# MA331 Intermediate Statistics

## Lecture 07 Inference for Two-Way Tables [1]

Xiaohu Li

Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, New Jersey 07030

Week 09

[1] Based on Chapter 9.

# 0. Topics to be covered

This lecture focuses on the inference on the statistical association between two categorical random variables, we will cover the following topics:

- Statistical association
- Contingency table of two categorical r.v.'s
- Inference for two-way table
- Test for goodness-of-fit

# 1. Statistical association

☞ Chapters 7 and 8 deal with the inference about proportions in one-sample and two-sample settings. Here we study two populations with each response variable has two or more categories and test whether two concerned categorical variables are independent.

☞ The methods in this lecture answer questions such as the following.

- Are men and women equally likely to suffer lingering fear symptoms after watching scary movies at a young age?
- Does the style (classic, country and rock etc) of a stores background music affect the purchase of French and Italian wine?
- Is vitamin A supplementation of young children in developing countries associated with areduction in death rates?

# 2. A simplified two-way contingency table

✍ To study the inference for two proportions, we summarize the raw data by listing the number of obs ($n$) and sample count ($N$) in each pop.

✍ Example: To compare the proportions of male and female college students who engage in frequent binge drinking, the data is summarized as a table.

| Population | $n$ | $X$ | $\hat{p} = X/n$ |
|---|---|---|---|
| 1 (men) | 5,348 | 1,392 | 0.260 |
| 2 (women) | 8,471 | 1,748 | 0.206 |
| Total | 13,819 | 3,140 | 0.227 |

✍ CI suggests that the men are $5.4\%$ more likely to be frequent binge drinkers, with a $95\%$ margin of error of $1.5\%$.

✍ Motivation: we wonder whether the proportions of man and woman are different. That is, it makes sense to classify student by gender in this study.

# 3. Two-way contingency tables

✍ To be clear we consider a different summary of the data. Rather than recording just the count of binge drinkers, we record counts of all the outcomes in a two-way contingency table.

✍ Students classified by gender and whether or not they are frequent binge drinkers. Two categorical variables are 'Frequent binge drinker', with values 'Yes' and 'No', and 'Gender', with values 'Man' and 'Woman'.

Two-way table for frequent binge drinking and gender

| | Gender | | |
|---|---|---|---|
| Frequent binge drinker | Men | Women | Total |
| Yes | 1,392 | 1,748 | 3,140 |
| No | 3,956 | 6,723 | 10,679 |
| Total | 5,348 | 8,471 | 13,819 |

✍ In general, for the so-called $r \times c$ table with $r$ categories in row and $c$ categories in column, our research interest lies in test the association between the row and column category variables.

# 4. Test for the association between row and column

✍ At a given significance level $\alpha$, we want to test

$H_0$: no association b/w row and column

based on a given two-way table. Eg., Frequent binge drinker and Gender have no association.

✍ The alternative hypothesis $H_a$ is that there is an association between Row and Column variables. Eg., Binge and Gender has an association. Because $H_a$ includes all sorts of possible (positive and negative) associations, it can't be described as either one-sided or two-sided.

✍ General idea: To test $H_0$, we compare the observed cell counts with expected cell counts calculated under the assumption that $H_0$ is true. Intuitively, reject $H_0$ if they are very different.

# 5. Test for the association between row and column

✍Expected cell counts: under $H_0$: no association b/w Row and Column, each column follows a binomial distribution with a common prob, cell counts in this column are expected to be the corresponding means.

