EXST 7005

Fall 2010

Lab #6: Chi-Squared Test of Independence

Chi-Squared Tests of Independence

In the previous labs, we mainly talked about how to get statistics for numerical data. Today we are going to learn how to deal with categorical data. Among the statistics dealing with categorical data, the chi-square test is the most frequently used. The chi-square test provides a method for testing the association between the row and column variables in a two-way table. The null hypothesis H_0 is that there is no association between the two variables. In other words, one variable does not vary according to the other variable. By contrast, the alternative hypothesis H_a claims that some association does exist although it does not specify the type of association.

The chi-square test is based on a test statistic that measures the divergence of the observed data from the values expected under the null hypothesis. The expected value for each cell in a two-way table is calculated through (row total*column total)/n, where n is the total number of observations in the table:

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

The distribution of the statistic X^2 is chi-square with $(r-1)(c-1)$ degrees of freedom, where r represents the number of rows and c represents the number of columns.

Example: We want to test whether there is a relationship between the type of bus and arriving on time. Then there are two variables we need to take into consideration: type of bus (E for express; R for regular) and promptness (L for late and O for on time). We have then collected 50 observations as follows (each line contains 10 observations):

```
E O E L E L R O E O
E O E O E O R L R O
R L E O R L E O R L
R O E O E O R L E L
E O R L E O R L E O
R O E L E O E O E O
E O E O E O E L E O
E O R L R L R O R L
E L E O R L R O E O
E O E O E O E L R O
R L
```

The following program runs PROC FREQ to test the independence of the two variables:

```
proc freq data=BUS;
tables BUSTYPE*ONTIME/chisq;
run;
```

The following table shows results from chi-squared test of independence. The degree of freedom is (rows-1) *(columns-1), in this case, 1*1=1. The probability is smaller than .001 which suggests that we should reject the null hypothesis. In other words, the bus type and punctuality are not independent of each other.

Statistic	DF	Value	Prob
Chi-Square	1	7.2386	0.0071
Likelihood Ratio Chi-Square	1	7.3364	0.0068
Continuity Adj. Chi-Square	1	5.7505	0.0165
Mantel-Haenszel Chi-Square	1	7.0939	0.0077

SAS also creates a table with variable 1 as the row and variable 2 as the column. The values represent the number of times the value was observed, percentage of observations, percentage of that row's observations, and percentage of that column's observations.

Table of BUSTYPE by ONTIME			
BUSTYPE	ONTIME		
Frequency Percent Row Pct Col Pct	L	O	Total
E	7	22	29
	14.00	44.00	58.00
	24.14	75.86	
	35.00	73.33	
R	13	8	21
	26.00	16.00	42.00
	61.90	38.10	
	65.00	26.67	
Total	20	30	50
	40.00	60.00	100.00

Assignment

1. The following table indicates the number of animals that survived a treatment. Calculate by hand to test whether the treatment influences animals' survival. Show the steps.

	Dead	Alive	Total
Treated	36	14	50
Not treated	30	25	55
Total	66	39	105

2. This following dataset is about the employers demographics of a local company. The main variables are gender (0=male, 1=female), pay (1=low, 2=intermediate, 3=high), education (1=college degree, 2=graduate degree), and year of hire. Conduct Chi-squared tests and then use the output to answer the following questions. You are also expected to interpret the results in this scenario. Use 0.05 as the significance level for all the tests.

id	sex	pay	edu	yoh
1	0	2	2	2
2	1	3	2	1
3	1	2	1	1
4	0	2	1	2
5	1	2	1	2
6	1	1	1	3
7	1	3	2	2
8	0	1	2	3
9	0	1	1	2
10	0	1	1	2
11	1	3	2	1
12	0	2	2	2
13	1	2	1	2
14	1	3	2	2
15	0	2	2	2
16	0	1	1	3
17	1	1	1	2
18	0	1	1	3
19	0	2	2	2
20	1	1	1	3
21	1	2	1	2
22	0	3	2	1
23	0	1	1	2
24	0	2	2	3
25	1	1	1	2

1) Does the amount of pay differ between genders?
2) Does the amount of pay differ between education levels?
3) Will an employer get more pay than his coworker who has less years of hire than him/her?