MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, we finished Module 10 the contingency tables.

- Today we cover the next Module 11 Multiple contingency tables (Chapter 16 of textbook)

Odds and Odds Ratio

- We first introduce two terms to formula the question.
- For a Bernoulli trial (binary outcome: Yes/No), the ratio of the two probabilities (Yes versus No) is called the odds.

Odds= $\frac{P}{1-\mathrm{P}}$ where P is the probability of getting 'Yes'.

- We now see what are the odds for two populations in a 2x2 contingency table where each column reflects one population.

Odds and Odds Ratio

- In a 2x2 table where each column reflects one population:
- The conditional probability of Yes in

p_{11}	p_{12}
p_{21}	p_{22}

first population is $P_1 = \frac{P_{11}}{P_{11}+P_{21}}$; hence

the odds is $Odds_1 = \frac{P_1}{1-P_1} = \frac{P_{11}}{P_{21}}$.

The conditional probability of Yes in second population is $P_2 = \frac{P_{12}}{P_{12}+P_{22}}$;

hence the odds is $Odds_2 = \frac{P_2}{1-P_2} = \frac{P_{12}}{P_{22}}$.

- Thus the odds-ratio is $OR = \frac{Odds_1}{Odds_2} = \frac{P_{11}P_{22}}{P_{21}P_{12}}$.

Inference of Odds Ratio

- Notice that marginal independence means $\mathbf{H_0}$: $OR = 1$.
- Testing this null hypothesis is in fact the independent two-samples proportions comparison testing in the last chapter, which can be done with the χ^2-test on this contingency table.

- Next (chapter 16), we study the inference of the odds ration (OR) in multiple contingency tables.

Chapter 16 Multiple 2x2 Contingency Tables

- First, why do we need statistical methods for inference of odds-ratio (OR) in multiple contingency tables?
- Can we simply merge those tables into one contingency tables?

- Simpson's Paradox: We can get misleading results if we simply merge the multiple 2x2 contingency tables.

Simpson's Paradox

- Salary data in two companies.

Company 1	Male employees	Female employees
Pay ≤$100,000	10	1
Pay >$100,000	100	10

$$OR=\frac{(10)(10)}{(100)(1)}=1.$$

Company 2	Male employees	Female employees
Pay ≤$100,000	10	100
Pay >$100,000	1	10

$$OR=\frac{(10)(10)}{(1)(100)}=1 \text{ also.}$$

Merged Together	Male employees	Female employees
Pay ≤$100,000	20	101
Pay >$100,000	101	20

$$OR=\frac{(20)(20)}{(101)(101)}\approx 0.04 \text{ when}$$

two tables are combined.

Simpson's Paradox

- Salary Example: There appears to be differences in employee's pay for different genders (OR≈0.04), but no such differences in each company (OR=1).

(No bias in paying in respect to gender, but maybe the bias is in recruiting?)

- Simpson's Paradox: the odds-ratio in combined contingency table can be misleading!

Inference of Odds-Ratio in Multiple 2x2 Tables

- The problem is separated in two steps:

(A) **Test** if the *odds-ratios* across all tables are the *same*. If they are different, then the inference should be done separately for each table. Otherwise, we use the methods in the next step to infer about the common odds-ratio.

(B) Statistical inferences for the *common odds-ratio* across multiple contingence tables.

- The methods in both steps are “Mantel-Haenszel methods”.

Inference of Odds-Ratio in Multiple 2x2 Tables

- **(A)** Test for common odds-ratios.

k=1,2,…,K tables each with T_k subjects.

Observed		X		
		Yes	No	
Y	Yes	a_k	b_k	N_{1k}
	No	c_k	d_k	N_{2k}
		M_{1k}	M_{2k}	T_k

True Parameters		X		
		Yes	No	
Y	Yes	$p_{11,k}$	$p_{12,k}$	$p_{1\cdot,k}$
	No	$p_{21,k}$	$p_{22,k}$	$p_{2\cdot,k}$
		$p_{\cdot 1,k}$	$p_{\cdot 2,k}$	1

- Odds-ratio for the k-th table:

Parameter $OR_k = \frac{p_{11,k}p_{22,k}}{p_{12,k}p_{21,k}}$; Estimate $\widehat{OR}_k = \frac{a_k d_k}{b_k c_k}$.

