STA4030: Categorical Data Analysis
Two-way Tables: Measures of Association

Instructor: Bojun Lu

School of Data Science
CUHK(SZ)

September 28, 2020

Agenda

4.1 Introduction

4.1.1 Example

Recall the Physician's Health Study data. Let us combine the two categories "Fatal Attack" and "Nonfatal Attack" into one category "Attack".

	Myocardial Infarction		
	Attack	No Attack	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

When we test this data we observe $X^2 = 24.4291$ and a P-value less than 0.0001.

4.1 Introduction

Now imagine we observed this data.

	Myocardial Infarction		
	Attack	No Attack	Total
Placebo	19	1085	1104
Aspirin	10	1093	1103

With this data, $X^2 = 2.822$ and the P-value is 0.0930. Our conclusion from the test is completely different. Yet all we have done is divided all the numbers by 10 and rounded.

What's going on?

4.1 Introduction

4.1.2 Large X^2 or G^2 $\neq$ Large effect size

The two 2×2 tables of Physician's Health Study data (one real, the other imagined) surely exhibited the same relationship between the X and Y variables. After all, their sample proportions are (more or less) exactly the same. The real data does not exhibit a *stronger* relationship than the imaginary data, but its X^2 statistic was much larger.

The difference between the X^2 is not caused by a sudden change in the strength of the relationship, but in the change of sample size.

4.1 Introduction

Say we apply the LR and Pearson's Chi-squared tests to some data and reject the null hypothesis of "independence". What can we say about the nature of the dependence between the categorical variables X and Y?

Nothing, really. G^2 and X^2 tell us nothing about the direction of the dependence, nor the magnitude of it. They only tell us if it is *statistically significant* or not. Even if G^2 and X^2 yield very small P-values, this does not mean that the dependence or effect size is very large.

Therefore we need statistics which do tell us about the size and direction of the independence. There are several, but we will consider three.

4.2 Difference of Proportions

4.2.1 Introduction

For the time being, we will restrict ourselves to 2×2 tables. We use generic terms *success* and *failure* for the response categories of a binary variable.

We would like to compare the probability of a successful response in row 1 with the probability of a successful response in row 2.

The population difference of proportions is given by
$\delta = \pi_{1(1)} - \pi_{1(2)} = \frac{\pi_{11}}{\pi_{1+}} - \frac{\pi_{21}}{\pi_{2+}}$.

4.2 Difference of Proportions

When the sampling is product binomial and the data exhibits homogeneity, $\delta = 0$ because rows have identical conditional distributions.

When the sampling is Poisson or Multinomial and the data exhibits independence, $\delta = 0$, again because the rows have identical conditional distributions.

Clearly $-1 \leq \delta \leq 1$.

4.2 Difference of Proportions

4.2.2 Inference

The MLE for δ is $\hat{\delta} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}$.

The large-sample $100(1-\alpha)\%$ Wald confidence interval for δ is

$$\hat{\delta} \pm z_{\alpha/2}\hat{\sigma}(\hat{\delta}),$$

where the estimated standard error of $\hat{\delta}$ (when treating two rows as independent binomial samples) is

$$\hat{\sigma}(\hat{\delta}) = \sqrt{\frac{\frac{n_{11}}{n_{1+}}(1-\frac{n_{11}}{n_{1+}})}{n_{1+}} + \frac{\frac{n_{21}}{n_{2+}}(1-\frac{n_{21}}{n_{2+}})}{n_{2+}}}$$

$$= \sqrt{\frac{n_{11}n_{12}}{n_{1+}^3} + \frac{n_{21}n_{22}}{n_{2+}^3}}$$

4.2 Difference of Proportions

4.2.3 Example - Physician's Health Study

Consider the (real) Physician's Health Study data again, with the two "Attack" categories combined.

	Myocardial Infarction		
Group	Attack	No Attack	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

We treat the two rows as independent binomial samples, then

$$\hat{\pi}_{1(1)} = \frac{189}{11034} = 0.0171, \quad \hat{\pi}_{1(2)} = \frac{104}{11037} = 0.0094.$$

The estimated difference is $\hat{\delta} = 0.0171 - 0.0094 = 0.0077$.

