Class Notes: STAT 501
Nonparametrics & Log-Linear Models
Relative Risk, Odds Ratio,
Fisher's Exact Test of 2x2 Tables

Da Kuang
University of Pennsylvania

Contents

1 Relative Risk

A difference between two proportions of a certain fixed size usually is more important when both proportions are near 0 or 1 than when they are near the middle of the range.

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

- The difference between 0.010 and 0.001 is the same as the difference between 0.410 and 0.401, namely 0.009.
- The first difference is more striking, since 10 times as many subjects had adverse reactions with one drug as the other.

In such cases, the ratio of proportions, namely relative risk, is a more relevant descriptive measure. For the above example,

- The proportions 0.010 and 0.001 have a relative risk of 0.010/0.001 = 10.
- The proportions 0.410 and 0.401 have a relative risk of 0.010/0.001 = 1.02.

A relative risk of 1.00 occurs when $p_1 = p_2$, that is when the response is independent of the group.

2 Review: Study Design

2.1 Conditional Design

When the rows of a contingency table refer to different groups, the sample sizes for those groups are often fixed by the sampling design. We assume a binomial distribution for the sample in each row, with number of trials equal to the fixed row total.

When the columns are a response variable Y and the rows are an explanatory variable X, it is sensible to divide the cell counts by the row totals to form conditional distributions on the response. In doing so, we inherently treat the row totals as fixed and analyze the data the same way as if the two rows formed separate samples.

2.2 Unconditional Design

When the total sample size n is fixed and we cross classify the sample on two categorical response variables, the multinomial distribution is the actual joint distribution over the cells. The cells of the contingency table are the possible outcomes, and the cell probabilities are the multinomial parameters.

3 Odds Ratio

In lecture 9, we assume the rows of contingency table are two independent binomial distributions, and we would like to check whether the two distribution have the same probability parameter.

In this lecture, we would like to exam the dependency. For conditional design, we check the dependency between the two binomial distribution for each row. For the unconditional design, we check the dependency among the four cells in the table.One common used measure for association/dependency is the sample odds ratio.

3.1 Odds

For a probability of success π_1, the odds of success are defined to be $\pi_1/(1-\pi_1)$. For instance, if $\pi_1 = 0.75$, the odds of success equal 3.

The odds are non-negative, with value greater than 1 when a success is more likely than a failure.

- When odd = 3, a success is 3 times as likely as a failure. We expect to observe 3 successes for every 1 failure.
- When odd = $1/3$, a failure is 3 times as likely as a success. We expect to observe 1 success for every 3 failure.

3.2 Odds Ratio

Suppose

- Random variable U has probability of success π_1.
- Random variable V has probability of success π_2.

Definition 3.1 (Odds Ratio) *Odds ratio is the ratio of two odds*

$$\theta = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} \geq 0$$

If $\pi_1 = \pi_2$, then $\theta = 1$. The independence of U and $V \iff \theta = 1$.

3.3 Conditional Design

Suppose the treatment group subjects iid Bernoulli(π_1), $i = 1, \cdots, n_1$, and the control group subjects iid Bernoulli(π_2), $j = 1, \cdots, n_2$.

	Success	Failure
Treatment	π_1	
Control	π_2	

	Success	Failure	Total
Treatment	n_{11}	n_{12}	n_1
Control	n_{21}	n_{22}	n_2

Figure 1: Conditional Contingency Table

The sample odd ratio is

$$\widehat{\theta} = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\frac{n_{11}}{n_{12}}}{\frac{n_{21}}{n_{22}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

3.4 Unconditional Design

Let π_{ij} denote the true unknown joint probability of falling into the ij-th cell, $i = 1, 2$, $j = 1, 2$. The contingency table is as follows.

