Application of Statistics in Educational Research I

MPU1034

Nonparametric Statistics*: Chi-Square

Prof. Dr. Mohd Salleh Abu

Dr. Hamidreza Kashefi

main source: Vernoy & Vernoy (1997)

Nonparametric statistics are tests that do not compare population parameters and make fewer assumptions than parametric statistics

In this chapter, we dealing with hypothesis testing involving nominal data

Some Commonly Used Jargons…

- Nominal Data
- Nonparametric Statistics
- Chi-Square (χ^2)
- Observed Frequency (f_o)
- Expected Frequency (f_c)
- Cells

The Needs of χ^2

- ☑ *t* Tests and ANOVA are applicable for data which take form in either interval or ratio.
- ☑ *t* Tests and ANOVA are not applicable for nominal data.
- ☑ Two examples where ANOVA are not applicable are as follows:

Eg. 1: Comparing The Number of Deaths of Population in Three States

Nos. of Deaths Per Thousand Population for 3 States

State 1	**State 2**	**State 3**
45	34	20

where the numbers are treated as NOMINAL scales; they only represent a list of numbers of people who fit into each category (ie. the state)

Eg. 2: Comparing The Professional Integrity (IP) Trait of School Principals Serving in Various Categories of Secondary Schools

Level of IP	Category of Schools		
	Not Effective	Effective	Very Effective
Low	5	7	0
Average	9	14	2
High	31	24	43

where all the numbers are in the NOMINAL scales; they only represent a list of numbers of people who fit into each cells

☑ In each of the cases mentioned above (where data involved are in nominal form), the comparison can be done using chi-square (denoted by χ^2)

The General Idea About χ^2

- ✓ Chi-square is a statistical technique that enables us to compare the observed frequencies in different categories (the actual numbers obtained) with the frequencies expected from some theory or hypothesis.
- ✓ We represent an observed frequency with the symbol f_o and we represent an expected frequency with the symbol f_e
- ✓ The formula for is χ^2 is

$$\boxed{\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}}$$

- ✓ Note that chi-square requires null hypothesis (as well as alternative hypothesis) to enable us to compute the f_e

Essential Steps in Performing χ^2

Setting The Hypotheses

Finding f_o, f_e *and* χ^2

Finding Degree of Freedom

Finding Critical χ^2 Value

Drawing up a Conclusion

Eg. 1 (One-Way)

Suppose you are required to investigate the rate of deaths of population in three states namely, State A, State B and State C. The number of deaths per one thousand population in each state are as follows:

State A	State B	State C
45	34	20

Run a χ^2 test to see whether the difference of the death rates is significant at $\alpha = 0.05$

Setting the Hypothesis

H_o : There is no difference in the number of deaths per every one thousand population in each state (i.e. the death rates for each state are all equal)

H_1 : There is at least one state with a different death rate per every one thousand pupulation

Finding f_o, f_e and χ^2

- ✓ Note that if H_o were true (i.e. the death rates were the same for each state), we would expect there would be 33 deaths (out of 99 deaths) in each state. Thus, the expected frequency f_e for each state would be 33.

- ✓ We can now create a table showing the values of f_o and f_e for each category as follows:

	State A	State B	State C
f_o	45	34	20
f_e	33	33	33

Computation of χ^2

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(45 - 33)^2}{33} + \frac{(34 - 33)^2}{33} + \frac{(20 - 33)^2}{33}$$

$$= \frac{12^2}{33} + \frac{1^2}{33} + \frac{-13^2}{33} = \frac{144}{33} + \frac{1}{33} + \frac{169}{33}$$

$$= 4.364 + 0.030 + 5.121$$

$$= 9.515$$

Finding Degree of Freedom (*df*)

- ✓ For one-way case, the degree of freedom is given by

  df = the number of categories - 1

In this case (refer to table),

$df = 3 - 1 = 2$

Finding The Critical Value of χ^2

Table 15.3 A Portion of Table χ, Critical Values of the χ^2 Statistic

df	.25	.10	.05	.025	.01	.005
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.071	12.833	15.086	16.750

Source: Owen, 1962.

From the Table χ with df = 2, α = 0.05, we would get the critical value of χ^2 equal to 5.991

Drawing a Conclusion

Comparing the computed value of $\chi^2 = 9.515$ and the critical value of $\chi^2 = 5.991$, we can safely reject H_o (i.e. there are at least two states with different death rates!)

Eg. 2 (Two-Way)

It was claimed that smoking is the single most preventable cause of death and disease in the developed world. So let's examine the relationship between cancer deaths and smoking behavior as well as social support networks. Now our table should look like Table 15.4, which is a 2 x 3 table with two rows and three columns. Note that we are not only dealing with the different types of social networks of the women, but we have further identified each woman by whether or not she smoked.

Table 15.4 Number of Deaths Per Thousand Women for Three Types of Social Networks and for Smokers and Nonsmokers

	Type of social network			
	Poor	Moderate	Good	Total
Smoker	25	26	18	69
Nonsmoker	20	8	2	30
Total	45	34	20	99

Run a χ^2 test to see whether the difference of the number of deaths over the types of social networks and smoking habit is significant at $\alpha = 0.05$

Setting the Hypothesis

H_o : There is no difference of death rates over the different types of social networks and smoking habit among the women

H_1 : There is at least one combination (involving the different types of social networks and smoking habit of the women) with different death rate as compared to the others