(A) Test for common odds-ratios.

k=1,2,…,K tables each with T_k subjects.

- Odds-ratio for the k-th table:

Parameter $OR_k = \frac{p_{11,k}p_{22,k}}{p_{12,kp21,k}}$; Estimate $\widehat{OR}_k = \frac{a_k d_k}{b_k c_k}$.

- Test $\mathbf{H_0}$: $OR_1 = OR_2 = \ldots = OR_K$ versus $\mathbf{H_A}$: Not all equal.
- Use χ^2-test on the logarithm of $\widehat{OR}_k$:

$$y_k = \ln(\widehat{OR}_k) = \ln(\frac{a_k d_k}{b_k c_k})$$

(A) Test for common odds-ratios.

- χ^2-test for $\mathbf{H_0}$: $OR_1 = OR_2 = \ldots = OR_K$ using $\ln(\widehat{OR}_k)$.
- Observed: $y_k = \ln(\widehat{OR}_k) = \ln(\frac{a_k d_k}{b_k c_k})$.
- Expected under $\mathbf{H_0}$: Naïve estimate $\bar{Y} = \frac{1}{K}\sum_{k=1}^{K} y_k$

(Issue: bigger table (T_k) has more accurate estimate.)

Use $Y = \frac{\sum_{k=1}^{K} w_k y_k}{\sum_{k=1}^{K} w_k}$ instead with $w_k = \frac{1}{\frac{1}{a_k}+\frac{1}{b_k}+\frac{1}{c_k}+\frac{1}{d_k}} = \frac{1}{\widehat{Var}(y_k)}$

(A) Test for common odds-ratios.

- **Mantel-Haenszel test** for $\mathbf{H_0}$: $OR_1 = OR_2 = \ldots = OR_K$
- Observed: $y_k = \ln(\widehat{OR}_k) = \ln(\frac{a_k d_k}{b_k c_k})$.
- Expected under $\mathbf{H_0}$: $Y = \frac{\sum_{k=1}^{K} w_k y_k}{\sum_{k=1}^{K} w_k}$ and $w_k = \frac{1}{\frac{1}{a_k}+\frac{1}{b_k}+\frac{1}{c_k}+\frac{1}{d_k}}$
- $\chi^2_{Obs} = \sum_{k=1}^{K} w_k (y_k - Y)^2 \approx \chi^2_{df=K-1}$
- Reject $\mathbf{H_0}$ if $\chi^2_{Obs} > \chi^2_{\alpha, df=K-1}$

(A) Test for common odds-ratios.

- **Mantel-Haenszel test** for $\mathbf{H_0}$: $OR_1 = OR_2 = \ldots = OR_K$
- Expected under $\mathbf{H_0}$: $Y = \frac{\sum_{k=1}^{K} w_k y_k}{\sum_{k=1}^{K} w_k}$ and $w_k = \frac{1}{\frac{1}{a_k}+\frac{1}{b_k}+\frac{1}{c_k}+\frac{1}{d_k}}$
- Adjustment for empty cell: if one entry is zero then $w_k = \frac{1}{\infty} = 0$ and the whole k-th table is discarded. If there are empty cells use instead

$$w_k = \frac{1}{\frac{1}{a_k+0.5}+\frac{1}{b_k+0.5}+\frac{1}{c_k+0.5}+\frac{1}{d_k+0.5}}$$

(A) Test for common odds-ratios.

- **Example** (Page 376 of textbook)

Smokers		Coffee drinker		
		Yes	No	
MI	Yes	1011	81	1092
	No	390	77	476
		1401	158	1559

Non-Smokers		Coffee drinker		
		Yes	No	
MI	Yes	383	66	449
	No	365	123	488
		748	189	937

- Smokers $\widehat{OR}_S = \frac{(1011)77}{(390)81} = 2.46$; Non-Smokers $\widehat{OR}_{NS} = \frac{(383)123}{(365)66} = 1.96$
- Are the odds-ratios the same?
- We test $\mathbf{H_0}: OR_S = OR_{NS}$ versus $\mathbf{H_A}: OR_S \neq OR_{NS}$

(A) Test for common odds-ratios.