	Factor Y		
Factor X	Yes	No	Total
Yes	π_{11}	π_{12}	$\pi_1 = \pi_{11} + \pi_{12}$
No	π_{21}	π_{22}	$\pi_2 = \pi_{12} + \pi_{22}$
Total	$\pi_{.1} = \pi_{11} + \pi_{21}$	$\pi_{.2} = \pi_{12} + \pi_{22}$	1

Figure 2: Unconditional Contingency Table

In the table,

- $\pi_{11} = \mathbf{P}(X = 1 \text{ and } Y = 1)$
- $\pi_{12} = \mathbf{P}(X = 1 \text{ and } Y = 1)$
- $\pi_{21} = \mathbf{P}(X = 0 \text{ and } Y = 1)$

- $\pi_{22} = \mathbf{P}(X = 0 \text{ and } Y = 0)$

The odd θ_1 in the first row is

$$\theta_1 = \frac{\mathbf{P}(Y = 1|X = 1)}{P(Y = 0|X = 1)} = \frac{\frac{\mathbf{P}(Y=1,X=1)}{\mathbf{P}(X=1)}}{\frac{\mathbf{P}(Y=0,X=1)}{\mathbf{P}(X=1)}} = \frac{\frac{\pi_{11}}{\pi_{11}+\pi_{12}}}{\frac{\pi_{12}}{\pi_{11}+\pi_{12}}} = \frac{\pi_{11}}{\pi_{12}}$$

The odd θ_2 in the second row is

$$\theta_2 = \frac{\mathbf{P}(Y = 1|X = 0)}{P(Y = 0|X = 0)} = \frac{\frac{\mathbf{P}(Y=1,X=0)}{\mathbf{P}(X=0)}}{\frac{\mathbf{P}(Y=0,X=0)}{\mathbf{P}(X=0)}} = \frac{\frac{\pi_{21}}{\pi_{21}+\pi_{22}}}{\frac{\pi_{22}}{\pi_{21}+\pi_{22}}} = \frac{\pi_{21}}{\pi_{22}}$$

Therefore, the population odd ratio is

$$\theta = \frac{\theta_1}{\theta_2} = \frac{\frac{\pi_{11}}{\pi_{12}}}{\frac{\pi_{21}}{\pi_{22}}} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{22}}$$

Hence, the sample odd ratio is

$$\widehat{\theta} = \frac{n_{11}n_{22}}{n_{21}n_{22}}$$

One may notice that the sample odd ratios are the same in two different studies.

4 Properties of Odds Ratio

4.1 Directed Association

The odds ratio θ measures the strength of association between X and Y.

- $1 < \theta < \infty$ implies that the odds of $Y = 1$, given that $X = 1$, is larger than the odds that $Y = 1$, given that $X = 0$.
- $0 < \theta < 1$ implies that the odds of $Y = 1$, given that $X = 1$, is smaller than the odds that $Y = 1$, given that $X = 0$.

Values of θ farther from 1 in a given direction represent stronger association.

Note that the association is directed.

- The odd ratio does not change value when the table orientation reverses so that the rows become the columns and the columns become the rows.
- Two values for θ represent the same strength of association, but in opposite directions, when one value is the inverse of the other.

4.2 Example: $\theta = 4$ vs. $\theta = 1/4$

- When the order of the rows is reversed or the order of the columns is reversed, the new value of θ is the inverse of the original value.
- This ordering is usually arbitrary, so whether we get 4 or 0.25 for the odds ratio is merely a matter of how we label the rows and columns.

4.3 Equivalent to independence

The odd ratio $\theta = 1 \iff X$ and Y are independent.

Proof.