4.2 Difference of Proportions

This difference has an estimated standard error of

$$\sqrt{\frac{(0.0171)(0.9829)}{11,034} + \frac{(0.0094)(0.9906)}{11,037}} = 0.0015.$$

A 95% confidence interval for the true difference $\pi_1 - \pi_2$ is

$$0.0077 \pm 1.96(0.0015) \quad \text{or} \quad (0.005, 0.011).$$

Since this interval contains only positive values, we conclude that

$$\pi_1 - \pi_2 > 0 \quad \text{or} \quad \pi_1 > \pi_2.$$

Taking aspirin is associated with diminishing the risk of heart attacks.

4.3 Relative Risk

4.3.1 Introduction

Consider a comparison of two drugs on the proportion of subjects who have adverse reactions when using the drug.

A difference between 0.010 and 0.001 is 0.009. The difference between 0.410 and 0.401 is also 0.009. The first difference seems more noteworthy, since ten times as many subjects have adverse reactions with one drug as the other.

The ratio of proportions of success in each row is called the relative risk.

4.3 Relative Risk

The relative risk, RR is defined by $RR = \frac{\pi_{1(1)}}{\pi_{1(2)}} = \frac{\pi_{11}/\pi_{1+}}{\pi_{21}/\pi_{2+}}$.

The relative risk can be any nonnegative real number.

The relative risk can have different interpretations to the difference of proportions. For instance, the proportions 0.010 and 0.001 have a relative risk of $0.010/0.001 = 10.0$, whereas the proportions 0.410 and 0.401 have a relative risk of $0.410/0.401 = 1.02$.

A relative risk of 1 occurs when $\pi_{1(1)} = \pi_{2(1)}$, which means that response is independent of group.

RR is probably a better measure of association the δ when proportions are extreme. In medical research, RR is much more common than δ.

4.3 Relative Risk

4.3.2 Inference

The MLE for RR is $\widehat{RR} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$.

$\widehat{RR}$ is skewed (it has to be a non-negative number), so it can require a very large sample size before it behave normally. We therefore consider $\log(\widehat{RR})$, which will be less skewed and approach normality much faster.

The large-sample $100(1-\alpha)\%$ Wald confidence interval for $\log(RR)$ is

$$\log(\widehat{RR}) \pm z_{\alpha/2}\hat{\sigma}(\log(\widehat{RR})),$$

where the estimated standard error of $\log(\widehat{RR})$ is

$$\hat{\sigma}(\log(\widehat{RR})) = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}}.$$

4.3 Relative Risk

4.3.3 Example - Physician's Health Study (again)

For Physician's Health Study data, the sample relative risk is

$$\widehat{RR} = \frac{0.0171}{0.0094} = 1.82$$

The sample proportion of heart attack cases was 82% higher for the group taking placebo. The 95% CI for $\log(RR)$ is $(0.3598, 0.8355)$. The 95% CI for RR is therefore $(e^{0.3598}, e^{0.8355}) = (1.4330, 2.3060)$.

This indicates that the risk of heart attacks is at least 43% higher for the placebo group. Compare this interpretation of the data to that provided by the difference of proportions.

4.4 Odds Ratios

4.4.1 Introduction

The odds of an event is the ratio of the probability of the event occurring to the probability the event does not occur. This is the statistical definition of the word "odds". Its meaning in gambling is different.

Thus for a probability of success π, the odds of success are

$$\text{odds} = \Omega = \pi/(1-\pi).$$

If $\pi = 0.75$, then the odds of success equal $0.75/0.25 = 3$. We then expect to observe three successes for every one failure. When odds = 1/3, we expect to observe one success for every three failures.

The odds are nonnegative. When odds are greater than 1, a success is more likely than a failure.