- For men, number of Binge drinkers
  $N_{Yes} \sim \mathrm{B}(3140/13819, 5348)$,
  $\mathrm{E}(N_{Yes}) = \frac{3140}{13819} \cdot 5348 = 1223.37$,
  $\mathrm{E}(N_{No}) = 5348 - 1223.37 = 4124.63$.
- For women, number of Binge drinkers
  $N_{Yes} \sim \mathrm{B}(3140/13819, 8471)$,
  $\mathrm{E}(N_{Yes}) = \frac{3140}{13819} \cdot 8471 = 1924.81$,
  $\mathrm{E}(N_{No}) = 8471 - 1924.81 = 6546.19$.
- Under $H_0$, we expect to observe the table

Two-way table for frequent binge drinking and gender

| Frequent binge drinker | Gender: Men | Gender: Women | Total |
|---|---|---|---|
| Yes | 1,392 | 1,748 | 3,140 |
| No | 3,956 | 6,723 | 10,679 |
| Total | 5,348 | 8,471 | 13,819 |

| Frequent binge drinker | Men | Women |
|---|---|---|
| Yes | 1223.37 | 1924.81 |
| No | 4124.63 | 6546.19 |

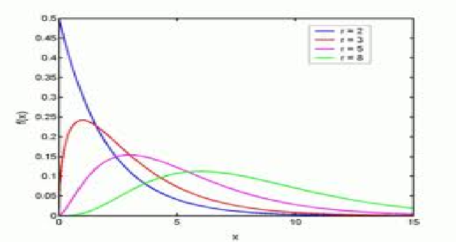
# 6. Test statistic for $H_0$: no association b/w row & column

✍ Difference between the observed table with cells counts $O_{i,j}$ and the expected to be observed table with cells counts $E_{i,j}$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

✍ Pearson proved that $\chi^2 \sim \chi^2_{(r-1)(c-1)}$ with degree of freedom $k = (r-1)(c-1)$, (Why?) and the density curve

$$f(x;k) = \frac{2^{1-\frac{k}{2}} x^{k-1} e^{-\frac{x^2}{2}}}{\Gamma(\frac{k}{2})}, \quad x \geq 0.$$




✍ The testing rule: reject $H_0$ if $p$-value of the observed statistic

$$\mathrm{P}\Big(\chi^2_k > \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}\Big) < \alpha.$$

# 7. Test for association b/w row and column – an example

✍ Test $H_0$: no association between Gender and Frequent binge drinker at significance level $\alpha = 0.01$.

✍ Testing statistic $\chi^2$ is observed as

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(o_{i,j}-e_{i,j})^2}{e_{i,j}} = \frac{(1392-1223.37)^2}{1223.37}+\frac{(1748-1924.81)^2}{1924.81} + \frac{(3956-4124.63)^2}{4124.62}+\frac{(6723-6546.19)^2}{6546.19} = 51.16.$$

✍ $p$-value

$$\mathrm{P}(\chi_1^2 > 51.16) = 1 - \mathrm{pchisq}(51.16, 1) = 8.53 \times 10^{-13} << 0.01 = \alpha.$$

So, reject $H_0$, i.e., Gender and Frequent binge drinker are associated.

✍ Critical value (upper $\alpha$ quantile)

$$\mathrm{qchisq}(1-0.01, 1) = 6.634897.$$

$\chi^2$ is observed as $51.16 >> 6.634897$. So, reject $H_0$ again.

# 8. Test for association of two-way tables: R example

```
                    Frequency of breast self-examination
                    ------------------------------------
Age                 Monthly      Occasionally       Never
---------------------------------------------------------
under 45              91             90              51
45 - 59              150            200             155
60 and over          109            198             172
---------------------------------------------------------
```

```
# Get data into an R data object.
 row1 = c(91,90,51)           # or col1 = c(91,150,109)
 row2 = c(150,200,155)        # and col2 = c(90,200,198)
 row3 = c(109,198,172)        # and col3 = c(51,155,172)
 2way = rbind(row1,row2,row3) # and 2way = cbind(col1,col2,col3)
 2way

     [,1] [,2] [,3]
row1   91   90   51
row2  150  200  155
row3  109  198  172

chisq.test(2way)    ## Do the Chi-square test.

Pearson Chi-squared test
data:  2way
X-squared=25.086, df=4, p-value=4.835e-05
```

# 9. Statistical models and the goodness-of-fit

✍ Statisticians crunch data to extract population's information (parameter estimation and hypothesis test) or discover population's pattern (data mining and statistical learning). This is achieved through building the statistical model for the data set.