- If X and Y are independent, then

$$\theta_1 = \frac{\mathbf{P}(Y = 1|X = 1)}{P(Y = 0|X = 1)} = \frac{\mathbf{P}(Y = 1)}{\mathbf{P}(Y = 0)}$$
$$\theta_2 = \frac{\mathbf{P}(Y = 1|X = 0)}{P(Y = 0|X = 0)} = \frac{\mathbf{P}(Y = 1)}{\mathbf{P}(Y = 0)}$$
$$\theta = \frac{\theta_1}{\theta_2} = 1$$

- If $\theta = 1$, then

$$\frac{\mathbf{P}(Y=1,X=1)}{\mathbf{P}(Y=0,X=1)} = \frac{\mathbf{P}(Y=1,X=0)}{\mathbf{P}(Y=0,X=0)}$$
$$\frac{\mathbf{P}(Y=1,X=1)}{\mathbf{P}(X=1)-\mathbf{P}(Y=1,X=1)} = \frac{\mathbf{P}(Y=1,X=0)}{\mathbf{P}(X=0)-\mathbf{P}(Y=1,X=0)}$$
$$\frac{\mathbf{P}(Y=1,X=1)}{\mathbf{P}(X=1)-\mathbf{P}(Y=1,X=1)} = \frac{\mathbf{P}(Y=1,X=0)}{1-\mathbf{P}(X=1)-\mathbf{P}(Y=1,X=0)}$$
$$\mathbf{P}(X=1,Y=1) = \mathbf{P}(X=1)\mathbf{P}(Y=1)$$

  Hence, we have X and Y are independent.

■

Therefore the null hypothesis about independence can be reduced to the odd ratio $\theta = 1$.

4.4 Example: Aspirin vs Cardiovascular Disease

The Physicians' Health Study was a five-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease. Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo. The study was "blind" — the physicians in the study did not know which type of pill they were taking.

	Myocardial Infarction	
Group	Yes	No
Placebo	189	10845
Aspirin	104	10933

Figure 3: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. New Engl. J. Med., 318: 262-264, 1988.

$$\theta = \frac{189 \times 10933}{104 \times 10845} = 1.83$$

The estimated odds of Myocardial Infarction for male physicians taking placebo equals 1.83 times the estimated odds for male physicians taking aspirin. The estimated odds were 83% higher for the placebo group.

4.5 Large Sample Inference

If the sample size is large enough, we can use central limit theorem (CLT) for hypothesis testing and estimate the confidence interval.

Note that the range of odd ratio for association is not symmetric, $0 < \theta < 1$ and $1 < \theta < \infty$. So before applying CLT, we need to transform the odd ratio into log-space, i.e. $\lambda = \ln \theta$.

Then the sample estimation will be $\widehat{\lambda} = \ln \widehat{\theta}$. The estimated standard deviation is $\widehat{\text{SD}}(\widehat{\lambda}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$.

The null hypothesis of indepence, $\theta = 1$ can be reduced to $\lambda = 0$.

We can apply CLT to compare $\widehat{\lambda}/\widehat{\text{SD}}(\widehat{\lambda})$ with normal quantile for hypothesis.

Two-sided confidence interval:

$$\widehat{\lambda} - z_{\alpha/2}\widehat{\text{SD}}(\widehat{\lambda}) \leq \lambda \leq \widehat{\lambda} + z_{\alpha/2}\widehat{\text{SD}}(\widehat{\lambda})$$

Transform to the odd ratio:

$$\exp\left(\widehat{\lambda} - z_{\alpha/2}\widehat{\text{SD}}(\widehat{\lambda})\right) \leq \theta \leq \exp\left(\widehat{\lambda} + z_{\alpha/2}\widehat{\text{SD}}(\widehat{\lambda})\right)$$

5 Fisher Exact Test

Sir Ronald A. Fisher, the English statistician, has been called the father of modern statistics.

5.1 The Lady Tasting Tea

A famous story, in Cambridge, England, in the late 1920is, that a colleague of Fisher's claimed that she could tell, while drinking tea with milk, whether milk or tea was poured into the cup first.

An experiment was designed to test her claim.

- 8 cups of tea were presented to her in a random order; 4 of these had milk poured first while the other 4 had tea poured first.
- She was told that there were 4 cups of each type.

Suppose we have the following data show the results of the experiment.