4.4 Odds Ratios

Consider any 2×2 table:

$X \backslash Y$	S	F
1	π_{11}	π_{12}
2	π_{21}	π_{22}

We can calculate the odds of success in the two rows:

$\Omega_1 = \dfrac{\pi_{11}/\pi_{1+}}{\pi_{12}/\pi_{1+}}$: odds of S to F for Y given level 1 of X

$\Omega_2 = \dfrac{\pi_{21}/\pi_{2+}}{\pi_{22}/\pi_{2+}}$: odds of S to F for Y given level 2 of X

Then compare these odds by taking their ratio:

$$\theta = \Omega_1/\Omega_2 = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \text{odds ratio (or cross-product ratio)}$$

4.4 Odds Ratios

4.4.2 Properties

1. The value of θ reflects the direction and the degree of association.

   If $0 \leq \theta < 1$, individuals in row 2 are less likely to fall in column 2 than are individuals in row 1.

   If $1 < \theta < \infty$, individuals in row 2 are more likely to fall in column 2 than are individuals in row 1.

2. $\theta = 1 \Longleftrightarrow X$ and Y are independent.

4.4 Odds Ratios

Proof:

$\Leftarrow$ If X and Y are independent, then $\theta = 1$.

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\Pr(X=1,Y=1)\Pr(X=2,Y=2)}{\Pr(X=1,Y=2)\Pr(X=2,Y=1)}$$
$$= \frac{\Pr(X=1)\Pr(Y=1)\Pr(X=2)\Pr(Y=2)}{\Pr(X=1)\Pr(Y=2)\Pr(X=2)\Pr(Y=1)} = 1.$$

$\Rightarrow$ If $\theta = 1$, then X and Y are independent.

$$\theta = 1 \Rightarrow \Omega_1 = \Omega_2 \Rightarrow \frac{\pi_{12}}{\pi_{11}} = \frac{\pi_{22}}{\pi_{21}} \qquad (*)$$

$$\Pr(Y=1|X=1) = \frac{\Pr(Y=1,X=1)}{\Pr(X=1)} = \frac{\pi_{11}}{\pi_{1+}} = \frac{\pi_{11}}{\pi_{11}+\pi_{12}} = \frac{1}{1+\frac{\pi_{12}}{\pi_{11}}}$$

$$\Pr(Y=1|X=2) = \frac{\Pr(Y=1,X=2)}{\Pr(X=2)} = \frac{\pi_{21}}{\pi_{2+}} = \frac{\pi_{21}}{\pi_{21}+\pi_{22}} = \frac{1}{1+\frac{\pi_{22}}{\pi_{21}}}$$

4.4 Odds Ratios

Based on $(*)$, $\Pr(Y = 1 \mid X = 1) = \Pr(Y = 1 \mid X = 2) = p^*$ (say).

On the other hand,

$X \backslash Y$	1	2	
1	π_{11}	π_{12}	π_{1+}
2	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	1

$$\begin{aligned}\Pr(Y = 1) &= \pi_{+1} = \pi_{11} + \pi_{21} \\ &= \pi_{1+} \Pr(Y = 1 \mid X = 1) + \pi_{2+} \Pr(Y = 1 \mid X = 2) \\ &= \pi_{1+} p^* + \pi_{2+} p^* \\ &= (\pi_{1+} + \pi_{2+}) p^* = p^*.\end{aligned}$$

So, $\Pr(Y = 1 \mid X = 1) = \Pr(Y = 1 \mid X = 2) = \Pr(Y = 1)$.
Similarly, $\Pr(Y = 2 \mid X = 1) = \Pr(Y = 2 \mid X = 2) = \Pr(Y = 2)$.
i.e., X and Y are independent.

4.4 Odds Ratios

3. θ does not change when rows become the columns and the columns become the rows:

$X\backslash Y$	S	F
1	π_{11}	π_{12}
2	π_{21}	π_{22}

$$\theta_1 = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}},$$

$Y\backslash X$	1	2
S	π_{11}	π_{21}
F	π_{12}	π_{22}

$$\theta_2 = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}},$$

$\therefore \theta_1 = \theta_2.$

Therefore it is unnecessary to identify one variable as the response in order to use θ.