- **MI/Coffee Example** test $\mathbf{H_0}$: $OR_S = OR_{NS}$
- $y_1 = \ln(\widehat{OR}_S) = \ln(2.46) = 0.9$, $y_2 = \ln(\widehat{OR}_{NS}) = 0.673$.

$$w_1 = \frac{1}{\frac{1}{1011}+\frac{1}{390}+\frac{1}{81}+\frac{1}{77}} = 34.62,\ w_2 = \frac{1}{\frac{1}{383}+\frac{1}{365}+\frac{1}{66}+\frac{1}{123}} = 34.93.$$

$$Y = \frac{(34.62)0.9 + (34.93)0.673}{34.62 + 34.93} = 0.786;$$

$$\chi^2_{Obs} = 34.62(0.9-0.786)^2 + 34.93(0.673-0.786)^2 = 0.896$$

1-pchisq(0.896, df=2-1) to get p-value=0.344

- Fail to reject $\mathbf{H_0}$: $OR_S = OR_{NS}$.

(A) Test for common odds-ratios.

- **MI/Coffee Example** Fail to reject $\mathbf{H_0}$: $OR_S = OR_{NS}$.

There is no strong evidence that the odds ratios are different in these two tables.

Can make inference of the common odds-ratio.

- The usage of χ^2-test in this stage is similar to the usage of F-test for equal variances in ANOVA: we hope to accept the null hypothesis and then proceed to next stage analysis (for ANOVA, the F-test for equal means).

Inference of Odds-Ratio in Multiple 2x2 Tables

- **(B) Mantel-Haenszel** methods for inference on the common odds-ratio of multiple 2x2 tables. Notice that there is only one parameter $OR=OR_1=\ldots=OR_K$ now.

- **(1) Point estimator** $\widehat{OR}=\frac{\sum_{k=1}^{K} a_k d_k/T_k}{\sum_{k=1}^{K} b_k c_k/T_k}$.

- **MI/Coffee Example**

$$\widehat{OR}=\frac{(1011)77/1559+(383)123/937}{(390)81/1559+(365)66/937}=2.18.$$

(B) Inference on the common odds-ratio

- **MI/Coffee Example**

Smokers		Coffee drinker		
		Yes	No	
MI	Yes	1011	81	1092
	No	390	77	476
		1401	158	1559

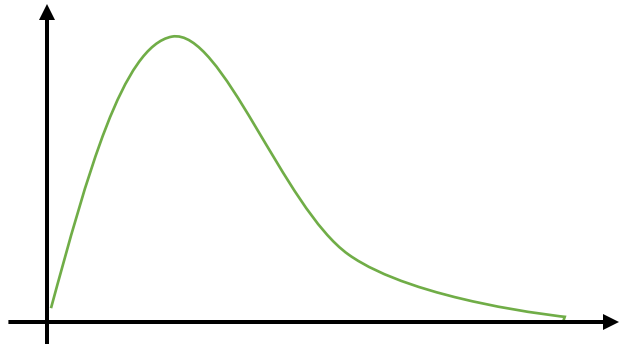
Non-Smokers		Coffee drinker		
		Yes	No	
MI	Yes	383	66	449
	No	365	123	488
		748	189	937

- Point estimator $\widehat{OR} = \dfrac{(1011)77\ /1559+(383)123/937}{(390)81/1559\ +(365)66/937} = 2.18.$
- This differs from the averaged log-odds-ratio Y in **(A)**

Y=0.786 ≠ $\ln(\widehat{OR})$= 0.779; e^{Y}=2.19 ≠ $\widehat{OR}$= 2.18.

(B) Inference on the common odds-ratio

- **(2) Confidence Interval** for OR.



- The range of OR is (0, ∞), and the distribution of $\widehat{OR}$ is skewed.
- $\widehat{OR} \pm \mathrm{L}$ is NOT a good C.I.
- Trick: do the symmetric C.I. on log-scale and then translate back. However, notice that the point estimator of $\ln(\mathrm{OR})$ is $Y = \frac{\sum_{k=1}^{K} w_k \ln(\widehat{OR}_k)}{\sum_{k=1}^{K} w_k}$ as in (A).

The log-scale C.I. centers around Y, not $\ln(\widehat{OR})$!