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Figure 4: Contingenvy Table of Lady Tasting Tea

One can observe that she was right 3 out of 4 times on both types. Is this sufficient evidence to support her claim of special power?

The contingency table has two difference comparing with the one we met.

- The columns are fixed.
- The sample size is very small.

5.2 Review: Hyper-geometric Distribution

Suppose there are m red balls and $n - m$ green balls. We randomly pick k balls without replacement. What is the probability that z balls are red?

Let Z be the number of red balls.

$$\mathbf{P}(Z = z) = \frac{\binom{m}{z}\binom{n-m}{k-z}}{\binom{n}{k}}, z = 0, 1, \cdots, k.$$

Here we assume that

- $z \leq k$
- $z \leq m$
- $k - z \leq n - m$

5.2.1 Example

Suppose we have the following contingency table, in other words, $m = 6, n-m = 19$, $n = 25$. Also, $k = 10$ and $z = 1$.

	picked	not picked	Total
Red	1		6
Green	9		19
Total	10		25

Figure 5: Hyper-geometric Distribution Example

We have

$$\mathbf{P}(Z=1)=\frac{\binom{6}{1}\binom{19}{9}}{\binom{25}{10}}$$

The pdf of hyper-geometric distribution in R is as follows.

```
choose(6,1)*choose(19,9)/choose(25,10)
dhyper(x=1, m=6, 19, k=10)
```

5.3 Fisher Exact Test

Applying the hyper-geometric distribution, conditional on 4 guesses of having milk added first, the probability of 3 correct guesses is

$$\mathbf{P}(Z=3)=\frac{\binom{4}{3}\binom{4}{4-3}}{\binom{8}{4}}=0.229$$

So if the lady cannot tell the difference between two types of mix, there is still 22.9% of probability for her to have 3 correct guesses by chance. Then how to exam her claim? Hypothesis Testing. To be more specific, Fisher Exact Test.

5.4 Exact Inference for Small Samples

The setup of Fisher Exact Test can be represented by the following contingency table.

	Factor Y		
Factor X	Level 1	Level 2	Total
Level 1	n_{11}		n_1
Level 2			$n_2 = n - n_1$
Total	$n_{.1}$		n

Figure 6: Contingency Table of Fisher Exact Test

In the table, n_1, n, and $n_{\cdot 1}$ are fixed. We can infer that $n_{\cdot 2} = n - n_{\cdot 1}$. n_{11} is the observed random variable.

Fisher's exact test is based on the conditional distribution of n_{11} given n_1, n, and $n_{\cdot 1}$. Given, $z \leq \min\{n_1, n_{\cdot 1}\}$, and $n_{\cdot 1} - z \leq n_2$,

$$\mathbf{P}(n_{11} = z) = \frac{\binom{n_1}{z}\binom{n_2}{n_{\cdot 1} - z}}{\binom{n}{n_{\cdot 1}}}.$$

Recall that the p-value is calculated by the sum of all event which are equal or rarer than the observed event.

Note that the estimated odd ratio is different from the calculation we have been using. It is because that the experiment design is different with the ones we have. The estimation of odd ratio can be found in R's fisher test manual.

```
TeaTasting=
  matrix(c(3, 1, 1, 3),
         nrow = 2,
         dimnames = list(Truth = c("Milk", "Tea"),
                         Guess = c("Milk", "Tea")))

TeaTasting
# Guess
# Truth Milk Tea
# Milk 3 1
# Tea 1 3
fisher.test(TeaTasting,alternative="g")
# Fisher's Exact Test for Count Data
#
# data: TeaTasting
# p-value = 0.2429
# alternative hypothesis: true odds ratio is greater than 1
# 95 percent confidence interval:
```

```
# 0.3135693 Inf
# sample estimates:
# odds ratio
# 6.408309
dhyper(x=3,m=4,4,k=4)
# [1] 0.2285714
dhyper(x=4,m=4,4,k=4)
# [1] 0.01428571
```