4.4 Odds Ratios

4. We will more often work on $\log\theta$, which is invariant under the interchange of rows or columns (except its sign).

$$\begin{vmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{vmatrix} \qquad \theta_1 = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

$$\begin{vmatrix} \pi_{12} & \pi_{11} \\ \pi_{22} & \pi_{21} \end{vmatrix} \qquad \theta_2 = \frac{\pi_{12}\pi_{21}}{\pi_{11}\pi_{22}} \quad \text{(interchange columns)}$$

$$\begin{vmatrix} \pi_{21} & \pi_{22} \\ \pi_{11} & \pi_{12} \end{vmatrix} \qquad \theta_3 = \frac{\pi_{12}\pi_{21}}{\pi_{11}\pi_{22}} \quad \text{(interchange rows)}$$

$$\theta_1 = \frac{1}{\theta_2} = \frac{1}{\theta_3}, \qquad \log\theta_1 = -\log\theta_2 = -\log\theta_3.$$

4.4 Odds Ratios

For example, from the Physician's Health Study

Group\Attack	Yes	No
Placebo	189	10,845
Aspirin	104	10,933

$$\theta_1 = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1.832, \log\hat{\theta}_1 = 0.605.$$

Group\Attack	No	Yes
Placebo	10,845	189
Aspirin	10,933	104

$$\theta_2 = \frac{\pi_{12}\pi_{21}}{\pi_{11}\pi_{22}} = 0.546, \log\hat{\theta}_2 = -0.605.$$

Group\Attack	Yes	No
Aspirin	104	10,933
Placebo	189	10,845

$$\theta_3 = \frac{\pi_{12}\pi_{21}}{\pi_{11}\pi_{22}} = 0.546, \log\hat{\theta}_3 = -0.605.$$

The interpretations of θ_1, θ_2, and θ_3 are the same.

4.4 Odds Ratios

5. The odds ratio and Relative Risk are related the following way:

$$\text{Odds ratio} = \text{Relative risk} \times \left(\frac{\pi_{22}/\pi_{2+}}{\pi_{12}/\pi_{1+}} \right).$$

When the proportion of success is close to zero for both groups, the odds ratio and relative risk take similar values. e.g. For the aspirin and heart attacks data, the relative risk is 1.82 and the odds ratio is 1.832.

Note: It is unnecessary to identify one classification as a response variable to estimate θ. By contrast, the relative risk requires this specification, and its value depends on whether it is applied to the first or to the second outcome category.

4.4 Odds Ratios

4.4.3 Inference

The MLE of θ is $\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$.

As with RR, θ is non-negative hence $\hat{\theta}$ has a skewed sampling distribution. We therefore use the log transform as this converges to normality faster.

In a large sample, it can also be shown that

$$\log\hat{\theta} \xrightarrow{L} N(\log\theta, \sigma^2(\log\hat{\theta})),$$

where

$$\sigma^2(\log\hat{\theta}) = \frac{1}{n\pi_{11}} + \frac{1}{n\pi_{12}} + \frac{1}{n\pi_{21}} + \frac{1}{n\pi_{22}}.$$

4.4 Odds Ratios

We can estimate $\sigma^2(\log\hat{\theta})$ by

$$\hat{\sigma}^2(\log\hat{\theta}) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}.$$

Therefore the large-sample $100(1-\alpha)\%$ Wald confidence interval for $\log\theta$ is

$$\log\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}(\log\hat{\theta})$$

Note $\hat{\theta}$ equals 0 or ∞ if any of the $n_{ij} = 0$. In this case, one may make the adjustment of adding 0.5 to each cell and recalculating the Wald C.I., for example

$$\hat{\theta} = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}.$$

4.4 Odds Ratios

4.4.4 Example - Physician's Health Study (again, again)

For the physicians taking placebo, the estimated odds of a heart attack equal $n_{11}/n_{12} = 189/10,845 = 0.0174 = 1.74/100$. There were 1.74 "yes" outcomes for every 100 "no" outcomes.

For the physicians taking aspirin, the estimated odds of a heart attack equal $n_{21}/n_{22} = 104/10,933 = 0.0095 = 0.95/100$. There were 0.95 "yes" outcomes for every 100 "no" outcomes.