(B) Inference on the common odds-ratio

- **(2) Confidence Interval** for OR.
- On log-scale $Y = \frac{\sum_{k=1}^{K} w_k \ln(\widehat{OR}_k)}{\sum_{k=1}^{K} w_k}$, and $w_k = \frac{1}{\widehat{Var}[ln(\widehat{OR}_k)]}$.
- How to estimate Var(Y)? If $w_k = \frac{1}{Var[ln(\widehat{OR}_k)]}$, then

$$\text{Var}(Y) = \sum_{k=1}^{K}\{w_k^2\, Var[ln(\widehat{OR}_k)]\}\left(\frac{1}{\sum_{k=1}^{K} w_k}\right)^2 = \sum_{k=1}^{K} w_k \left(\frac{1}{\sum_{k=1}^{K} w_k}\right)^2$$

$$= \frac{1}{\sum_{k=1}^{K} w_k}.$$

So $\widehat{Var}(Y) = \frac{1}{\sum_{k=1}^{K} w_k}$.

(B) Inference on the common odds-ratio

- **(2) Confidence Interval** for OR.
- On log-scale $Y = \frac{\sum_{k=1}^{K} w_k \ln(\widehat{OR}_k)}{\sum_{k=1}^{K} w_k}$, and $\widehat{Var}(Y) = \frac{1}{\sum_{k=1}^{K} w_k}$.
- $Y \approx$normal when sample size is big
- (1-α) C.I. for ln(OR) is $Y \pm z_{\alpha/2} \sqrt{\widehat{Var}(Y)} = Y \pm z_{\alpha/2} \frac{1}{\sqrt{\sum_{k=1}^{K} w_k}}$.
- (1-α) C.I. for OR is $\mathbf{exp}\left(Y \pm z_{\alpha/2} \frac{1}{\sqrt{\sum_{k=1}^{K} w_k}}\right)$.

(B) Inference on the common odds-ratio

- **(2) (1-α) Confidence Interval** for OR **is**

$$\exp\left(Y \pm z_{\alpha/2} \frac{1}{\sqrt{\sum_{k=1}^{K} w_k}}\right),$$

Observed		X		
		Yes	No	
Y	Yes	a_k	b_k	N_{1k}
	No	c_k	d_k	N_{2k}
		M_{1k}	M_{2k}	T_k

where $Y = \frac{\sum_{k=1}^{K} w_k \ln(\widehat{OR}_k)}{\sum_{k=1}^{K} w_k}$, $w_k = \frac{1}{\frac{1}{a_k}+\frac{1}{b_k}+\frac{1}{c_k}+\frac{1}{d_k}}$.

- This use that Y ≈normal when *sample size is big*

Rule of thumb: $\sum_{k=1}^{K} \frac{M_{ik}N_{jk}}{T_k} \geq 5$ for i=1,2 and j=1,2.

(B) Inference on the common odds-ratio

- **MI/Coffee Example**

Smokers		Coffee drinker		
		Yes	No	
MI	Yes	1011	81	1092
	No	390	77	476
		1401	158	1559

Non-Smokers		Coffee drinker		
		Yes	No	
MI	Yes	383	66	449
	No	365	123	488
		748	189	937

- 95% C.I. for $\ln(\mathrm{OR})$ is $0.786 \pm 1.96 \frac{1}{\sqrt{34.62+34.93}}$

$= (0.551, 1.021)$.

- 95% **C.I. for OR** is $(e^{0.551}, e^{1.021})$= (1.73, 2.78).

(B) Inference on the common odds-ratio

- **(3) Hypothesis test for H_0:** $OR = C_0$ versus H_A: $OR \neq C_0$.
- We focus on H_0: $OR = 1$ versus H_A: $OR \neq 1$.
- (a) We can always use the confidence interval.

  For the MI/coffee example, $1 \notin (1.73, 2.78)$.

  So reject H_0.

- (b) But we want a formal test also, which will be done using a χ^2-test.

(B) Inference on the common odds-ratio

- **(3) Hypothesis test for H_0:OR=1 versus H_A: OR≠1.**
- (b) A Mantel-Haenszel χ^2-test.

Observed		X		
		Yes	No	
Y	Yes	a_k	b_k	N_{1k}
	No	c_k	d_k	N_{2k}
		M_{1k}	M_{2k}	T_k

- $\chi^2_{Obs} = \frac{(\sum_{k=1}^{K} a_k - \sum_{k=1}^{K} \frac{M_{1k}N_{1k}}{T_k})^2}{\sum_{k=1}^{K} \frac{M_{1k}M_{2k}N_{1k}N_{2k}}{T_k^2}} \approx \chi^2_{df=1}$
- Reject H_0 if $\chi^2_{Obs} > \chi^2_{\alpha,df=1}$

(B3b) Ideas of Mantel-Haenszel test for H_0: $OR=1$.