✍ A statistical model is the probability distribution assumed for the data. E.g., Binomial distribution for flipping a coin $n$ times, and exponential distribution for system's lifetime etc.

✍ A new model usually must be confirmed first by statistical analysis and then verified by professional researchers in the area. For example,

- Engineering reliability: the lifetime of a component or system is of Weiull distribution.
- Network security: the node's degree in a network is of power law distribution.
- Operations research: the number of bugs detected is of poisson distribution.

✍ A new model or a model with a new background must be confirmed through testing the goodness-of-fit.

# 10. Measure the goodness-of-fit of a categorical variable

✍ At significance level $\alpha$, test

$H_0$: the population $X$ has the distribution $\mathrm{P}(X = x_i) = p_i,\ i = 1, \cdots, k$.

✍ Model (probability mass function) and the sample

- Model: probability masses $p_i,\ i = 1, \cdots, k.$
- Model: expected counts $E_i$'s $np_i,\ i = 1, \cdots, k.$
- Data: observed counts $O_i$'s $N_i = n_i,\ i = 1, \cdots, k.$

✍ Difference/distance between the model ($E_i$'s) and the sample ($O_i$'s)

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}.$$

✍ Under $H_0$, $\chi^2 \sim \chi^2_{k-1}$ with degree of freedom $k - 1$.

✍ The testing rule: reject $H_0$ if $p$-value of the observed statistic

$$\mathrm{P}\Big(\chi^2_{k-1} > \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}\Big) < \alpha.$$

# 11. Test for the goodness-of-fit – an example

✍ Adequate Calcium Today (ACT) study examines relationships b/w bone growth and calcium intake. Based on a sample we get

| State | AZ | CA | HI | IN | NV | OH | Total |
|---|---|---|---|---|---|---|---|
| participants | 167 | 257 | 257 | 297 | 107 | 482 | 1567 |

✍ Let us see how well the sample reflects state population proportions, i.e., test

$H_0$: the sample proportions coincide with population proportions.

| State | AZ | CA | HI | IN | NV | OH | Total |
|---|---|---|---|---|---|---|---|
| population proportion | 0.105 | 0.172 | 0.164 | 0.188 | 0.070 | 0.301 | 1.000 |
| expected counts | 164.54 | 269.52 | 256.99 | 294.60 | 109.69 | 471.67 | 1567.01 |

✍ Testing statistic $\chi^2$ is observed as $\frac{(167-164.535)^2}{164.535} + \cdots + \frac{(482-471.57)^2}{471.67} = 0.0369.$

✍ $p$-value $\mathrm{P}(\chi^2_5 > 0.0369) = 1 - \mathrm{pchisq}(0.0369, 5) = 0.965 >> 0.05 = \alpha$.

✍ Conclusion: Not reject $H_0$. That is, no evidence against the consistency between the sample proportions and population proportions.

# 12. Concluding remarks

✍ Pearson's $\chi^2$ test only tells whether there is the statistical association between row and column variables.

✍ If $H_0$ is not rejected, we tentatively tend to no association between row and column variables. In this occasion (e.g., Frequent Binge Drinker), one variable (e.g., gender) may be ignored. This is very useful in

- selecting dependent variable in logistic regression, and
- reducing dimension in big data analysis.

✍ Pearson's $\chi^2$ test can not tell the direction of the association (positive or negative) when it is confirmed.

# 13. More concluding remarks

✍ In statistical sense,

$$P(C_1 \mid R_1) > (<) P(C_1)$$

$$\Longrightarrow \quad \text{positive (negative) association.}$$

To confirm the direction we have to test the above inequality instead.

✍ Pearson's $\chi^2$ test can not directly handle the association between two continuous random variables.

✍ If a less accurate result is desired, one can consider the discretization:

- cut the range of concerned continuous variables into some categories,
- and then apply the $\chi^2$ test.