4.4 Odds Ratios

$\hat{\theta} = 0.0174/0.0095 = 1.832$. The estimated odds of a heart attack were 83% higher for the placebo group.

The 95% CI for $\log\theta$ is $(0.3647, 0.8462)$. The 95% CI for θ is therefore $(e^{0.3647}, e^{0.8462}) = (1.4401, 2.3308)$.

This indicates that the odds of heart attacks is at least 44% higher for the placebo group. Note the similarity to the RR calculations, because the probability of "success" is small for both groups.

4.5 For $r \times c$ Tables

4.5.1 Introduction

All the measures of association we have so far discussed have been defined for 2×2 tables. That's because a single number, like an odds ratio, could do the job.

However, with a general $r \times c$ table, one number cannot be enough because there are more than one way for the data to exhibit dependence. After all, the LR and Pearson's Chi-squared tests for independence/homogeneity use $(r-1)(c-1)$ degrees of freedom. It follows we'll need $(r-1)(c-1)$ numbers to measure the association in the table.

In fact, you could calculate more than $(r-1)(c-1)$ numbers to describe the association, but you only need $(r-1)(c-1)$.

4.5 For $r \times c$ Tables

4.5.2 Odds Ratios in $r \times c$ Tables

Consider the set of $(r-1)(c-1)$ *local odds ratios*

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}, \; i = 1, \ldots, r-1, \; j = 1, \ldots, c-1.$$

These $(r-1)(c-1)$ odds ratios determine all odds ratios formed from pairs of rows and pairs of columns.

Another basic set is

$$\alpha_{ij} = \frac{\pi_{ij}\pi_{rc}}{\pi_{rj}\pi_{ic}}, \; i = 1, \ldots, r-1, \; j = 1, \ldots, c-1.$$

The $(r-1)(c-1)$ parameters in above two formulae can describe any association in an $r \times c$ table. Independence is equivalent to all $(r-1)(c-1)$ odds ratios equaling 1.0.

4.5 For $r \times c$ Tables

Recall the original data from the Physician's Health Study.

	Myocardial Infarction			
	Fatal Attack	Nonfatal Attack	No Attack	Total
Placebo	18	171	10845	11034
Aspirin	5	99	10933	11037

Here, the sample local odds ratios are $\frac{18 \times 99}{171 \times 5} = 2.08$ from the first two columns and $\frac{171 \times 10933}{10845 \times 99} = 1.74$ from the second two columns. The product of these is 3.63, the odds ratio from the first and third columns.

These odds ratios both suggest that more serious events are more common in the placebo group. *Note: we could have also calculated two differences in proportions or two relative risks to investigate the data.*

4.5 For $r \times c$ Tables

4.5.3 Row Fractions

Sometimes, a simple comparison of the conditional distributions can be revealing.

For each row, divide the observed frequency in every cell by the sum of the frequencies of the row. These are called *row fractions*.

Compare the row fractions among the rows and if the row fractions are identical among all rows, the variables are not associated.

4.5 For $r \times c$ Tables

For example, say we observe

$X_1 \backslash X_2$	A	B	C	Total
A	1	9	10	20
B	9	81	90	180
C	10	90	100	200
Total	20	180	200	400

and we calculate their row fractions

$X_1 \backslash X_2$	A	B	C	Total
A	1/20	9/20	10/20	20
B	9/180	81/180	90/180	180
C	10/200	90/200	100/200	200
Total	20	180	200	400

and since they are the same in each row, we can say X_1 and X_2 are are not associated.

4.5 For $r \times c$ Tables

For another example, consider the table below

		Grade					
		A	B	C	D	E	Total
Preliminary	pass	14	22	3	0	0	39
Course	fail	1	3	4	2	1	11
	Total	15	25	7	2	1	50

It's clear from these row fractions there might be an association between the result in the preliminary course and the final grade.

In this case, the variables are ordinal. We shall see that for ordinal data in $r \times c$ tables, more measures of association and tests are available.