Observed		X: Yes	X: No	
Y	Yes	a_k	b_k	N_{1k}
	No	c_k	d_k	N_{2k}
		M_{1k}	M_{2k}	T_k

True Parameters		X: Yes	X: No	
Y	Yes	$p_{11,k}$	$p_{12,k}$	$p_{1\cdot,k}$
	No	$p_{21,k}$	$p_{22,k}$	$p_{2\cdot,k}$
		$p_{\cdot 1,k}$	$p_{\cdot 2,k}$	1

- Notice $E(a_k \mid \mathbf{H_0})=T_k p_{11,k}=T_k p_{1\cdot,k}p_{\cdot 1,k}$ is estimated by

$$T_k \frac{M_{1k}}{T_k}\frac{N_{1k}}{T_k}=\frac{M_{1k}N_{1k}}{T_k}.\quad \chi^2_{Obs}=\frac{(\sum_{k=1}^K a_k-\sum_{k=1}^K \frac{M_{1k}N_{1k}}{T_k})^2}{\sum_{k=1}^K \frac{M_{1k}M_{2k}N_{1k}N_{2k}}{T_k^2}}$$ comes

from $$\frac{\left[\sum_{k=1}^K a_k-E(\sum_{k=1}^K a_k \mid \mathbf{H_0})\right]^2}{\widehat{Var}(\sum_{k=1}^K a_k \mid \mathbf{H_0})}.$$

(B3b) Ideas of Mantel-Haenszel test for H_0: $OR=1$.

- This test statistic is based on

$\left[\sum_{k=1}^{K} a_k - \mathrm{E}\left(\sum_{k=1}^{K} a_k \,|\mathbf{H_0}\right)\right]^2$ rather than the standard χ^2-test $\sum_{k=1}^{K}[a_k - \mathrm{E}(a_k\,|\mathbf{H_0})]^2$.

- The standard χ^2-test statistic combines entries in different cells of one table.
- Here we are combining entries in same cells (a_k) across different tables.

(B3b) Test H_0: $OR=1$ in MI/Coffee Example

Smokers		Coffee drinker: Yes	Coffee drinker: No	
MI	Yes	1011	81	1092
	No	390	77	476
		1401	158	1559

Non-Smokers		Coffee drinker: Yes	Coffee drinker: No	
MI	Yes	383	66	449
	No	365	123	488
		748	189	937

a_1=1011, $\frac{M_{11}N_{11}}{T_1}=\frac{(1401)1092}{1559}$=981.3, a_2=383, $\frac{M_{12}N_{12}}{T_2}=\frac{(748)449}{937}$=358.4.

$$\chi^2_{Obs}=\frac{[(1011+383)-(981.3+358.4)]^2}{\frac{(1401)1092(158)476}{(1559)^2(1559-1)}+\frac{(748)449(189)488}{(937)^2(937-1)}}=43.68.$$

1-pchisq(43.68, df=2-1) to get p-value=3.87×10^{-11}. Reject H_0: $OR=1$

- Drinking coffee is associated with myocardial infarction.

Chapter 16 Inference on Multiple 2x2 Tables

- Notice that there are two χ^2-test in this chapter
- **(1) Mantel-Haenszel test** for $\mathbf{H_0}$: $OR_1 = OR_2 = \ldots = OR_K$

This is a goodness-of-fit test. We want large p-value so that we can proceed to inference on common odds-ratio.

- **(2) Mantel-Haenszel test** for $\mathbf{H_0}$: $OR = 1$.

  Small p-value ⇒ the marginals are not independent.

- The test in (2) is based on point estimator $\widehat{OR}$. But confidence interval is built on $Y = \ln(\widehat{OR})$, not $\ln(\widehat{OR})$.

Summary

Today, we finished Module 11 Multiple contingency tables (Chapter 16 of textbook)

- Simpson's Paradox: the odds-ratio in combined contingency table can be misleading!
- **(A)** Test for common odds-ratios.
- **(B)** Inference on the common odds-ratio.

- Homework 8 includes topics in this module and the next module.