 1	2	3	Total
1	15	~~12~~	~~8~~	35
2	~~8~~	15	9	32
3	~~6~~	8	7	21
Total	29	35	24	88

Let us being with n_{11} to compute C (for concordant pairs). 15(15+9+8+7) Notice all four cells has larger indices than n_{11}.

row	col 1	2	3	Total
1	~~15~~	12	~~8~~	35
2	~~8~~	~~15~~	9	32
3	~~6~~	~~8~~	7	21
Total	29	35	24	88

12(9+7). n_{13} does not have concordant pairs whose indices are larger than those of n_{13}; let us skip.

row	col 1	2	3	Total
1	~~15~~	~~12~~	~~8~~	35
2	8	~~15~~	~~9~~	32
3	~~6~~	8	7	21
Total	29	35	24	88

In the second row, 8(8+7)

row	col 1	2	3	Total
1	~~15~~	~~12~~	~~8~~	35
2	~~8~~	15	~~9~~	32
3	~~6~~	~~8~~	7	21
Total	29	35	24	88

15(7). The row 3 does not have any concordant pairs.

Therefore, C=15(15+9+8+7)+12(9+7)+8(8+7)+15(7)=1002

Let us do the reverse way to compute D (for discordant pairs) For example,

row	col 1	2	3	Total
1	~~15~~	~~12~~	8	35
2	8	15	~~9~~	32
3	6	8	~~7~~	21
Total	29	35	24	88

For n_{13}, 8(8+15+6+8)

D=8(8+15+6+8)+12(8+6)+9(6+8)+15(6)=680

$\gamma = \frac{C-D}{C+D} = \frac{1002-680}{1002+680} = .1914$. This gamma indicates a slightly positive relationship which is not statistically discernable (p<.557)

6. Measure of Association: Lambda λ for Nominal Variable(s)

Goodman-Kruskal Lambda is the representative measure of association for nominal data. The Lambda is PRE (Proportional Reduction in Error) measure indicating the percentage

reduction in errors in predicting variable 2 when knowing information of variable 1. It is an asymmetric measure that depends on the predictor (row or column).

The Lambda ranges from 0 to 1, where zero means knowing first variable is not helpful in predicting second variable. Lambda close to 1 indicates that the information (knowing distribution of the predictor) is able to help reduce error substantially.

Lambda is computed as $\lambda_{c|r} = 1 - \frac{\sum_{i=1}^{r}[n_{i*} - Max(n_{ij})]}{N - Max(n_{*j})} = \frac{\sum_{i=1}^{r} Max(n_{ij}) - Max(n_{*j})}{N - Max(n_{*j})}$, where n_{i*} is the marginal row frequency of the *i*th row, $Max(n_{ij})$ is the largest frequency of the *i*th row, and $Max(n_{*j})$ is the largest marginal column frequency.

See the example on pages 577-580. $Max(n_{1j})$ is the largest frequency in the first row, which is 100.

row	col 1	2	3	Total
1	100	80	20	200
2	50	120	30	200
Total	150	200	50	400

$Max(n_{2j})$ is 120

row	col 1	2	3	Total
1	100	80	20	200
2	50	120	30	200
Total	150	200	50	400

$Max(n_{*j})$ is the largest marginal column frequency, which is 200.

row	col 1	2	3	Total
1	100	80	20	200
2	50	120	30	200
Total	150	200	50	400

The lambda is $\lambda_{c|r} = \frac{\sum_{i=1}^{r} Max(n_{ij}) - Max(n_{*j})}{N - Max(n_{*j})} = \frac{(100+120) - 200}{400 - 200} = .1$. Although two nominal variables are not independent, they have a marginal association; knowing variable 1 does not help substantially reduce error occurring when predicting variable 2.

Flow Chart of Chi-square Test and Measure of Association