Finding f_o , f_e and χ^2

- ✓ Note that there are two rows and three columns which together make up six cells. Therefore, there would be are six combination of categories which require six different f_e to be fitted in.
- ✓ For each combination of categories, the f_e are computed using the following rule:

$$f_e = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}}$$

for column 1, row 1

$$f_e = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}} = \frac{69 \cdot 45}{99} = \frac{3105}{99}$$
$$= 31.36$$

for column 2, row 1

$$f_e = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}} = \frac{30 \cdot 45}{99} = \frac{1350}{99}$$
$$= 13.64$$

for column 1, row 2

$$f_e = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}} = \frac{69 \cdot 34}{99} = \frac{2346}{99}$$
$$= 23.70$$

for column 2, row 2

$$f_e = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}} = \frac{30 \cdot 34}{99} = \frac{1020}{99}$$
$$= 10.30$$

for column 1, row 3

$$f_e = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}} = \frac{69 \cdot 20}{99} = \frac{1380}{99}$$
$$= 13.94$$

for column 2, row 3

$$f_e = \frac{(\text{row total}) \cdot (\text{column total})}{\text{grand total}} = \frac{30 \cdot 20}{99} = \frac{600}{99}$$
$$= 6.06$$

Table 15.5 Frequency Observed and Frequency Expected for Three Types of Social Networks and for Smokers and Nonsmokers

	Type of social network: Poor	Moderate	Good
Smoker	$f_o = 25$ $f_e = 31.36$	$f_o = 26$ $f_e = 23.70$	$f_o = 18$ $f_e = 13.94$
Nonsmoker	$f_o = 20$ $f_e = 13.64$	$f_o = 8$ $f_e = 10.30$	$f_o = 2$ $f_e = 6.06$

Computation of χ^2

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(25 - 31.36)^2}{31.36} + \frac{(26 - 23.70)^2}{23.70} + \frac{(18 - 13.94)^2}{13.94}$$

$$+ \frac{(20 - 13.64)^2}{13.64} + \frac{(8 - 10.30)^2}{10.30} + \frac{(2 - 6.06)^2}{6.06}$$

$$= \frac{-6.36^2}{31.36} + \frac{2.3^2}{23.70} + \frac{4.06^2}{13.94} + \frac{6.36^2}{13.64} + \frac{-2.3^2}{10.30} + \frac{-4.06^2}{6.06}$$

$$= \frac{40.45}{31.36} + \frac{5.29}{23.70} + \frac{16.48}{13.94} + \frac{40.45}{13.64} + \frac{5.29}{10.30} + \frac{16.48}{6.06}$$

$$= 1.29 + 0.22 + 1.18 + 2.97 + 0.51 + 2.72$$

$$= 8.89$$

Finding Degree of Freedom (*df*)

- ✓ For two-way case, the degree of freedom is given by

  df = (number of rows -1) x (number or columns -1)

In this case (refer to table),

$df = (2 - 1) \times (3 - 1) = 2$

Finding The Critical Value of χ^2

Table 15.3 A Portion of Table χ, Critical Values of the χ^2 Statistic

df	.25	.10	.05	.025	.01	.005
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.071	12.833	15.086	16.750

Source: Owen, 1962.

From the Table χ with df = 2, α = 0.05, we would get the critical value of χ^2 equal to 5.991

Drawing a Conclusion

Comparing the computed value of χ^2 = 8.89 and the critical value of χ^2 = 5.991 , we can safely reject H_o (i.e. there is at least one combination of categories with different death rates)

Some Notes Pertaining to Chi-Square

Chi-square must be used with the following assumptions and limitations:

- ☑ It is assumed that the sample is randomly selected from the population.

- ☑ It is assumed that all observations are independent. (This assumption is usually met if only one observation is made for each subject)

- ☑ The chi-square test is limited to nominal data.

- ☑ The chi-square test tends to be less accurate with very small expected frequencies. (This is especially true with expected frequencies of less than 5. A good rule of thumb is not to conduct the chi-square test on data with expected frequencies of less than 5)

- ☑ The chi-square test tends to be less accurate for small degrees of freedom and a small N.

Performing Chi-Square Using SPSS

(go to Data (Chi-Square)

Data source: Data (Chi-Square)

The Crosstabulation: Category (CPAs) vs MUET Band

Category (CPAs)	MUET						Total
	1	2	3	4	5	6	
CPA <1	0	0	2	0	1	0	3
1.00 <= CPA < 1.69	1	4	8	3	1	0	17
1.70 <= CPA < 1.99	8	19	48	21	2	0	98
2.00 <= CPA < 2.19	12	62	133	49	5	0	261
2.20 <= CPA < 2.39	34	133	263	115	12	2	559
2.40 <= CPA < 2.59	44	258	480	187	19	0	988
2.60 <= CPA < 2.79	63	340	656	295	33	3	1390
2.80 <= CPA < 2.99	69	371	820	363	47	1	1671
3.00 <= CPA < 3.19	69	382	843	422	76	2	1794
3.20 <= CPA < 3.39	55	235	668	400	60	1	1419
3.40 <= CPA < 3.59	36	96	437	345	89	2	1005
3.60 <= CPA < 3.79	13	38	186	228	73	3	541
3.80 <= CPA < 3.89	2	5	33	58	22	2	122
CPA >= 3.90	0	0	7	27	21	1	56
Total	406	1943	4584	2513	461	17	9924

Setting the Hypothesis

H_o : There is no difference of academic achievement as measured by CPA categories over the different bands of MUET

H_1 : There is at least one combination (involving the different bands of MUET) with academic achievement (as measured by CPA categories)

The Chi-Square Test

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	834.8094	65	0.0000
Likelihood Ratio	727.5772	65	0.0000
Linear-by-Linear Association	439.7924	1	0.0000
N of Valid Cases	9924		

Note: 28 cells (33.3%) have expected count less than 5. The minimum expected count is .01

Conclusion: Reject H_0 as $p < \alpha$ (with $\alpha = 0.05$)

The Chi-Square Distribution